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

Survey Methodology

Survey Methodology
47-2

Release date: January 6, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

Email at infostats@statcan.gc.ca

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

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

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

December 2021

•

Volume 47

•

Number 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

Survey Methodology usually publishes innovative theoretical or applied research papers, and sometimes review papers, that provide new insights on statistical methods relevant to National Statistical Offices and other statistical organizations. Topics of interest are provided on the journal web site (www.statcan.gc.ca/surveymethodology). Authors can submit papers either to the regular section of the Journal or to the short notes section for contributions under 3,000 words, including tables, figures and references. Although the review process may be streamlined for short notes, all papers are peer-reviewed. However, the authors retain full responsibility for the contents of their papers, and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada
Volume 47, Number 2, December 2021

Contents

Waksberg Invited Paper Series

Regular Papers

Short notes

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 227-228
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. The recipient of the Waksberg Award is also invited to give the Waksberg Invited Address, usually at the Statistics Canada Symposium, and receives an honorarium.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2023 Waksberg Award.

This issue of *Survey Methodology* opens with the 21th paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Bob Fay (Chair), Jean Opsomer, Jack Gambino and Elizabeth Stuart for having selected Sharon Lohr as the author of 2021 Waksberg Award paper.

2021 Waksberg Invited Paper

Author: Sharon Lohr

Sharon Lohr has published widely about survey sampling, design of experiments, and statistical methods for education, public policy, law, and crime. She is the author of numerous articles in statistics journals and of the books *Sampling: Design and Analysis*, now in its third edition, and *Measuring Crime: Behind the Statistics*. Formerly Dean's Distinguished Professor of Statistics at Arizona State University and a Vice President at Westat, she is now a statistical consultant and writer.

Sharon is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute, and currently serves as a member of the Committee on National Statistics of the U.S. National Academies of Sciences, Engineering, and Medicine. She was the inaugural recipient of the Gertrude M. Cox award for contributions to statistical practice, and has been honored by being selected to give the Morris Hansen Lecture at the Washington Statistical Society and the Deming Lecture at the Joint Statistical Meetings.

Waksberg Award honorees and their invited papers since 2001

2022 Roderick **Little**, Manuscript topic under consideration, (expected for vol. 48, 2).

2021 Sharon **Lohr**, “Multiple-frame surveys for a multiple-data-source world”. *Survey Methodology*, vol. 47, 2, 229-263.

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, December 2021
Vol. 47, No. 2, pp. 229-263
Statistics Canada, Catalogue No. 12-001-X

Multiple-frame surveys for a multiple-data-source world

Sharon L. Lohr[1]

Abstract

Multiple-frame surveys, in which independent probability samples are selected from each of Q sampling frames, have long been used to improve coverage, to reduce costs, or to increase sample sizes for subpopulations of interest. Much of the theory has been developed assuming that (1) the union of the frames covers the population of interest, (2) a full-response probability sample is selected from each frame, (3) the variables of interest are measured in each sample with no measurement error, and (4) sufficient information exists to account for frame overlap when computing estimates. After reviewing design, estimation, and calibration for traditional multiple-frame surveys, I consider modifications of the assumptions that allow a multiple-frame structure to serve as an organizing principle for other data combination methods such as mass imputation, sample matching, small area estimation, and capture-recapture estimation. Finally, I discuss how results from multiple-frame survey research can be used when designing and evaluating data collection systems that integrate multiple sources of data.

Key Words: Combining data; Data integration; Dual-frame survey; Indirect sampling; Mass imputation; Misclassification; Survey design; Undercoverage.

1. Introduction

Throughout his 33-year career at the Census Bureau and subsequent 32-year career at Westat, Joe Waksberg repeatedly relied on multiple data sources to improve the quality of estimates while reducing costs. He used external data sources to evaluate coverage in the U.S. decennial census (Marks and Waksberg, 1966; Waksberg and Pritzker, 1969), to calibrate survey weights, and to improve efficiency or oversample rare populations when designing surveys (Hendricks, Igra and Waksberg, 1980; Cohen, DiGaetano and Waksberg, 1988; DiGaetano, Judkins and Waksberg, 1995; Waksberg, 1995; Waksberg, Judkins and Massey, 1997b).

On several occasions, Waksberg integrated data from two or more surveys directly in order to improve coverage or to obtain larger sample sizes for subpopulations (Waksberg, 1986; Burke, Mohadjer, Green, Waksberg, Kirsch and Kolstad, 1994; Waksberg, Brick, Shapiro, Flores-Cervantes and Bell, 1997a). In these multiple-frame surveys, independent samples were selected from sampling frames that together were thought to cover all, or almost all, of the target population. The data from the samples were combined to obtain estimates for the population as a whole and for subpopulations of interest. Waksberg approached the design of these multiple-frame surveys from the perspective of controlling both sampling and nonsampling errors, and found that using multiple frames met the challenges of producing reliable estimates in the face of increased data collection costs (with higher nonresponse for less expensive collection methods) and incomplete frame coverage.

---

[1] Sharon Lohr is Professor Emerita at Arizona State University. E-Mail: sharon.lohr@asu.edu.

Statistical agencies and survey organizations today face the same types of challenges that Waksberg addressed – declining response rates and increasing costs of survey data collection – but at an intensified level. At the same time, the emergence of new data sources provides opportunities for obtaining information about parts of populations of interest – sometimes with amazing rapidity. Many organizations are now using or researching methods for integrating data from multiple sources to improve the accuracy or timeliness of population estimates.

I feel tremendously honored to be asked to give the Waksberg lecture, and in this paper I want to build on Waksberg's insights about multiple-frame surveys by discussing their use as an organizing principle for combining information from multiple sources. Traditionally, multiple-frame surveys have integrated data from Q probability samples $S_1, \ldots, S_Q$ that are selected independently from Q frames. But the general structure can be expanded to include frames that consist of administrative records or nonprobability samples. The structure can also be expanded to situations in which some data sources do not measure the variables of interest y but they measure covariates $\mathbf{x}$ that can be used to predict y.

A number of authors have reviewed methods for combining data from multiple sources; see, for example, Citro (2014), Lohr and Raghunathan (2017), National Academies of Sciences, Engineering, and Medicine (2017, 2018), Thompson (2019), Zhang and Chambers (2019), Beaumont (2020), Yang and Kim (2020), and Rao (2021). The sources include traditional probability samples, administrative data sets, sensor data, social network data, and general convenience samples.

Although the types of data (and the speed with which some types of data can be collected) have changed in recent years, the basic structure of the problem for combining data sources is unchanged from the earliest dual-frame surveys. Section 2 discusses the structure and assumptions for traditional multiple-frame surveys through the example of the National Survey of America's Families, a dual-frame survey that Waksberg worked on during the 1990s. Section 3 reviews methods for calculating estimates of population characteristics from traditional multiple-frame surveys where all assumptions are met, including the special case in which one sample is a census of a subset of the population. Section 4 then discusses how the multiple-frame structure incorporates many of the methods currently used for combining data, sometimes with relaxed assumptions. Section 5 addresses issues for designing data collection systems that control sampling and nonsampling errors, with a discussion of possible future directions for research.

2. Classical multiple-frame survey structure and assumptions

First, let's look at an example of what I shall call a "classical" multiple-frame survey – a survey that is designed to take probability samples from each of a fixed number of frames – and define the notation and assumptions that will be used to describe estimators and their properties.

2.1 National Survey of America's Families

The goal of the 1997 National Survey of America's Families (NSAF) was to provide information on social and economic characteristics of the U.S. civilian noninstitutional population under age 65, with emphasis on obtaining reliable estimates for persons and families – particularly families with children – below 200 percent of the poverty threshold. Estimates were desired for the nation as a whole; in addition, separate estimates were desired for 13 states that were purposively selected to vary by geographic region, dominant political party, size, and fiscal capacity.

To meet the precision requirements for estimates, it was desired to have an effective sample size of about 800 poor children in each state. This goal could have been met by taking a household sample from an area frame. Waksberg et al. (1997b) had determined that screening households for income and subsampling nonpoor households would be the most cost-effective way of achieving the desired sample sizes in an area-frame sample, but the cost would be high because only about one in eight families was expected to have children and be under 200 percent of the poverty threshold.

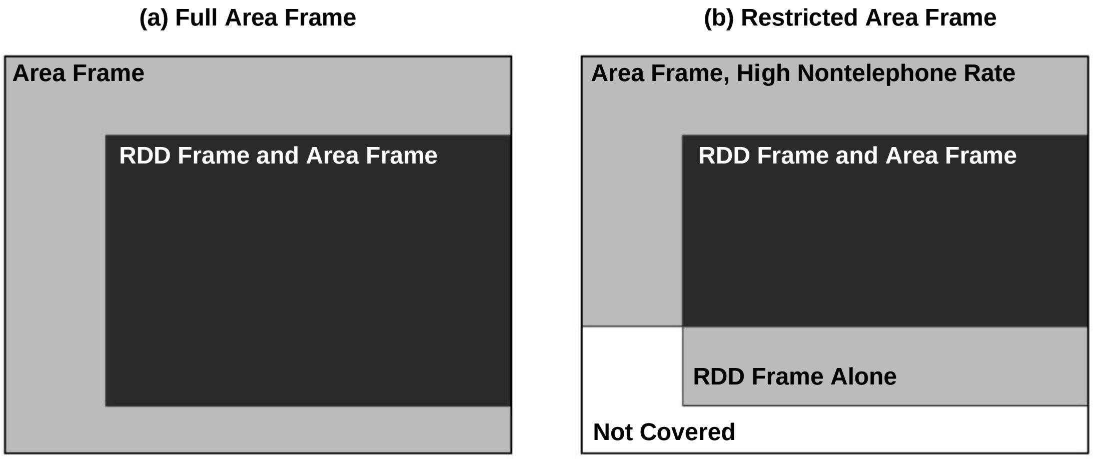
Screening costs would be greatly reduced if the survey could be conducted by telephone using random digit dialing (RDD). But Current Population Survey data indicated that about 20 percent of families living in poverty did not have telephones, so the RDD frame was expected to have substantial undercoverage of the target population. Moreover, households under 200 percent of poverty without telephones might have different income levels or health characteristics than households under 200 percent of poverty with telephones.

Thus, a sample from the area frame would provide high coverage but also come with unacceptably high costs. An RDD survey would have lower costs but would have substantial undercoverage of the population of interest. Waksberg et al. (1997a) used a dual-frame survey, with one sample from the area frame and a second sample chosen independently from the RDD frame, to take advantage of the lower costs of an RDD sample yet also cover nontelephone households. Figure 2.1(a) shows the structure of the two frames.

To further reduce costs, Waksberg et al. (1997a) excluded census block groups with few nontelephone households from the area frame; according to the 1990 census, the excluded areas accounted for less than ten percent of the nontelephone households in each state. With this exclusion, the area and RDD frames each contained households not found in the other frame, as shown in Figure 2.1(b).

Households with telephones that were in the non-excluded block groups were present in both frames. If a probability sample were taken from each frame, households in that overlap (the dark shaded area in Figure 2.1(b)) could be selected in both samples. The survey designers could either conduct the interview with all households in each sample and then deal with the multiplicity in the estimation (an overlap design), or screen out the households in one of the frames that were also in the other frame (a screening design).

Figure 2.1 Frame coverage for the NSAF. The dark shaded area is in both frames.



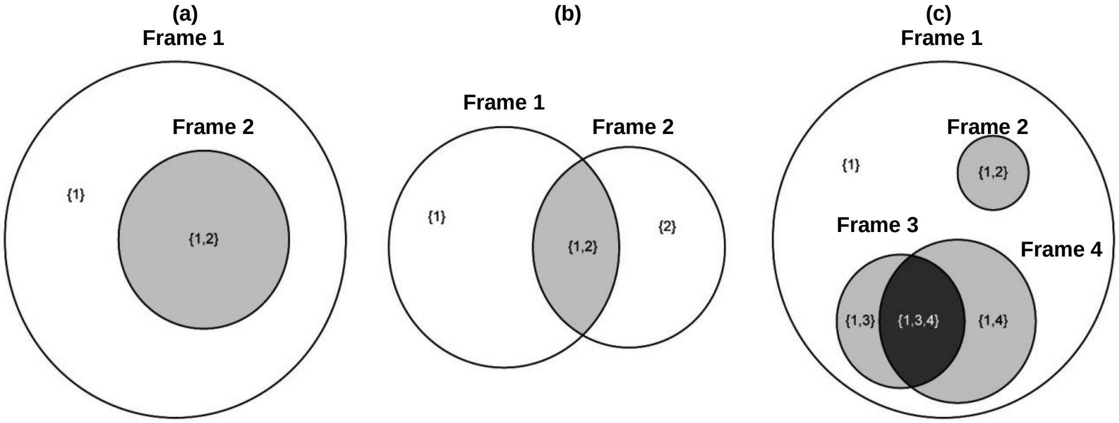
Waksberg and his colleagues chose to use screening. Households in the area sample were asked if they had a telephone, and only those without telephones were administered the detailed interview. The detailed interview was lengthy and expensive to conduct; screening out the telephone households during a short interview saved resources that could be used to increase the number of nontelephone households in the sample. Households with telephones were sampled only through the RDD frame; households in the RDD sample with no children and above 200 percent of the poverty line were subsampled. Because a screening survey was used, the combined sample from the two surveys was a stratified sample, and resources were allocated to the two samples using stratified sampling formulas that accounted for the higher cost of sampling from the area frame.

2.2 Notation and assumptions for multiple-frame surveys

In classical multiple-frame surveys such as the NSAF, a number of assumptions are needed to be able to obtain unbiased estimates of population characteristics along with confidence intervals having approximately correct coverage probabilities.

Suppose there are Q frames. A population domain d is defined by the intersections of the frames: domain {1, 3, 4}, for example, contains the population units that are in Frames 1, 3, and 4 but not in any of the other frames. Let D denote the set of possible domains; depending on the overlap of units, D can contain between 1 and $2^Q - 1$ domains. Figure 2.2 shows three examples of frame relationships. When Frame 1 is complete but Frame 2 is incomplete as in Figure 2.2(a), $D = \{\{1\}, \{1, 2\}\};$ any population unit in Frame 2 is also in Frame 1. For an overlapping dual-frame survey such as that in Figure 2.2(b), $D = \{\{1\}, \{2\}, \{1, 2\}\}.$

Figure 2.2 Three frame structures. (a) Frame 1 has complete coverage and Frame 2 is incomplete. (b) Frames 1 and 2 are both incomplete but overlap. (c) Frame 1 is complete; Frames 2, 3, and 4 are all incomplete but Frames 3 and 4 overlap.



Define $\delta_i(d) = 1$ if unit i is in domain d and 0 otherwise, and let $\delta_i^{(q)} = 1$ if unit i is in Frame q and 0 otherwise. Frame q has population size $N^{(q)}$ and domain d has population size N_d; these sizes may be known or unknown. The target population has a total of N units.

The following assumptions are typically made in order to draw inferences from classical multiple-frame surveys.

(A1) The union of the Q frames covers the target population.

(A2) The sample S_q taken from Frame q is a probability sample where unit i has probability $\pi_i^{(q)}$ of being in S_q. Let $w_i^{(q)}$ represent the final weight for unit i in S_q; options for $w_i^{(q)}$ include the design weight $1/\pi_i^{(q)}$, the Hájek weight $N^{(q)}/[\hat{N}^{(q)}\pi_i^{(q)}]$ with $\hat{N}^{(q)} = \sum_{j \in S_q} 1/\pi_j^{(q)}$, or a nonresponse-adjusted weight.

(A3) The samples $S_1, \ldots, S_Q$ are selected independently.

(A4) The domain membership of each unit i in S_q, $\{\delta_i(d), d \in D\}$, is known.

(A5) The estimator of the population total in domain d from S_q, $\hat{Y}_d^{(q)} = \sum_{i \in S_q} \delta_i(d) w_i^{(q)} y_i$, is approximately unbiased for $Y_d = \sum_{i=1}^{N} \delta_i(d) y_i$, for all Frames q containing domain d and for all variables y.

(A6) There is no measurement error. If unit i is in Frame q and Frame q', y_i will have the same value if measured in S_q as it will if measured in $S_{q'}$.

These are strong assumptions; some relaxation of individual assumptions is possible for specific estimators, as discussed in Section 3. But they are weaker than assumptions needed for some of the other possible data integration methods. Record linkage, for example, has an implicit assumption that unit i in Frame q can be matched with a specific unit in Frame q'. For multiple-frame surveys, one must know whether a unit sampled from Frame q is also in other frames, but does not need to identify the matched unit.

2.3 Were the assumptions met in the NSAF?

Survey assumptions are rarely met exactly in practice, and the NSAF was no exception. Assumption (A1) was not met because of the exclusion of block groups with high telephone ownership. The sample from the area frame yielded fewer nontelephone households than expected, perhaps because of measurement error in the 1990 census or population changes since 1990. In addition, post-survey investigations using data from the 1997 Current Population Survey indicated that the block groups excluded from the frame may have had more nontelephone households than anticipated (Waksberg, Brick, Shapiro, Flores-Cervantes, Bell and Ferraro, 1998).

Although independent probability samples were taken from each frame, each sample had nonresponse. The estimated response rates for children were 65 percent in the RDD sample and 84 percent in the area sample. The weighting procedure attempted to address potential bias from undercoverage and nonresponse. The weights of the nontelephone households in the area sample were ratio-adjusted to attempt to compensate for undercoverage from the block group exclusions. Nonresponse-adjusted weights were calculated separately for the area- and RDD-frame samples, and then the combined samples were poststratified to Census Bureau control totals (Brick, Shapiro, Flores-Cervantes, Ferraro and Strickler, 1999). Groves and Wissoker (1999) found little evidence of residual bias in their nonresponse bias analysis; one of the few differences they reported was that households in the RDD sample that required more calls for contact, and households in a subsample taken of nonrespondents, were slightly less likely to be receiving food assistance.

In the NSAF, the domain membership was determined by asking household respondents in the area sample if they had a working telephone. If that question was answered accurately, then Assumption (A4) was met. The investigators attempted to reduce measurement error for Assumption (A6) by having centralized telephone interviewers conduct all of the detailed interviews; households in the area frame were interviewed over a cellular telephone brought by the field representative. Because interviews in domain $\{1, 2\}$ were obtained only from the RDD sample, however, no data are available for evaluating possible measurement error or relative nonresponse bias for the two samples.

Waksberg had used dual-frame surveys several times prior to the NSAF, mostly to increase sample sizes when sampling rare populations, but he recommended using them only when a simpler design would not meet the survey objectives. He wrote: “The price is additional complexity in the sampling operations and the possibility of error if the matching of the two frames is not done carefully.... My instincts are that a more complex scheme should not be used unless there is a reasonably good pay-off” (Waksberg, 1986).

Was the extra complication and expense of the dual-frame design worth the effort in the NSAF? Because telephone households were screened out of the area sample, and because the yield of nontelephone households was less than anticipated, only 1,488 of the total of 44,461 interviewed households came from the area sample. But because of the high poverty rate of the nontelephone households, the estimated percentage of children in households under 200 percent of the poverty threshold was about 3.6 percentage points higher with the full sample than with the RDD sample alone. Even though for many variables there was only a small difference between the full-sample estimate and the RDD-sample estimate, that difference could not have been evaluated without the area sample.

3. Estimation in classical multiple-frame surveys

The main problem for inference in a classical multiple-frame survey – one that is designed so as to satisfy Assumptions (A1) to (A6) – is how to account for potential overlap among the samples. In the NSAF, telephone households were screened out of the area sample, but in many applications screening is infeasible or it is more cost-effective to obtain data from the full sample selected from each frame. When separate surveys or data sources are not designed with data combination in mind, the overlap depends on the coverage of the individual data sources.

With an overlap design, units that are contained in more than one frame have multiple chances for being selected in the sample. An estimator constructed by summing the weighted observations from each of the Q samples,

$$\hat{Y}_{\text{concat}} = \sum_{q=1}^{Q} \sum_{i \in S_q} w_i^{(q)} y_i,$$

will be a biased estimator of $Y = \sum_{i=1}^{N} y_i$ because the individual sample weights do not reflect the multiple chances of selection for units in overlap domains. Methods for estimating population totals thus typically multiply the survey weights $w_i^{(q)}$ by a multiplicity adjustment $m_i^{(q)}$ that satisfies $\sum_{q=1}^{Q} \delta_i^{(q)} m_i^{(q)} \approx 1$ for each unit $i,$ resulting in the estimator

$$\hat{Y} = \sum_{q=1}^{Q} \sum_{i \in S_q} w_i^{(q)} m_i^{(q)} y_i = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i, \tag{3.1}$$

where $\tilde{w}_i^{(q)} = w_i^{(q)} m_i^{(q)}$ is the multiplicity-adjusted weight.

3.1 Hartley's composite estimator

Hartley (1962) was the first author to present a rigorous theory of estimation in dual-frame surveys where units in the overlap domain {1, 2} might be sampled from both frames. This four-page paper made several important contributions. First, Hartley defined the problem in statistical terms. Second, he proposed an optimal estimator for combining the estimates from the two surveys. And third, he studied the design problem of allocating the resources to the different samples, with a joint consideration of the allocation and the estimator that minimize the variance of the estimated population total subject to a fixed cost.

Hartley (1962) estimated the population total $Y = \sum_{i=1}^{N} y_i$ by

$$\hat{Y}(\theta) = \hat{Y}^{(1)}_{\{1\}} + \hat{Y}^{(2)}_{\{2\}} + \theta \hat{Y}^{(1)}_{\{1,2\}} + (1-\theta)\,\hat{Y}^{(2)}_{\{1,2\}}. \tag{3.2}$$

He proposed choosing θ to minimize $V\left[\hat{Y}(\theta)\right].$ This resulted in the value

$$\theta_H = \frac{V\left(\hat{Y}^{(2)}_{\{1,2\}}\right) + \mathrm{Cov}\left(\hat{Y}^{(2)}_{\{2\}}, \hat{Y}^{(2)}_{\{1,2\}}\right) - \mathrm{Cov}\left(\hat{Y}^{(1)}_{\{1\}}, \hat{Y}^{(1)}_{\{1,2\}}\right)}{V\left(\hat{Y}^{(1)}_{\{1,2\}}\right) + V\left(\hat{Y}^{(2)}_{\{1,2\}}\right)}. \tag{3.3}$$

The estimator in (3.2) is of the form in (3.1) with multiplicity weight adjustments

$$m^{(1)}_i = \delta_i\left(\{1\}\right) + \delta_i\left(\{1,2\}\right)\theta, \quad m^{(2)}_i = \delta_i\left(\{2\}\right) + \delta_i\left(\{1,2\}\right)\left(1-\theta\right).$$

If it is desired to use the optimal compositing factor θ_H, estimators may be substituted for the unknown covariances in (3.3). Because θ_H depends on covariances involving y, however, the optimal multiplicity adjustment may differ for different variables, giving a different set of weights for each. In addition, θ_H can be less than 0 or greater than 1, possibly resulting in negative weights for some observations. These features carry over to the Q-frame generalization of Hartley's optimal estimator studied by Lohr and Rao (2006).

The estimator in (3.2), with fixed value of θ, is approximately unbiased for Y under Assumption (A5). If the estimated domain totals and the estimates of the covariances in (3.3) are consistent, then the estimator with $\hat{\theta}_H$ is consistent for Y. Saegusa (2019) studied Hartley's estimator from the perspective of empirical process theory, establishing a law of large numbers and a central limit theorem when S_1 and S_2 are both simple random samples.

Hartley's application was in agriculture, and many of the early applications of dual-frame surveys were for agriculture or business surveys (Kott and Vogel, 1995), where list frames existed that contained the larger business or agricultural operations. A dual-frame survey with a disproportionately larger sample from the list frame reduced costs because (1) obtaining data from an operation in the list frame was often less expensive than obtaining data from an operation in the area frame and (2) oversampling the list frame was analogous to oversampling high-variance strata in stratified sampling and thus resulted in greater efficiency.

Later, as cellular telephones became more prevalent, concern about bias from using landline telephone samples alone led to use of dual-frame telephone surveys, with one sample from a landline frame and a second sample from a cellular telephone frame. Here, both frames are incomplete but together cover the population of persons with telephones. For these surveys, an important consideration is how to deal with persons having both kinds of telephones. The next section reviews choices for the compositing.

3.2 Multiplicity weighting adjustments

Hartley's optimal estimator, with θ_H, uses a different set of weights for each response variable, which can lead to internal inconsistencies among estimators. Various authors have proposed estimators that use a single set of weights for all analyses. Here, I briefly list some of the multiplicity adjustment factors $m_i^{(q)}$ that result in one set of weights for the general estimator of the population total in (3.1). The resulting estimators are approximately unbiased for the population total Y under Assumptions (A1), (A4), and (A5). These and additional estimators are reviewed in detail by Lohr (2011), Lu, Peng and Sahr (2013), Ferraz and Vogel (2015), Arcos, Rueda, Trujillo and Molina (2015), and Baffour, Haynes, Western, Pennay, Misson and Martinez (2016).

- Screening estimator, with $m_i^{(1)} = 1, m_i^{(2)} = 1 - \delta_i^{(1)}, \ldots, m_i^{(Q)} = \prod_{q=1}^{Q-1} \left(1 - \delta_i^{(q)}\right).$ A unit sampled from Frame q is discarded if it is in any of Frames $1, \ldots, q-1.$ This estimator is automatically used with a screening design such as the NSAF; with an overlap design, its use means that some data observations are thrown away.
- Multiplicity estimator, with $m_i^{(q)} = 1/$ (number of frames containing unit i) $= 1 \Big/ \sum_{q=1}^{Q} \delta_i^{(q)}$. In a dual-frame survey, this gives the estimator in (3.2) with $\theta = 1/2.$ Mecatti (2007) noted that with the multiplicity estimator, Assumption (A4) can be replaced by the slightly less restrictive assumption that $\sum_{q=1}^{Q} \delta_i^{(q)}$ is known for each sampled unit $i.$

  The multiplicity estimator can also be viewed as a special case of the generalized weight share method (Deville and Lavallée, 2006) using the standardized link matrix, since the number of links to population unit i is the number of frames containing that unit.
- Single-frame estimator (Bankier, 1986; Kalton and Anderson, 1986), which considers the observations as if they had been sampled from a single frame. If inverse probability weights are used, with $w_i^{(q)} = 1 \big/ \pi_i^{(q)}$, then $m_i^{(q)} = \pi_i^{(q)} \Big/ \sum_{f=1}^{Q} \delta_i^{(f)} \pi_i^{(f)}$. This estimator requires that the inclusion probability for unit i be known for all Q frames, including frames from which the unit was not sampled. The multiplicity adjustments consider the inclusion probabilities for the designs but not the relative variances, which are affected by clustering and stratification in the individual samples.
- Effective sample size (ESS) estimator (Chu, Brick and Kalton, 1999; O'Muircheartaigh and Pedlow, 2002), where the domain estimator from each frame is weighted by the relative effective sample size from that frame. Let $n^{(q)}$ be the sample size from Frame q and let $\text{deff}^{(q)}$ denote the design effect for a key variable or a smoothed design effect for multiple variables. The effective sample size for S_q is $\tilde{n}^{(q)} = n^{(q)} \big/ \text{deff}^{(q)}$ and the multiplicity adjustment for unit i is

$$m_i^{(q)} = \frac{\tilde{n}^{(q)}}{\sum_{f=1}^{Q} \delta_i^{(f)} \tilde{n}^{(f)}}.$$

  This estimator considers the relative variances of estimators from different samples and is often more efficient than the screening, multiplicity, and single-frame estimators.

> The pseudo-maximum-likelihood (PML) estimator of Skinner and Rao (1996) is of this type when the frame sizes $N^{(q)}$ and domain sizes N_d are unknown; Skinner and Rao (1996) recommended using the design effect for estimating $N_{\{1,2\}}$ to establish the effective sample size for the dual-frame case. The PML estimator is asymptotically equivalent to an ESS estimator that poststratifies to the domain sizes N_d when those are known; when the frame sizes $N^{(q)}$ are known but not $N_{\{1,2\}}$, the PML estimator is asymptotically equivalent to calibrating the ESS estimator to estimated domain sizes calculated from the pseudo-likelihood function.

Approximately unbiased estimates of the variances for all estimators considered in this section can be derived under Assumptions (A1) to (A6) and additional regularity conditions that ensure consistency of estimated totals and variance estimators from the Q samples. Skinner and Rao (1996) studied linearization variance estimators; Chauvet (2016) derived linearization variance estimators for the French housing survey that accounted for the variance reduction due to high sampling fractions from some of the frames. Lohr and Rao (2000) developed theory for using the jackknife with multiple frames, and Lohr (2007) and Aidara (2019) considered bootstrap variance estimators. These methods rely on Assumption (A3) of independent samples; Chauvet and de Marsac (2014) considered the situation in which the samples share primary sampling units but independent samples are taken at the second stage of the design.

Calculating linearization variance estimates requires special software that implements the partial derivative calculations for the multiple frames. Replication variance estimation methods such as jackknife and bootstrap, however, can be calculated in standard survey software by creating a single data set that contains all the concatenated observations and weights $\tilde{w}_i^{(q)}$ from the Q samples and creating replicate weights using standard methods for stratified multistage samples (Metcalf and Scott, 2009). The concatenated data set has $\sum_{q=1}^{Q} H_q$ strata, where H_q is the number of strata for S_q; observations from different samples are in different strata. The replicate weight methods also can include effects of calibration (see Section 3.3) on the variance.

Of course, many applications call for estimates of quantities other than population totals, and the multiple-frame theory applies to parameters that are smooth functions of domain totals. A different compositing factor may be desired when quantities other than population totals are of primary interest, however, and there may be special considerations for other types of analyses. Other types of statistical analyses that have been studied in the multiple-frame setting include linear (Lu, 2014b) and nonparametric (Lu, Fu and Zhang, 2021) regression, logistic regression with ordinal data (Rueda, Arcos, Molina and Ranalli, 2018), empirical distribution functions (Arcos, Martínez, Rueda and Martínez, 2017), gross flow estimation with missing data (Lu and Lohr, 2010), and chi-squared tests (Lu, 2014a).

Lu (2014b) noted that linear regression parameters estimated using the multiplicity-adjusted weights are the finite population regression coefficients $\mathbf{B}$ that minimize the sum of squares $\sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \mathbf{B})^2$. However, one of the reasons for taking a multiple-frame survey, rather than using an incomplete frame, is a concern that population characteristics may differ across domains. Lu (2014b) suggested examining the residuals separately by domain and also fitting separate regression models by domain to assess the appropriateness of the regression model.

3.3 Calibration

The PML estimator is calibrated to population counts that are known for the frames and domains. In a dual-frame survey where $N^{(1)}$ and $N^{(2)}$ are known, $\sum_{q=1}^{2}\sum_{i \in S_q} w_i^{(q)} m_{i,\,\mathrm{PML}}^{(q)} \delta_i^{(f)} = N^{(f)}$ for $f = 1, 2.$ If the overlap domain size $N_{\{1,\,2\}}$ is also known, the PML estimator is calibrated to all three domain sizes. Skinner (1991) used calibration with the single-frame estimator, raking the estimator to the population frame counts.

Ranalli, Arcos, Rueda and Teodoro (2016) studied general calibration theory for dual-frame surveys. They assumed that a vector of auxiliary information $\mathbf{x}$ is available with known population totals $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i,$ and calculated multiple-frame generalized regression weights as

$$c_i^{(q)} = \tilde{w}_i^{(q)} \left[1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{f=1}^{Q} \sum_{k \in S_f} \alpha_k \tilde{w}_k^{(f)} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \alpha_i \mathbf{x}_i \right], \tag{3.4}$$

where α_k is an arbitrary constant and $\hat{\mathbf{X}} = \sum_{f=1}^{Q} \sum_{k \in S_f} \tilde{w}_k^{(f)} \mathbf{x}_k$ estimates $\mathbf{X}$ using the multiplicity-adjusted weights. Under regularity conditions, they showed that for the dual-frame estimator in (3.2) with fixed $\theta,$ the variance of the generalized regression estimator $\hat{Y}_{\mathrm{GR}} = \sum_{q=1}^{2} \sum_{i \in S_q} c_i^{(q)} y_i$ is approximated by

$$V(\hat{Y}_{\mathrm{GR}}) \approx V \left[\sum_{q=1}^{2} \sum_{i \in S_q} \tilde{w}_i^{(q)} (y_i - \mathbf{x}_i^T \mathbf{B}) \right], \tag{3.5}$$

where $\mathbf{B} = \left(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i y_i.$ The variance of the estimator depends on the residuals from the regression model just as in the single-frame case.

Särndal and Lundström (2005) distinguished among types of auxiliary information that can be used in calibration. InfoU is information available at the population level. A vector $\mathbf{x}^*$ can be considered as InfoU if the population total $\mathbf{X}^* = \sum_{i=1}^{N} \mathbf{x}_i^*$ is known and $\mathbf{x}^*$ is observed for every respondent in the sample. InfoS is information available at the level of the sample, but not at the population level. Vector $\mathbf{x}^o$ is InfoS if it is known for every member of the sample, both respondents and nonrespondents, but $\sum_{i=1}^{N} \mathbf{x}^o$ is unknown.

In a multiple-frame survey, the variables available for InfoU and InfoS may differ across frames. For the NSAF, little auxiliary information was known for nonrespondents in the RDD sample but address-related information (for example, characteristics of the block group) was known for all members of the area-frame sample. The reverse may be true for a dual-frame survey in which Frame 1 is an area frame and Frame 2 is a list frame. The list frame may have rich information that can be used for weighting class adjustments or calibration, while the auxiliary information for the area frame may be restricted to information measured in the survey for which population totals are known from an external source such as a census or population register.

Ranalli et al. (2016) allowed for differing InfoU information across the frames; some of the auxiliary variables may be known for units from all samples and for the full population, while other variables may be of the form $x_i^* = x_i \delta_i^{(q)}$ with total $X^* = \sum_{i=1}^N x_i \delta_i^{(q)}$, the total of variable x in Frame q. Calibration to frame counts $N^{(q)}$ is thus a special case of the general calibration theory.

But the differing amounts of information for the frames may also have a bearing on the multiplicity adjustments. Suppose that Frame 2 has rich auxiliary information for calibration while Frame 1 has little information. Calibrating the weights $w_i^{(2)}$ before compositing may increase the relative effective sample size from S_2 and thus increase the value of $\tilde{n}^{(2)} \big/ \left(\tilde{n}^{(1)} + \tilde{n}^{(2)}\right)$ that would be used for the ESS estimator.

Haziza and Lesage (2016) argued that a two-step weighting procedure offers several advantages for single-frame surveys with nonresponse. The first step divides the design weight for unit i by its estimated response propensity (often calculated from InfoS information) and the second step calibrates the nonresponse-adjusted weights to population control totals (available from InfoU information). When there is substantial nonresponse, weighting adjustment factors from step 1 are often much higher than those from step 2; if the response propensity model is correct, the weighting adjustments in step 2 converge to 1 as $n \to \infty$. The two-step procedure is thus more robust toward misspecification of the calibration model.

The same considerations apply for multiple-frame surveys. A two-step procedure, where step 1 adjusts the samples separately for nonresponse and step 2 calibrates the combined samples, provides robustness to the calibration model. Suppose that S_1 has full response; S_2 has nonresponse but the response propensities can be predicted perfectly from variable x. Then, performing a separate nonresponse adjustment for each sample in step 1 removes the bias for S_2 so that Assumption (A5) is satisfied. If the data are combined first and then calibrated using (3.4), however, the calibration may change the weights for units in S_1 in order to meet the calibration constraints – introducing bias for the estimates from S_1 while not removing it for estimates from S_2. More research is needed on the ordering of steps for weight adjustments. It may be better to perform two steps of nonresponse adjustments and calibration on each sample separately, then adjust the weights for multiplicity, and then calibrate to population totals (including re-calibrating on the individual frame variables).

One consequence of using an overlap estimator for a multiple-frame survey is that the multiplicity adjustments may introduce more weight variation, with observations belonging to one frame having much larger weights than observations belonging to more than one frame. If, for example, a list frame (Frame 2 in Figure 2.2(a, b)) is disproportionately oversampled, then the sampling weights for observations in domain $\{1\}$, which are sampled only from Frame 1, may be large relative to the weights for the other domains. Wolter, Ganesh, Copeland, Singleton and Khare (2019) suggested using a shrinkage estimator, estimating $Y_{\{1\}}$ by $\kappa \hat{Y}_{\{1\}}^{(1)} + (1 - \kappa)\, N_{\{1\}}\, (\hat{Y}_{\{2\}}^{(2)} + \hat{Y}_{\{1,2\}}) \big/ N^{(2)}$, but the shrinkage may introduce bias – after all, the reason for using a more complicated multiple-frame design instead of just sampling from Frame 2 is to avoid potential bias from omitting domain {1}. A better solution, if feasible, is to address the weight variation when designing the survey, as discussed in Section 5.

3.4 Probability sample combined with census of a population subset

Lohr (2014) and Kim and Tam (2021) noted that the situation in Figure 2.2(a) includes the special case in which a probability sample S_1 is taken from Frame 1 having full coverage, and the sample S_2 from Frame 2 is a census of domain {1, 2}. The overlap domain is thus defined to be the units in S_2, which may be from administrative records or a convenience sample. Although S_2, considered by itself, may have undercoverage bias, in the multiple-frame setting the bias is eliminated by the presence of a sample from Frame 1. The units in S_2 have $w_i^{(2)} = 1$ and represent themselves alone; they do not represent any units in other parts of the population. When $N^{(2)} / N$ is small, say from a small convenience sample, S_2 will have little effect on dual-frame estimators – almost all of the population is in domain {1}. But when $N^{(2)} / N$ is large, as may occur when Frame 2 consists of administrative records, the availability of those records may improve the precision of $\hat{Y}$ if Assumptions (A1) to (A6) are met.

When S_2 is a census with no measurement error, $\hat{Y}_{\{1,2\}}^{(2)} = Y_{\{1,2\}}$. The estimator in (3.2) is

$$\hat{Y}(\theta) = \hat{Y}_{\{1\}}^{(1)} + \theta \hat{Y}_{\{1,2\}}^{(1)} + (1-\theta) Y_{\{1,2\}}; \tag{3.6}$$

taking $\theta = 0$ uses the known population total from Frame 2 and relies on Frame 1 only for estimation of the part of the population not in Frame 2.

Kim and Tam (2021) noted that since $Y_{\{1,2\}}$ is known, it can be used as an InfoU calibration total. They proposed two calibration estimators: a ratio estimator $\hat{Y}_{\text{ratio}} = \hat{Y}^{(1)} Y_{\{1,2\}} \big/ \hat{Y}_{\{1,2\}}^{(1)}$ and a generalized regression calibration estimator. For many designs, however, the ratio estimator will be less efficient than $\hat{Y}(0)$ from (3.6) because

$$V\left(\hat{Y}_{\text{ratio}}\right) \approx V\left(\hat{Y}_{\{1\}}^{(1)}\right) + \left(\frac{Y_{\{1\}}}{Y_{\{1,2\}}}\right)^2 V\left[\hat{Y}_{\{1,2\}}^{(1)}\right] - 2\,\frac{Y_{\{1\}}}{Y_{\{1,2\}}}\,\text{Cov}\left(\hat{Y}_{\{1\}}^{(1)}, \hat{Y}_{\{1,2\}}^{(1)}\right);$$

the ratio adjustment can introduce extra variability from $\hat{Y}_{\{1,2\}}^{(1)}$ that is excluded from $\hat{Y}(0)$.

Calibrating $\hat{Y}(\theta)$ to $Y_{\{1,2\}} = \sum_{i=1}^{N} x_i$, for $x_i = \delta_i^{(2)} y_i$, the generalized regression weights in (3.4) become

$$c_i^{(q)} = \tilde{w}_i^{(q)} \left[1 + \left(Y_{\{1,2\}} - \hat{Y}_{\{1,2\}}(\theta) \right) \left(\sum_{f=1}^{Q} \sum_{k \in S_f} \tilde{w}_k^{(f)} \delta_k^{(2)} y_k^2 \right)^{-1} \delta_i^{(2)} y_i \right], \tag{3.7}$$

resulting in $\hat{Y}_{\text{GR}} = \hat{Y}(0)$ from (3.6). Similarly, calibrating on the vector $\mathbf{x}_i = \left(1, \delta_i^{(2)}, \delta_i^{(2)} y_i\right)^T$ results in $\hat{Y}_{\text{GR}} = \hat{Y}_{\{1\}}^{(1)} N_{\{1\}} \big/ \hat{N}_{\{1\}}^{(1)} + Y_{\{1,2\}}$.

For some designs, the variance can be reduced even further. Montanari (1987, 1998) proposed using the regression coefficient $\boldsymbol{\beta} = \left[V(\hat{\mathbf{X}})\right]^{-1} \text{Cov}\,(\hat{Y}, \hat{\mathbf{X}})$ for calibration, resulting in the estimator

$$\hat{Y}_{\text{opt}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \boldsymbol{\beta}. \tag{3.8}$$

Rao (1994) called (3.8) the optimal regression estimator and showed that $V(\hat{Y}_{\text{opt}}) \le V(\hat{Y}_{\text{GR}})$. For the dual-frame situation considered in this section, with $x_i = \delta_i^{(2)} y_i$,

$$\boldsymbol{\beta} \;=\; \frac{\text{Cov}\left(\hat{Y}^{(1)}, \hat{Y}^{(1)}_{\{1,\,2\}}\right)}{V\left(\hat{Y}^{(1)}_{\{1,\,2\}}\right)} \;=\; 1 + \frac{\text{Cov}\left(\hat{Y}^{(1)}_{\{1\}}, \hat{Y}^{(1)}_{\{1,\,2\}}\right)}{V\left(\hat{Y}^{(1)}_{\{1,\,2\}}\right)}$$

and

$$\begin{aligned} \hat{Y}_{\text{opt}} \;&=\; \hat{Y}^{(1)} + \left(Y_{\{1,\,2\}} - \hat{Y}^{(1)}_{\{1,\,2\}}\right)\left[1 + \frac{\text{Cov}\left(\hat{Y}^{(1)}_{\{1\}}, \hat{Y}^{(1)}_{\{1,\,2\}}\right)}{V\left(\hat{Y}^{(1)}_{\{1,\,2\}}\right)}\right] \\ &=\; \hat{Y}^{(1)}_{\{1\}} + \theta_H \hat{Y}^{(1)}_{\{1,\,2\}} + (1 - \theta_H)\, Y_{\{1,\,2\}}, \end{aligned} \tag{3.9}$$

where $\theta_H = -\,\text{Cov}\left(\hat{Y}^{(1)}_{\{1\}}, \hat{Y}^{(1)}_{\{1,\,2\}}\right) \big/ V\left(\hat{Y}^{(1)}_{\{1,\,2\}}\right)$ is Hartley's optimal value for θ from (3.3).

Although we usually think of the compositing factor θ as being between 0 and 1, θ_H can be outside of this range. For a conceptual example, suppose that Frame 2 is a list of children receiving food assistance at school and the sample from Frame 1 is a cluster sample of households. Then households in which one or more children are receiving food assistance have some household members in domain $\{1, 2\}$ and other members in domain $\{1\}$. If y exhibits high intra-household correlation, then we would expect $\hat{Y}^{(1)}_{\{1\}}$ and $\hat{Y}^{(1)}_{\{1,\,2\}}$ to be positively correlated. In this case, Hartley's optimal estimator results in negative weights for units in domain $\{1, 2\}$ from the probability sample.

Even though $\hat{Y}_{\text{opt}}$ is more efficient for special situations such as the cluster sample described above, it depends in practice on an estimate of the covariance, is optimal only for this particular y variable, and may have negative weights. Negative weights can also occur if one does optimal calibration with auxiliary variable $\left(1, \delta_i^{(2)}, \delta_i^{(2)} y_i\right)$; in fact, that calibration results in the estimator proposed by Fuller and Burmeister (1972). These optimal regression estimators are sensitive to the model assumptions, and in general I do not recommend their use.

When the Frame-2 sample is a census and Assumptions (A1) to (A6) are met, the precision of population estimates depends entirely on the design of S_1. When the samples are not designed to be part of a multiple-frame survey (and sometimes even when they are), it is likely that one or more of the assumptions is violated. Assumptions (A4) and (A6) are particularly suspect when it is desired to combine data from surveys that were not designed with combination in mind. Even if both surveys measure unemployment, they may use different questions so that the unemployment statistics from S_2 measure a different concept than the statistics from S_1. Domain misclassification may also occur. A unit in the census S_2 is known to also be in complete Frame 1, but it may be difficult to tell whether a unit in S_1 is also in the administrative records or convenience sample that serves as S_2. These problems are discussed in the next section.

4. Multiple-frame surveys and data integration

Rao (2021) reviewed a number of data integration methods for combining information from a probability sample S_1, assumed to come from a frame with complete coverage, with information from a nonprobability sample S_2, often a census of part of the population as in Section 3.4. Rao considered two cases for making inferences about y: (1) y is observed in both samples, and (2) auxiliary information $\mathbf{x}$ is observed in both samples but y is observed only in S_2. In this section I examine various data integration methods from the perspective of the multiple-frame paradigm and the assumptions in Section 2.2.

4.1 Small area estimation

Small area estimation can be considered to be a special case of a dual-frame estimation problem in which Assumption (A6) is not met. Here, S_1 is a probability sample from Frame 1 and Frame 2 is often an administrative data source. Both frames are assumed to have complete coverage of the population, but the variable of interest y is measured only in S_1. Auxiliary information $\mathbf{x}$ used to predict y is measured in both samples. Beaumont and Rao (2021) discussed integrating probability and nonprobability samples through the use of the Fay-Herriot (1979) estimator with small area estimation techniques.

A composite small area estimator (Rao and Molina, 2015) of the population mean η_a in area a is of the form

$$\hat{\eta}_a = \theta_a \hat{\eta}_a^{(1)} + (1 - \theta_a)\, \hat{\eta}_a^{(2)},$$

where $\hat{\eta}_a^{(1)}$ is the direct estimator for the sample mean in area a from S_1 (which may have large variance or may not exist), $\hat{\eta}_a^{(2)} = \mathbf{x}_a^T \hat{\boldsymbol{\beta}}$ is a predicted value from a regression model, and θ_a is a compositing factor. For the Fay-Herriot estimator, θ_a depends on the relative precision of the two estimators under an assumed regression model whose parameters are estimated from S_1. For the estimator $\hat{\eta}_a$, the variable y is measured differently in the two frames – predicted values are used for Frame 2 – and different compositing factors are used in different areas.

4.2 Mass imputation and sample matching

Suppose that S_1 is a full-response probability sample from Frame 1, but the variable of interest y is not measured in S_1. However, y is measured in S_2 from Frame 2, and auxiliary variables $\mathbf{x}$ are measured in both samples. Let $\tilde{y}_i$ be the predicted value of y_i from an imputation model, relating y_i to

$\mathbf{x}_i$, that is developed on S_2 and let $\tilde{Y}^{(1)} = \sum_{i \in S_1} w_i^{(1)} \tilde{y}_i$ and $\tilde{Y}_d^{(1)} = \sum_{i \in S_1} w_i^{(1)} \delta_i(d) \tilde{y}_i$ be the estimated population and domain-d totals from S_1 using the imputed values.

Similarly to small area estimation, mass imputation fits into the dual-frame context by relaxing Assumption (A6) of no measurement error. Kim and Rao (2012) and Chipperfield, Chessman and Lim (2012) considered the situation where both frames are complete and S_1 and S_2 are both probability samples. The frames can differ – Frame 1, for example, might be an area frame and Frame 2 might be a population register – but both are assumed to have full coverage. Chipperfield et al. (2012) used a composite estimator

$$\hat{Y}_{\text{imp}} = \theta \tilde{Y}^{(1)} + (1 - \theta) \hat{Y}^{(2)}, \tag{4.1}$$

where the optimal value of the compositing factor θ minimizes the variance (considering both the sampling and imputation variability). Kim and Rao (2012) proposed adding a correction for bias with the estimator

$$\tilde{Y}^{(1)} + \sum_{i \in S_2} w_i^{(2)} (y_i - \tilde{y}_i);$$

this estimator is of the same form as (4.1) with $\theta = 1$ if the estimated parameters in the imputation model are required to satisfy $\sum_{i \in S_2} w_i^{(2)} (y_i - \tilde{y}_i) = 0.$

If the imputation model produces unbiased and accurate predictions for y_i, combining the samples augments the effective sample size for calculating estimates. When both samples are probability samples with full coverage, it is possible to perform model diagnostics on S_2. Chipperfield et al. (2012) suggested several diagnostics, including testing the imputation model on small areas, investigating whether it is possible to predict survey membership from the value of y_i (for S_2) or $\tilde{y}_i$ (for S_1), and studying the sensitivity of the mean squared error to different levels of bias in $\tilde{Y}^{(1)}$. The sensitivity of the diagnostics, however, depends on the quality and size of S_2. If S_2 is small relative to S_1, S_1 may contain subpopulations that are not well represented in S_2 and are poorly fit by the imputation model.

The situation becomes more complicated when Frame 2 is incomplete or when S_2 has selection bias. When domain $\{1\}$ is nonempty as in Figure 2.2(a), then the composite estimator with imputed values becomes

$$\hat{Y}_{\text{imp}} = \tilde{Y}_{\{1\}}^{(1)} + \theta \tilde{Y}_{\{1,2\}}^{(1)} + (1 - \theta) \hat{Y}_{\{1,2\}}^{(2)}. \tag{4.2}$$

The properties of the estimator in (4.2) depend on how well the imputation model predicts the values of y_i in S_1. Several imputation methods have been proposed. With sample matching (Rivers, 2007), $\tilde{y}_i$ for observation i in S_i is set equal to the value of y_i of the observation's nearest neighbor (with respect to the values of $\mathbf{x}$) in S_2. Rivers (2007), considering the situation in which S_2 is a convenience sample, took $\theta = 1$ in (4.2) and used the information in S_2 for the sole purpose of finding the imputed values $\tilde{y}_i$ for S_1. Yang, Kim and Hwang (2021) studied theoretical properties of mass-imputed estimators that employ nearest neighbor methods.

Chen, Li and Wu (2020), building on the work of Lee (2006), Lee and Valliant (2009), and Valliant and Dever (2011) on using propensity score weighting to estimate population characteristics from a nonprobability sample, proposed a “doubly-robust” estimator for the situation where $\mathbf{x}_i$ is measured in both surveys but y_i is measured only in nonprobability sample S_2. Let $R_i^{(2)} = 1$ if population unit i is in S_2 and 0 otherwise. Under strong assumptions that (1) $R_i^{(2)}$ and y_i are independent given covariates $\mathbf{x}_i$, (2) $\pi_i^{(2)} = P\left(R_i^{(2)} = 1\right) > 0$ for all population units i, and (3) $R_i^{(2)}$ and $R_j^{(2)}$ are conditionally independent given $\mathbf{x}$, they estimated $\pi_i^{(2)}$ as a function of $\mathbf{x}_i$, using information in S_1, and proposed the estimator

$$\hat{Y}_{\text{DR}} = \sum_{i \in S_1} w_i^{(1)} \tilde{y}_i + \sum_{i \in S_2} \frac{1}{\hat{\pi}_i^{(2)}} (y_i - \tilde{y}_i),$$

where $\tilde{y}_i$ is an imputation prediction for the unknown values of y in S_1 (developed using the information in S_2). The estimator $\hat{Y}_{\text{DR}}$ is approximately unbiased for Y if either the imputation model or the model predicting $\pi_i^{(2)}$ is correct. If the imputation model is correct, then the first term of $\hat{Y}_{\text{DR}}$ is approximately unbiased for Y and the second term has expected value 0. If the model predicting $\pi_i^{(2)}$ is correct, then $\sum_{i \in S_2} y_i \big/ \hat{\pi}_i^{(2)}$ is approximately unbiased for Y and $E\left[\sum_{i \in S_1} w_i^{(1)} \tilde{y}_i - \sum_{i \in S_2} \tilde{y}_i \big/ \hat{\pi}_i^{(2)}\right] \approx 0.$ If neither model is correct, however, $\hat{Y}_{\text{DR}}$ may have large bias.

Kim and Tam (2021) considered an extension of the situation in Section 3.4 in which y_i is not measured in S_1, or is measured differently than in S_2, and proposed substituting an imputed value $\tilde{y}_i$ for y_i in the estimators from S_1 in (3.6), obtaining the estimator in (4.2) with $\theta = 0$; they calibrated this estimator to the known domain size $N_{\{1,\,2\}}$.

4.3 Imputation and the NSAF

The estimators in Section 4.2 impute a predicted value $\tilde{y}_i$ for the unknown value of y_i in S_1. All have the strong assumption that the imputation model developed on S_2 applies to the units in domain $\{1\}$. As Lu (2014b) noted when studying regression for dual-frame surveys, relationships between $\mathbf{x}$ and y may differ across domains. Thus, an imputation model developed on a sample from an incomplete frame, or on a sample with selection bias, may provide poor predictions for y in other parts of the population. Moreover, without data on y in the part of the population that is imputed, it may not be possible to assess the quality of the predictions.

A dual-frame survey was taken for the NSAF because of concern that characteristics of interest might differ for telephone and nontelephone households. Let $y_i = 1$ if child i is in a household that is below 200 percent of the poverty threshold, and 0 otherwise. Using the full sample from both frames (Urban Institute and Child Trends, 2007) an estimated 42.2 percent of children lived in households below 200 percent of the poverty threshold, with standard error 0.5 percent. The estimated percentage from the RDD sample was 38.6 percent and the estimated percentage from the area sample was 93.4 percent. Children in the nontelephone households, sampled from the area frame, were much more likely to be living in poverty.

Now suppose that the NSAF had not measured poverty and income variables in the area sample, and y_i was imputed using regression relationships developed in the RDD sample. In many surveys, the only information available for developing an imputation model is demographic variables. Fitting a logistic regression model to the RDD sample that predicts y from race (with categories white, black, and other), and assigning each child in the area sample to the category with highest predicted probability, results in an estimate of 30.5 percent of children in the area sample living in poverty – a lower value than in the RDD sample. Adding an indicator variable for living in a single-parent household to the model, the estimated percentage for the area sample goes up to 51.9 percent. Both of these estimates, and estimates calculated using cell-mean imputations, are far below the percentage of 93.4 percent from the real data.

The problem, of course, is that the auxiliary information is not rich enough to provide a good prediction of poverty in the area sample. The key feature of the data, and the reason that Waksberg and his colleagues used a dual-frame survey, is that being without a telephone is highly associated with poverty. That association cannot be estimated from the RDD sample where all households have telephones. It might be possible to develop an imputation model using information from other surveys such as the Current Population Survey, where both telephone and non-telephone households are sampled, but I could not find an imputation model predicting y from non-income variables in the RDD sample that provided good predictions.

The nontelephone households were a small part of the population for the NSAF, but the differences between the multivariate relationships in the telephone and nontelephone households were so great that the imputation only slightly reduced bias. If poverty had not been measured for the nontelephone sample, however, and the published statistics had relied only on the imputations, there would have been no way to detect the bias.

4.4 Domain misclassification

One major challenge for combining data using a multiple-frame approach is identifying the domain membership (or multiplicity) of units in the data sources. This is challenging even for surveys that are designed to make use of multiple frames.

The NSAF was designed as a screening survey where telephone households were excluded from the area sample. All households sampled from Frame 2, the RDD frame, were correctly classified since they were contacted by telephone. The more difficult part was obtaining the correct domain classification for households in the area-frame sample. Initial prescreening questions asked whether the household had any working telephones; those that answered no were transferred to the telephone interviewer who conducted the detailed interview. The telephone interviewer administered another brief screening interview and asked again about telephone service. An additional 7 percent of households were excluded after answering the more detailed questions about telephone ownership. Some had told the in-person interviewer that they did not have a telephone because they thought the interviewer wanted to borrow it. Others had misunderstood the question about telephone ownership – one respondent, answering the prescreening questions in the living room, thought they applied only to telephones in the living room and did not mention the telephone in the bedroom (Cunningham, Shapiro and Brick, 1999). Although the second screening interview may have corrected for misclassification from respondents who mistakenly said they did not have a telephone during the prescreening, there was no remedy for potential misclassification from respondents who responded in prescreening that they had a telephone when in fact they did not. Misclassification in this direction may have been part of the reason the investigators had a smaller sample size of nontelephone households than they had anticipated.

In dual-frame telephone surveys, the domain for Figure 2.2(b) (cell only, landline only, or both) is usually determined by asking the respondent about other available telephones and, sometimes, the relative amount each type of telephone is used. Brick, Flores-Cervantes, Lee and Norman (2011) found that their landline samples and cell samples both had smaller estimated proportions of dual users than expected from statistics collected on telephone ownership in the National Health Interview Survey. They conjectured that this was because of persons who had access to both types of telephones but rarely used one of them.

Domain membership may be unknown or difficult to estimate when combining existing data sources. In some cases, as when administrative lists are combined, it may be possible to link records, or the data files may contain information that indicates whether the unit is in other frames. In others, there may be little or no information available on domain membership. How can one know whether a participant in an opt-in panel survey is also in a frame of Medicare recipients if no questions about Medicare are asked in the survey?

Lohr (2011) found that even a small amount of domain misclassification could create large biases in dual-frame estimators; moreover, calibration to domain counts that were based on misclassifications could worsen the bias. She proposed a method for adjusting for bias due to domain misclassification, assuming that misclassification probabilities P (observation classified in domain $d' \mid$ observation actually in domain d) are known or can be accurately estimated for different population subgroups. Lin, Liu and Stokes (2019) studied a similar method using misclassification probabilities P (observation actually in domain $d \mid$ observation classified in domain d').

It may be possible to use multiple-frame methods when domain membership is unknown if the probability that unit i is in domain d can be estimated from auxiliary information $\mathbf{x}_i$ known for all sampled units. Kim and Tam (2021) proposed substituting an estimator for the unknown domain membership for the situation in Section 3.4 where S_2 is a census of a subset of the population. They set $\tilde{\delta}_i(\{1, 2\}) = 1$ if the predicted probability that unit $i \in S_1$ was in domain $\{1, 2\}$, $\hat{P}\left[\delta_i(\{1, 2\}) = 1 \mid \mathbf{x}_i\right]$, exceeded 1/2, and estimated the population total for domain $\{1\}$ as $\sum_{i \in S_1} w_i^{(1)} \left[1 - \tilde{\delta}_i(\{1, 2\})\right] y_i$.

When domain membership is imputed, the mean squared error depends on the accuracy of the domain imputations as well as design features and nonresponse bias in S_1. More research is needed to establish statistical properties of estimators when domain membership is estimated. It may also be desired to study alternative estimators that use the predicted probabilities directly to estimate the total in domain $\{1\}$ as $\sum_{i \in S_1} w_i^{(1)} \hat{P}\left[\delta_i(\{1, 2\}) = 0 \mid \mathbf{x}_i\right] y_i$.

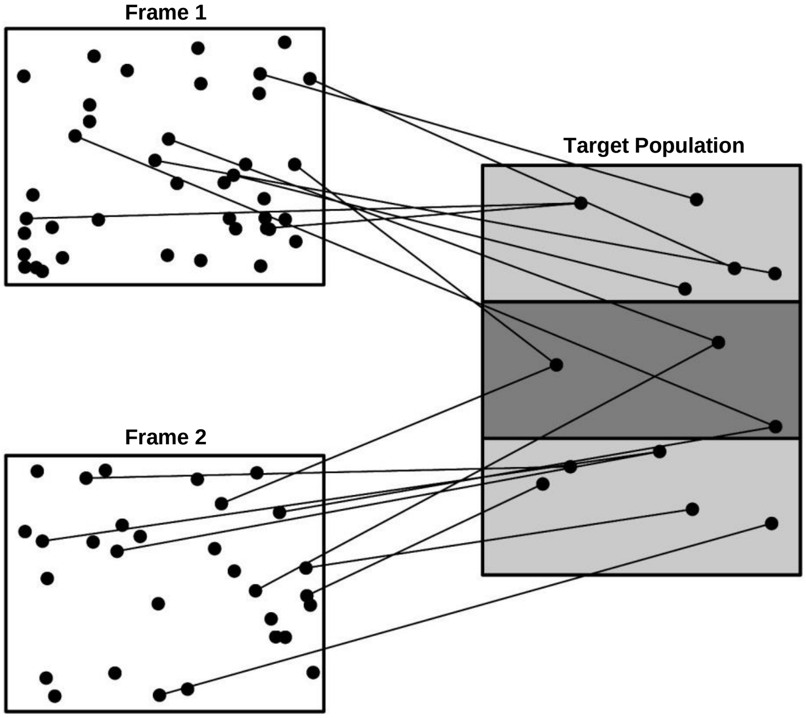
Dever (2018) used sample matching to evaluate the frame overlap for a probability sample S_1, taken from an address-based sampling frame, and a nonprobability sample S_2 recruited from social media sites. She investigated the percentage of respondents in S_1 who had no close match in S_2. Although this procedure does not provide an unbiased estimate of the size of domain {1}, a large percentage of unmatched cases for large samples can indicate that S_2 represents a different population than S_1.

4.5 Indirect sampling and capture-recapture estimation

Sections 4.2 to 4.4 looked at extensions of multiple-frame estimators that relaxed Assumptions (A2), (A4), and (A6). All of these, though, assumed that at least one of the frames, or their union, had full coverage. Let's now look at an example where Assumption (A1) of full coverage is relaxed, and the multiple frames are used to estimate the population size.

In indirect sampling, the target population consists of units that are linked to units in the sampling frame but are not necessarily in the frame (Lavallée, 2007) – units in the target population are sampled indirectly through the links to the sampling units in the frame. Lavallée and Rivest (2012) extended the idea to multiple-frame sampling. As an example, suppose the target population consists of home care workers, who provide paid care for elderly, ill, or disabled persons in their homes. Frame 1 might be a list of persons receiving Medicare benefits, and Frame 2 might be a list of home health care aides from employment or licensing agencies. Persons in the Frame-1 sample are asked to identify workers who provide them with home care, who are then interviewed. A sample of workers from Frame 2 is also interviewed. The home care workers identified from the Frame-1 sample may have links to multiple persons in Frame 1 and may also be in Frame 2. Similarly, persons in the Frame-2 sample may also have links to units in Frame 1. An example of linkage structure is shown in Figure 4.1.

Figure 4.1 Indirect sampling with two frames linked to the target population. Units in the dark shaded area have links to both frames.



With indirect sampling, the Q frames can contain different types of units (the situation with different types of units was also considered by Hartley, 1974). We are not interested in overlap of the sampling frames (shown as nonoverlapping in Figure 4.1 because they contain different types of units) but in the overlap for the units in the target population. Sampled units in the target population have multiple chances of selection if they are linked to multiple units in one or both sampling frames.

Let $l_{j,k}^{(q)} = 1$ if unit j from Frame q is linked to unit k in the target population, and let $L_k^{(q)}$ be the total number of links between unit k in the target population and Frame q (assumed to be knowable from asking unit k). Then an estimator $\hat{Y}^{(q)}$ can be found for each frame using the links as

$$\hat{Y}^{(q)} = \sum_{j \in S^{(q)}} w_j^{(q)} \sum_k \frac{l_{j,k}^{(q)}}{L_k^{(q)}} y_k = \sum_k u_k^{(q)} y_k,$$

where

$$u_k^{(q)} = \sum_{j \in S^{(q)}} w_j^{(q)} \frac{l_{j,k}^{(q)}}{L_k^{(q)}}.$$

In the context of our example, person j in S_1 would say they receive paid home care from provider k, resulting in $l_{j,k}^{(1)} = 1.$ Then the linked home care provider would be asked about how many other persons they work for who receive Medicare (assume they would know this or it could be determined from other sources), giving the value $L_k^{(1)}.$ The quantity $u_k^{(q)}$ sums the weights of the units in S_q with links to unit $k,$ adjusting for the multiplicity of the links to that frame. If $w_j^{(q)} = 1 \big/ \pi_j^{(q)},$ then

$$E[u_k^{(q)}] = E\left[\sum_{j \in S^{(q)}} w_j^{(q)} \frac{l_{j,k}^{(q)}}{L_k^{(q)}} \right] = a_k^{(q)},$$

where $a_k^{(q)} = 1$ if target population member k is linked to at least one unit in Frame q and 0 otherwise.

Multiple-frame methods may then be used to estimate characteristics of the population of home care providers, assuming that unit k linked from S_q provides accurate information on (1) the number of links to members of Frame $q\,(L_k^{(q)}),$ needed for multiplicity adjustments with Frame $q,$ and (2) whether they are also linked to the other frame(s) $(a_k^{(f)}$ for $f \neq q),$ needed to adjust for the multiplicity of linkage from different frames.

Lavallée and Rivest (2012) noted that if the union of the two frames has incomplete coverage – Assumption (A1) is violated – the samples from the two frames can be used to estimate the size of the target population. Let $\hat{T}^{(q)} = \sum_k u_k^{(q)}$ for $k = 1, 2.$ Then $E\big[\hat{T}^{(q)}\big]$ is the number of target population members that can be linked from Frame q. Each sample also provides an estimate of the number of target population units that can be linked from both frames: $\hat{T}_{\{1,2\}}^{(1)} = \sum_k u_k^{(1)} a_k^{(2)}$ and $\hat{T}_{\{1,2\}}^{(2)} = \sum_k u_k^{(2)} a_k^{(1)}.$ These can be composited to obtain an estimator $\hat{T}_{\{1,2\}}$ of the number of persons in the target population who can be captured from both frames.

The Lincoln-Petersen capture-recapture estimator of population size can then be used. Under the strong assumption that being captured by Frame 1 is independent of being captured by Frame 2, the total number of home care providers can be estimated by $\hat{T}^{(1)}\hat{T}^{(2)}\big/\hat{T}_{\{1,2\}}$. In some cases where the independence assumption is not met for the entire population, it may be approximately met on subpopulations whose estimated numbers can be summed. If there are more than two frames, loglinear models may be used to explore associations among the frames (Lohr, 2022, Chapter 14); Zhang (2019) presented a model for the situation in which frames may contain misclassified units.

Alleva, Arbia, Falorsi, Nardelli and Zuliani (2020) proposed using indirect multiple-frame sampling to estimate the number of people infected by SARS-CoV-2 during the early stages of the COVID-19 pandemic in 2020 – information needed for estimating transmissibility and infection parameters in epidemiologic models. In this application, Frame 1 consists of persons with verified infections (perhaps obtained from hospitals, quarantine centers, or clinics), and Frame 2 consists of other persons; the persons in S_2 are administered a test for SARS-CoV-2. The linked sample consists of persons who had contact within the past 14 days with anyone in S_1 or with a member of S_2 who tested positive.

5. Design of data collection systems

Section 4 discussed how estimators for integrated data can be thought of within a multiple-frame survey structure. This structure can also be used when designing data collection systems that make use of multiple sources. Hartley (1962) derived the values of $n^{(1)}$, $n^{(2)}$, and θ that minimize the variance of $\hat{Y}(\theta)$ in (3.2) when S_1 and S_2 are both simple random samples. His basic method can be extended to explore effects of sample design choices for other situations by considering mean squared errors under a range of potential bias assumptions.

There has been a substantial amount of work on optimal design and effects of nonresponse for dual-frame cellular/landline telephone surveys. Brick, Dipko, Presser, Tucker and Yuan (2006) and Brick et al. (2011) investigated nonsampling errors; Lu, Sahr, Iachan, Denker, Duffy and Weston (2013) performed a simulation study to calculate the anticipated mean squared error under various cost models and potential biases. Lohr and Brick (2014), studying allocation of resources in dual-frame telephone surveys with nonresponse, found that for some cost structures a screening survey, in which respondents with landlines are screened out of the cell phone sample, was more cost-efficient than an overlap survey. Levine and Harter (2015) presented graphical results to provide allocation guidance, considering the variance inflation from weight variation. Chen, Stubblefield and Stoner (2021) considered the design problem of oversampling minority populations in dual-frame telephone surveys, using optimal allocation methods from stratified sampling. Most of these articles focus on minimizing the variance of estimates for a fixed cost, and do not consider the effects of potential bias.

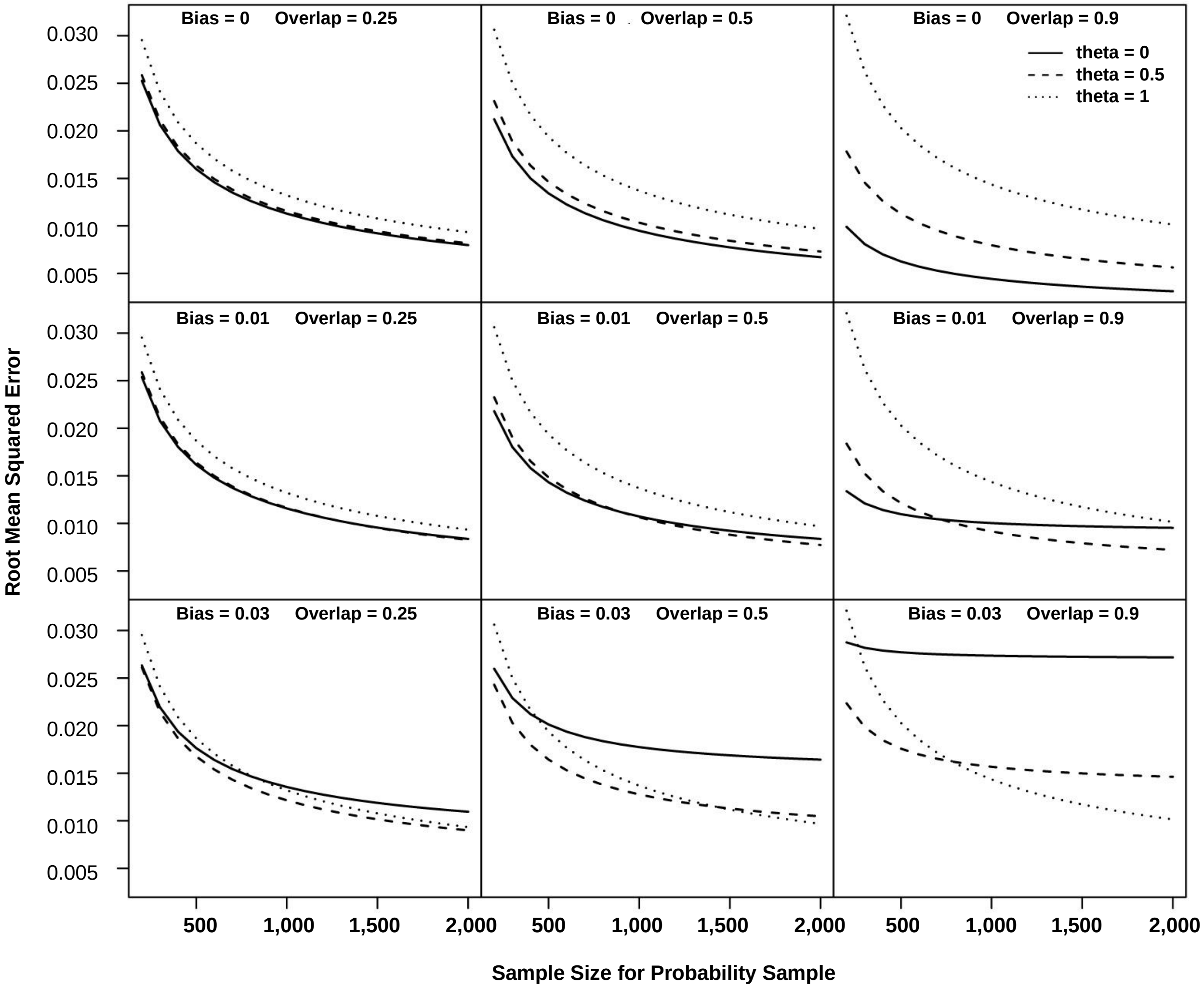
A number of papers in the 1980s studied error structures and designs for dual-frame surveys, typically supplementing a sample from an RDD frame with a sample from an area frame that was assumed to have full coverage. Biemer (1984) and Choudhry (1989) explored optimal designs theoretically and through simulation studies. Groves and Lepkowski (1985, 1986), and Traugott, Groves and Lepkowski (1987) investigated dual-frame designs with a view to minimizing mean squared error when estimates from the RDD frame may be biased. Lepkowski and Groves (1986) found that as the amount of bias increased in the RDD sample, its optimal allocation decreased, reaching an allocation of zero when the bias was 9 percent of the anticipated estimated percentage.

A small amount of bias can have similar effect for the situation considered in Section 3.4, where a census is taken from incomplete Frame 2 and a high-quality probability sample is taken from complete Frame 1. The plots in Figure 5.1 show the root mean squared error (RMSE) for an estimated proportion when S_1 is a simple random sample of size n and S_2 is a census of domain $\{1, 2\},$ for combinations of overlap size $N_{\{1, 2\}} \big/ N$ in $\{0.25, 0.5, 0.9\}$ and bias in $\{0, 0.01, 0.03\}$. The population proportion is 0.2 in domain $\{1\}$ and 0.3 in domain $\{1, 2\},$ and the overall population proportion is estimated using $\hat{Y}(\theta) \big/ N$ for $\hat{Y}(\theta)$ in (3.2). The lines show the RMSE for each n for $\theta = 1$ (S_2 is not used at all), $\theta = 0$ (the estimated proportion in domain $\{1, 2\}$ comes from S_2 and S_1 contributes only for estimating the proportion in domain $\{1\}$), and $\theta = 1/2$. In the bottom row of plots, the bias from S_2 begins dominating the RMSE even for relatively small sample sizes from S_1. A small amount of measurement bias can cancel the supposed advantage from data integration. This example assumes the error in S_2 is from measurement bias, but is similar in spirit to the example in Meng (2018), which shows that even when the selection bias from a convenience sample is small, a simple random sample of size 400 may have more useful information than a convenience sample of size 500 million.

As Thompson (2019) noted, many of the methods that have been developed for combining data from multiple sources have been situation-specific, with solutions tailored to the particular circumstances of that problem. One would not expect these methods to perform as well, on average, for other situations because of regression-to-the-mean effects. Before adopting a data combination method, it may be desirable to perform additional simulation studies that consider outcomes when the model assumptions are not met.

Lohr and Raghunathan (2017) discussed issues for designing data collection systems that leverage multiple data sources, focusing on the situation in which a probability survey is used in conjunction with administrative data sources that cover parts of the population. They considered using administrative data sources for (1) improving the frame for the probability sample, (2) providing contextual information for interpreting the survey data, (3) providing information for nonresponse follow-up and bias assessment, and (4) designing the entire data collection system to take advantage of inexpensive data collection afforded by some of the frames while obtaining complete coverage from the probability survey. Thinking of the design problem in the multiple-frame paradigm can be helpful for the last point. Lohr and Raghunathan (2017) suggested that when Frame 1 is complete but expensive to sample, while Frame 2 is incomplete but less expensive to sample – this includes the situation considered in Section 3.4 of this paper – it may be desirable to use a two-phase screening survey for the sample from Frame 1 and rely on the sample from Frame 2 to supply information for domain $\{1, 2\}$. That is the strategy that Waksberg and colleagues followed for designing the NSAF.

Figure 5.1 Root mean squared error of estimated population proportion under differing amounts of overlap and bias.



When there may be measurement error or domain misclassification, however, a more robust design may be preferred. Optimal designs for dual-frame surveys allocate resources so as to minimize the variance of estimated population totals of interest for fixed cost. Designs that are optimal under Assumptions (A1) to (A6) are not necessarily optimal when some of those assumptions are violated. The multiple-frame structure allows consideration of potential design performance under relaxation of the assumptions.

Hartley (1962) showed that a dual-frame survey resulted in substantial improvements in efficiency for the situations in Figure 2.2(a, b) when data can be inexpensively obtained from Frame 2 and $N_{\{1,2\}} / N$ is large. But when Frame 1 is complete, and the costs are comparable or $N_{\{1,2\}}/N$ is small, the extra complexity from using a dual-frame survey may outweigh its advantages. If, in addition, there is likely to be domain misclassification or if y is measured differently across the surveys, a dual-frame survey will be more complicated than a single sample from Frame 1 and may produce biased estimates.

On the other hand, using multiple data sources can also help assess nonsampling errors. Hartley (1974) wrote that when he presented his work on multiple-frame surveys at a conference, a discussant suggested that a "fairer" comparison would be to compare the variance from a dual-frame sample with that from a single sample of the same cost from the incomplete but cheap frame. Hartley responded (page 107): "The difficulty about this is, of course, that the bias through incompleteness may be of a magnitude which would make the single frame survey useless. If no a priori information on this bias is available, the two frame survey can in fact be regarded as an economical method of *measuring* this bias *and* eliminating it."

Thus, it may be desirable to design the data collection system with multiple goals of (1) obtaining estimates of key population quantities with small mean squared error, (2) assessing nonsampling errors from data sources, and (3) providing information to improve future survey designs. Some of the issues to consider include:

- Quality and stability of data sources. Classical multiple-frame survey design theory assumes that the frames are fixed. But it may be desired to use alternative data sources in which the frame is changing over time (for example, web-scraped prices) to help provide more timely information in coordination with a probability survey. Theory is needed on how to do this. If relying on data supplied by an external source, will those data continue to be available, and in the same form?
- Measurement of domain membership. If possible, information should be collected from each source to allow accurate determination of domain membership. If the information items collected in administrative sources cannot be altered, sometimes items can be added to probability samples that allow domain determination.
- Redundancy. For the situation in Section 3.4, where a census of part of the population is supplemented by a probability sample, a screening design might be optimal for S_1. But a screening design does not allow assessment of potential differences in measurement from the two samples. Some degree of overlap may be desired among the data sources in order to assess differences among the domain estimates from different sources.

  When an imputation model is developed for y based on relationships between y and $\mathbf{x}$ from a data source with incomplete coverage, there is a danger that this model will not apply to the other parts of the population. It may be desired to take a small sample from the uncovered part of the population for purposes of evaluating the model.
- Relative amounts of information for different domains. When data sources include administrative records or large convenience samples, there may be much more information about some parts of the population than others. The issue becomes how to obtain reliable information on the missing parts of the population. When that information comes from a

sample, there may be high weight variation. Levine and Harter (2015) studied the issue of weight variation in dual-frame telephone surveys. Some of the weight variation may be reduced by obtaining additional administrative data sources on underrepresented subpopulations, but there is a danger that, as organizations move away from expensive probability samples, some subpopulations will be omitted from all sources.

- Robustness to design assumptions. Designs that are optimal in theory often turn out to be less so in practice. Exploring the anticipated design performance under violations of the assumptions can be helpful for modifying a theoretically optimal design. In some cases, combining information across sources may result in worse estimates than using a single source, or it may be decided that the gains from combining data are not worth the extra trouble.

  Waksberg (1998) advised: "Do not treat statistical procedures as mechanical operations; be prepared for the unexpected." Having a design with some robustness to the assumptions gives flexibility for unexpected problems.
- Auxiliary information. Many of the methods for integrating data rely on auxiliary information to perform imputations or predict domain membership. Mercer, Lau and Kennedy (2018) argued that for calibration, the richness of the auxiliary information is far more important than the particular method used to calibrate, and the same is true for other data combination methods. Having rich auxiliary information (beyond demographic variables) allows for better data integration models – and for better assessment of their performance.

Waksberg argued that a survey statistician needs to look at the entirety of the problem, not just the optimal design for measuring a single variable. He said that a sampling statistician should "think not only about the specific questions that are asked, but the broader aspects of these questions: whether the questions make sense and can be solved, or whether they should be modified or changed. This is how I've tried to have people with whom I work think about the issues: Here's a question, how do you respond to this specific question? Is it the right question? What statistics will you get by a narrow interpretation of the question, and is there a better way to proceed?" (Morganstein and Marker, 2000, page 304).

In this paper, I have suggested that multiple-frame surveys can serve as an organizing structure for designing and evaluating data-integration systems. This can help clarify the strengths and weaknesses of each source and, perhaps, result in a better way to proceed.

Acknowledgements

I am grateful to the Waksberg Award Committee for selecting me for this honor, to Mike Brick for helpful discussions, and to the associate editor and two referees whose constructive suggestions improved this paper.

References

Aidara, C.A.T. (2019). Quasi random resampling designs for multiple frame surveys. *Statistica*, 79, 321-338.

Alleva, G., Arbia, G., Falorsi, P.D., Nardelli, V. and Zuliani, A. (2020). A sampling approach for the estimation of the critical parameters of the SARS-CoV-2 epidemic: An operational design. https://arxiv.org/ftp/arxiv/papers/2004/2004.06068.pdf, last accessed March 28, 2021.

Arcos, A., Martínez, S., Rueda, M. and Martínez, H. (2017). Distribution function estimates from dual frame context. *Journal of Computational and Applied Mathematics,* 318, 242-252.

Arcos, A., Rueda, M., Trujillo, M. and Molina, D. (2015). Review of estimation methods for landline and cell phone surveys. *Sociological Methods & Research,* 44, 458-485.

Baffour, B., Haynes, M., Western, M., Pennay, D., Misson, S. and Martinez, A. (2016). Weighting strategies for combining data from dual-frame telephone surveys: Emerging evidence from Australia. *Journal of Official Statistics*, 32, 549-578.

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

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

Beaumont, J.-F., and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.

Brick, J.M., Flores-Cervantes, I.F., Lee, S. and Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology*, 37, 1, 1-12. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11443-eng.pdf.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Shapiro, G., Flores-Cervantes, I., Ferraro, D. and Strickler, T. (1999). *1997 NSAF Snapshot Survey Weights*. Washington, DC: Urban Institute.

Burke, J., Mohadjer, L., Green, J., Waksberg, J., Kirsch, I.S. and Kolstad, A. (1994). Composite estimation in national and state surveys. In *Proceedings of the Survey Research Methods Section*, 873-878. Alexandria, VA: American Statistical Association.

Chauvet, G. (2016). Variance estimation for the 2006 French housing survey. *Mathematical Population Studies*, 23, 147-163.

Chauvet, G., and de Marsac, G.T. (2014). Estimation methods on multiple sampling frames in two-stage sampling designs. *Survey Methodology*, 40, 2, 335-346. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14090-eng.pdf.

Chen, S., Stubblefield, A. and Stoner, J.A. (2021). Oversampling of minority populations through dual-frame surveys. *Journal of Survey Statistics and Methodology*, 9, 626-649.

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

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.

Choudhry, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. In *Proceedings of the Survey Research Methods Section*, 566-571. Alexandria, VA: American Statistical Association.

Chu, A., Brick, J.M. and Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute*, 2, 103-104.

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.

Cohen, S.B., DiGaetano, R. and Waksberg, J. (1988). Sample design of the NMES survey of American Indians and Alaska Natives. In *Proceedings of the Survey Research Methods Section*, 740-745. Alexandria, VA: American Statistical Association.

Cunningham, P., Shapiro, G. and Brick, J.M. (1999). *1997 NSAF In-Person Survey Methods*. Washington, DC: Urban Institute.

Dever, J.A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*. https://nces.ed.gov/FCSM/pdf/A4_Dever_2018FCSM.pdf, last accessed July 7, 2021.

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

DiGaetano, R., Judkins, D. and Waksberg, J. (1995). Oversampling minority school children. In *Proceedings of the Survey Research Methods Section*, 503-508. Alexandria, VA: American Statistical Association.

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

Ferraz, C., and Vogel, F. (2015). Multiple frame sampling. In *Handbook on Master Sampling Frames for Agricultural Statistics: Frame Development, Sample Design and Estimation*, 89-106. Rome: Food and Agriculture Organization of the United Nations.

Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, 245-249. Alexandria, VA: American Statistical Association.

Groves, R.M., and Lepkowski, J.M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics,* 1, 263-286.

Groves, R.M., and Lepkowski, J.M. (1986). An experimental implementation of a dual frame telephone sample design. In *Proceedings of the Survey Research Methods Section*, 340-345. Alexandria, VA: American Statistical Association.

Groves, R.M., and Wissoker, D. (1999). *1997 NSAF Early Nonresponse Studies*. Washington, DC: Urban Institute.

Hartley, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, 203-206. Alexandria, VA: American Statistical Association.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36, 99-118.

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

Hendricks, S., Igra, A. and Waksberg, J. (1980). Ethnic stratification in the California Hypertension Survey. In *Proceedings of the Survey Research Methods Section*, 680-685. Alexandria, VA: American Statistical Association.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General),* 149, 65-82.

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. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89, 382-401.

Kott, P.S., and Vogel, F.A. (1995). Multiple-frame business surveys. In *Business Survey Methods,* (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), 185-203. New York: John Wiley & Sons, Inc.

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

Lavallée, P., and Rivest, L.-P. (2012). Capture-recapture sampling and indirect sampling. *Journal of Official Statistics*, 28, 1-27.

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

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.

Lepkowski, J.M., and Groves, R.M. (1986). A mean squared error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.

Levine, B., and Harter, R. (2015). Optimal allocation of cell-phone and landline respondents in dual-frame surveys. *Public Opinion Quarterly*, 79, 91-104.

Lin, D., Liu, Z. and Stokes, L. (2019). A method to correct for frame membership error in dual frame estimators. *Survey Methodology*, 45, 3, 543-565. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00008-eng.pdf.

Lohr, S.L. (2007). Recent developments in multiple frame surveys. In *Proceedings of the Survey Research Methods Section*, 3257-3264. Alexandria, VA: American Statistical Association.

Lohr, S.L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, 37, 2, 197-213. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11608-eng.pdf.

Lohr, S.L. (2014). When should a multiple frame survey be used? *The Survey Statistician*, 69 (January), 17-21.

Lohr, S.L. (2022). *Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: CRC Press.

Lohr, S.L., and Brick, J.M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology*, 2, 388-409.

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

Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, B., Peng, J. and Sahr, T. (2013). Estimation bias of different design and analytical strategies in dual-frame telephone surveys: An empirical evaluation. *Journal of Statistical Computation and Simulation*, 83, 2352-2368.

Lu, B., Sahr, T., Iachan, R., Denker, M., Duffy, T. and Weston, D. (2013). Design and analysis of dual-frame telephone surveys for health policy research. *World Medical & Health Policy*, 5, 217-232.

Lu, Y. (2014a). Chi-squared tests in dual frame surveys. *Survey Methodology*, 40, 2, 323-334. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14096-eng.pdf.

Lu, Y. (2014b). Regression coefficient estimation in dual frame surveys. *Communications in Statistics – Simulation and Computation*, 43, 1675-1684.

Lu, Y., and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36, 1, 13-22. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11248-eng.pdf.

Lu, Y., Fu, Y. and Zhang, G. (2021). Nonparametric regression estimators in dual frame surveys. *Communications in Statistics – Simulation and Computation*, 50, 854-864.

Marks, E., and Waksberg, J. (1966). Evaluation of coverage in the 1960 Census of Population through case-by-case checking. In *Proceedings of the Social Statistics Section*, 62-70. Alexandria, VA: American Statistical Association.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33, 2, 151-157. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10492-eng.pdf.

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

Mercer, A., Lau, A. and Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Washington, DC: Pew Research.

Metcalf, P., and Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.

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

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 1, 69-77. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998001/article/3911-eng.pdf.

Morganstein, D., and Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15, 299-312.

National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine (2018). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: National Academies Press.

O'Muircheartaigh, C., and Pedlow, S. (2002). Combining samples vs. cumulating cases: A comparison of two weighting strategies in NLSY97. In *Proceedings of the Survey Research Methods Section*, 2557-2562. Alexandria, VA: American Statistical Association.

Ranalli, M.G., Arcos, A., Rueda, M.d.M. and Teodoro, A. (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications*, 25, 321-349.

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

Rao, J.N.K. (2021). On making valid inferences by combining data from surveys and other sources. *Sankhyā, Series B*, 83-B, 242-272.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation, 2nd Ed*. Hoboken, NJ: Wiley.

Rivers, D. (2007). Sampling for web surveys. Paper presented at the Joint Statistical Meetings.

Rueda, M.d.M., Arcos, A., Molina, D. and Ranalli, M.G. (2018). Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. *International Statistical Review*, 86, 51-67.

Saegusa, T. (2019). Large sample theory for merged data from multiple sources. *The Annals of Statistics*, 47, 1585-1615.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Hoboken, NJ: Wiley.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *International Statistical Review*, 87, S79-S89.

Traugott, M.W., Groves, R.M. and Lepkowski, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, 51, 522-539.

Urban Institute and Child Trends (2007). *National Survey of America's Families (NSAF), 1997*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

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

Waksberg, J. (1986). Discussion of papers on new approaches in telephone sample design. In *Proceedings of the Survey Research Methods Section*, 367-369. Alexandria, VA: American Statistical Association.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. In *Proceedings of the Survey Research Methods Section*, 497-502. Alexandria, VA: American Statistical Association.

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the US Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 119-135.

Waksberg, J., and Pritzker, L. (1969). Changes in census methods. *Journal of the American Statistical Association*, 64, 1141-1149.

Waksberg, J., Judkins, D. and Massey, J.T. (1997b). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 1, 61-71. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3107-eng.pdf.

Waksberg, J., Brick, J.M., Shapiro, G., Flores-Cervantes, I. and Bell, B. (1997a). Dual-frame RDD and area sample with particular focus on low-income population. In *Proceedings of the Survey Research Methods Section*, 713-718. Alexandria, VA: American Statistical Association.

Waksberg, J., Brick, J.M., Shapiro, G., Flores-Cervantes, I., Bell, B. and Ferraro, D. (1998). Nonresponse and coverage adjustment for a dual-frame survey. *Proceedings: Symposium 97, New Directions in Surveys and Censuses*, 193-198. Ottawa: Statistics Canada.

Wolter, K.M., Ganesh, N., Copeland, K.R., Singleton, J.A. and Khare, M. (2019). Estimation tools for reducing the impact of sampling and nonresponse errors in dual-frame RDD telephone surveys. *Statistics in Medicine*, 38, 4718-4732.

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 Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.

Zhang, L.-C. (2019). Log-linear models of erroneous list data. In *Analysis of Integrated Data* (Eds., L.-C. Zhang and R.L. Chambers), 197-218. Boca Raton, FL: CRC Press.

Zhang, L.-C., and Chambers, R.L. (Eds.) (2019). *Analysis of Integrated Data*. Boca Raton, FL: CRC Press.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 265-277
Statistics Canada, Catalogue No. 12-001-X

Replication variance estimation after sample-based calibration

Jean D. Opsomer and Andreea L. Erciulescu[1]

Abstract

Sample-based calibration occurs when the weights of a survey are calibrated to control totals that are random, instead of representing fixed population-level totals. Control totals may be estimated from different phases of the same survey or from another survey. Under sample-based calibration, valid variance estimation requires that the error contribution due to estimating the control totals be accounted for. We propose a new variance estimation method that directly uses the replicate weights from two surveys, one survey being used to provide control totals for calibration of the other survey weights. No restrictions are set on the nature of the two replication methods and no variance-covariance estimates need to be computed, making the proposed method straightforward to implement in practice. A general description of the method for surveys with two arbitrary replication methods with different numbers of replicates is provided. It is shown that the resulting variance estimator is consistent for the asymptotic variance of the calibrated estimator, when calibration is done using regression estimation or raking. The method is illustrated in a real-world application, in which the demographic composition of two surveys needs to be harmonized to improve the comparability of the survey estimates.

Key Words: Fishing; Hunting and wildlife watching surveys; Raking; Regression estimation; Replicate construction.

1. Introduction

Variance estimation methods for complex surveys include linearization and replication methods. Some of the practical advantages of replication methods include the facts that multiple weight adjustments such as nonresponse adjustments and calibration are readily incorporated into the estimates, that detailed design information does not need to be released in the public-use datasets, and that data users can readily obtain variance estimates for wide classes of estimators without the need for derivations. There are numerous replication methods in use, with the appropriate choice of method dictated by the sampling design and the estimation objectives of the survey. We refer to Wolter (2007) for an overview of the types of variance estimation replication methods.

The problem we are addressing in this article is how to incorporate calibration into replication variance estimation, when the calibration control totals are themselves random and their variance is also estimated by a replication method. This problem occurred because we (the authors) were working with two surveys on the same topic and for the same target population, for which we were tasked with producing a unified set of estimates.

The first survey is the 2016 National Survey of Fishing, Hunting, and Wildlife-Associated Recreation (FHWAR). This survey, conducted by the U.S. Census Bureau, used successive difference replication (SDR), which is a variant of balanced repeated replication (BRR). SDR was originally proposed in Fay

---

[1] Jean D. Opsomer, Westat, Inc., 1600 Research Blvd., Rockville, MD, USA, 20850. E-mail: JeanOpsomer@westat.com; Andreea L. Erciulescu, Westat, Inc., 1600 Research Blvd., Rockville, MD, USA, 20850. E-mail: AndreeaErciulescu@westat.com.

and Train (1995) and is frequently used for Census Bureau surveys. The second survey is the 2016 50-state Survey of FHWAR, conducted by the Rockville Institute, the nonprofit affiliate of Westat. This survey used Delete-A-Group Jackknife (DAGJK) as the replication method (Kott, 2001).

The two 2016 FHWAR surveys were fielded concurrently using different modes of data collection, specifically to allow for comparison between the two and for subsequent reconciliation of the estimates. The National survey used a combination of telephone and in-person data collection and had a sample size sufficient to produce estimates at the census division level. The 50-state survey was a mail-based survey and, as its name implies, had a sample size sufficient to produce estimates at the state level. However, these differences in mode, together with further differences including other survey implementation aspects, subsampling strategies and estimation methods, led to substantial and often statistically significant differences in the estimates, with typically higher estimates in the 50-State Survey than in the National Survey. See Fish and Wildlife Service and Census Bureau (2018) and Rockville Institute (2018) for more details about the two FHWAR surveys.

As noted above, we were responsible for developing a calibration approach to “align” the estimates from the two surveys, in the sense of producing estimates at the state level based on the 50-state survey but compatible with those obtained from the National Survey. This, in turn, would make it possible to compare the 2016 state-level estimates to those from prior iterations of the National survey, which has been conducted since 1955 and with survey results that are directly comparable since 1991. One of the key steps in reconciling the estimates involved calibrating the demographic composition of the 50-state survey to that of the National survey, given that the latter was considered the “gold standard” in this application. To this end, a set of demographic estimates from the National survey were used as control totals for calibration of the 50-State survey. Because these control totals are themselves estimates, however, it was necessary to make sure that their variability is reflected in the variance estimates of the calibrated 50-State Survey estimates. This is an application of *sample-based calibration* (calibrating to random control totals). Sample-based calibration is typically seen in multi-phase surveys, in which the samples and the estimation methods can be coordinated. In the current setting, the two surveys are independent and have two sets of replicates created using different replication methods.

There is a limited literature on how to account for sample-based calibration in replicate variance estimation. Fuller (1998) developed a replication variance estimator for two-phase samples, in which the phase two estimates are calibrated to phase one control totals. In this approach, the phase two replicates are modified by adjustments derived from the spectral decomposition of the phase one estimated variance-covariance matrix of the control totals. Dever and Valliant (2010) and Dever and Valliant (2016) studied weight calibration to estimated control totals under a scenario where a (benchmark) survey is used to calibrate another (analytic) survey, which is more closely related to our setting. In the latter article, their simulation studies were developed for a generalized regression estimator, and linearization and jackknife replication variance estimation methods were compared. For the jackknife replication, the authors compared the performance of the Fuller (1998) adjustment and two adjustments based on draws from a multivariate normal distribution: one using the full variance-covariance matrix of the control totals, and one using only the diagonal of this matrix. The latter approach had been proposed by Nadimpalli, Judkins and Chu (2004), but no theoretical justification was provided. The method was motivated by considering the asymptotic distribution of the estimated control totals, which is then used to generate "synthetic" versions of these estimates for use as replicate control totals.

In this paper, we describe an approach to modify the replicates of the survey to be calibrated by using the replicates from the control survey directly. We show how this method can be used even when the replication methods and/or the number of replicates differ between the two surveys. Interestingly, Kott (2005) already made a brief mention of an approach that likewise uses the replicates directly, in the special case of both surveys using DAGJK with the same number of replicates. Unlike the methods in Fuller (1998) and Nadimpalli et al. (2004), these approaches do not require explicit calculation of the variance-covariance matrix of the control survey, greatly simplifying implementation in practice. In addition, they use valid calibrated totals, unlike the methods relying on draws from a normal distribution which can result in unstable or even unfeasible calibrated totals.

More generally, methods for harmonizing estimates from two surveys can be viewed as an application of *statistical data integration* (SDI), (Lahiri, 2020), a set of methods used to combine multiple data sources to create improved or new estimates compared to what can be obtained from the separate datasets. While they did not use the term SDI, Lohr and Raghunathan (2017) give an overview of the state-of-the-art tools available to perform most of the commonly encountered SDI activities. In a typical SDI application, the goal is the optimal combination of the information in the multiple data sources, which almost always involves creating an estimator that is different from those that are obtained from the separate sources. Methods to achieve this can be design-based, as in multi-frame estimation (Lohr and Rao, 2006) and composite regression estimation (Merkouris, 2004), or model-based (e.g., Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer, 2007). Sample-based calibration falls in the design-based category, but also aims to reproduce the estimates from one of the data sources exactly.

The remainder of the paper is as follows. The proposed method is developed under the setting of regression estimation in Section 2. Raking is another common calibration method and the one used for the two surveys of interest, so we extend the results to this setting in Section 3. In Section 4 we illustrate both the Fuller (1998) method and the proposed method using data from the two 2016 surveys of FHWAR. Section 5 provides overall conclusions.

2. Sample-based regression calibration

We consider a survey of a population U with sample s, weights w_i, target variables y_i. For a given survey estimator $\hat{\theta}$ constructed using the weights w_i, inference is conducted by replication, implemented through the provision of R sets of replicate weights $w_i^{(r)}, r = 1, \ldots, R,$ and variance estimation formula

$$\hat{V}(\hat{\theta}) = A\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \hat{\theta})^2, \tag{2.1}$$

where the $\hat{\theta}^{(r)}$ are computed in the same manner as $\hat{\theta}$ but replacing w_i by the $w_i^{(r)}$. The constant A depends on the replication method. For simplicity, we focus in what follows on the Horvitz-Thompson estimator of $t_y = \sum_U y_i$, denoted by $\hat{t}_y = \sum_s y_i / \pi_i$. In this case, many replication methods of the form (2.1) lead to a design consistent estimator of $\operatorname{Var}(\hat{t}_y)$. We will refer to this survey as the "primary survey".

We are interested in creating adjusted weights w_i^* that are calibrated to a set of control totals from a secondary survey of U with sample s_C, weights w_{Ci}. An estimator from this survey is denoted by $\hat{\theta}_C$. For the second survey, a replication-based variance estimator is also provided,

$$\hat{V}_C(\hat{\theta}_C) = A_C\sum_{r=1}^{R_C}(\hat{\theta}_C^{(r)} - \hat{\theta}_C)^2,$$

with replicate weights $w_{Ci}^{(r)}, r = 1, \ldots, R_C$, and replication-specific constant A_C. The control variables will be denoted by $\mathbf{x}_i$, with estimated totals $\hat{\mathbf{t}}_{Cx}$. Using regression estimation as a framework for calibration, the adjusted estimator is

$$\hat{t}_{y,\,\text{reg}} = \hat{t}_y + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T \hat{\boldsymbol{\beta}} = \sum_s w_i^* y_i \tag{2.2}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{W}_s \mathbf{Y}_s$ with $\mathbf{X}_s$ a matrix with i^{th} row equal to $\mathbf{x}_i^T$, $\mathbf{W}_s$ a diagonal matrix with i^{th} entry w_i and $\mathbf{Y}_s$ a vector containing the $y_i, i \in s$. Hence, the calibrated weights can be written as

$$w_i^* = w_i\left(1 + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T (\mathbf{X}_s^T \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{x}_i\right). \tag{2.3}$$

We note that post-stratification is a special case of regression estimation, see Särndal, Swensson and Wretman (1992, Chapter 7.6).

To obtain a variance estimator, we follow the traditional linearization approach for regression estimators with respect to the sampling design (see e.g., Särndal et al., 1992, Chapter 5.5). Under mild regularity conditions (such as design consistency of Horvitz-Thompson estimators and invertibility of required matrices), the linearized version of the regression estimator (2.2) is equal to the difference estimator,

$$\hat{t}_{y,\,\text{diff}} = \hat{t}_y + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x)^T \boldsymbol{\beta}_N = \hat{\mathbf{t}}_{Cx}^T \boldsymbol{\beta}_N + \hat{t}_e \tag{2.4}$$

where $\boldsymbol{\beta}_N = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$ is the population target of $\hat{\boldsymbol{\beta}}$ and $\hat{t}_e = \sum_s w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_N)$. The variance of $\hat{t}_{y,\,\text{diff}}$ is equal to

$$\operatorname{Var}(\hat{t}_{y,\text{diff}}) = \operatorname{Var}(\hat{t}_e) + \boldsymbol{\beta}_N^T \operatorname{Var}(\hat{\mathbf{t}}_{Cx})\, \boldsymbol{\beta}_N, \tag{2.5}$$

since the two surveys are independent. This "linearized variance" is the variance of the asymptotic distribution of the regression estimator (2.2). In expression (2.5), the first variance term can be estimated using the replicates from the primary survey and the variance-covariance of the control totals in the second term can be estimated using the replicates from the secondary survey. Hence, the plug-in variance estimator

$$\tilde{V}(\hat{t}_{y,\mathrm{reg}}) = V(\hat{t}_{\hat{e}}) + \hat{\boldsymbol{\beta}}^T \hat{V}_C(\hat{\mathbf{t}}_{Cx})\,\hat{\boldsymbol{\beta}}, \tag{2.6}$$

where $\hat{t}_{\hat{e}} = \sum_s w_i(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}),$ can be used for asymptotically valid inference for $\hat{t}_{y,\mathrm{reg}}.$

However, it is often not practical to maintain the two datasets and associated sets of replicates for variance estimation purposes. In the context of survey calibration, the organization in charge of creating the adjusted weights for the primary survey would often prefer to continue providing their dataset unchanged except for the new calibrated weights and associated replicate weights, so that data users can perform their analyses using traditional survey tools. Hence, it is of interest to create a single set of replicates for the primary survey that can be used to estimate the variance, while accounting for the fact that the control totals are themselves estimated from a different survey.

We therefore propose to construct new replicates for the primary survey to estimate (2.5). Assume for now that $R_C \le R.$ Starting from the replicate weights $w_i^{(r)}$ for the primary survey variance estimator, a replicate variance estimator of the first term in (2.6) is obtained by using the calibrated replicate weights

$$w_{1i}^{*(r)} = w_i^{(r)} \left(1 + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x^{(r)})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i\right). \tag{2.7}$$

These replicate weights are obtained by repeating the calibration for each of the replicate weights $w_i^{(r)}$ and lead to consistent variance estimation for regression estimators, as discussed for the general case in Fuller (2009, Chapter 4). See also Valliant (1993) for the special case of post-stratification.

The replicate weights $w_{1i}^{*(r)}$ can be further modified to capture the second term in (2.6) as follows:

$$w_i^{*(r)} = w_{1i}^{*(r)} + a_r w_i^{(r)} (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i, \tag{2.8}$$

with the constants a_r to be further defined below. Combining (2.7) and (2.8), the resulting replicate weights are

$$w_i^{*(r)} = w_i^{(r)} \left(1 + (\hat{\mathbf{t}}_{Cx} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx}) - \hat{\mathbf{t}}_x^{(r)})^T (\mathbf{X}_s^T \mathbf{W}_s^{(r)} \mathbf{X}_s)^{-1} \mathbf{x}_i\right). \tag{2.9}$$

These weights are again obtained by applying the same calibration as for the original weights to each of the replicates, but with replicate control totals $\hat{\mathbf{t}}_{Cx}^{*(r)} = \hat{\mathbf{t}}_{Cx} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx}).$ The resulting replicate estimates are

$$\hat{t}_{y,\mathrm{reg}}^{(r)} = \sum_s w_i^{*(r)} y_i = \hat{t}_y^{(r)} + (\hat{\mathbf{t}}_{Cx} - \hat{\mathbf{t}}_x^{(r)})^T \hat{\boldsymbol{\beta}}^{(r)} + a_r (\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}}^{(r)}.$$

The constants a_r are chosen to account for the difference between the primary and control replication methods, in particular between R_C and R and A_C and $A,$ by letting

$$a_r = \begin{cases} \sqrt{\dfrac{A_C}{A}} & r = 1, \ldots, R_C \\ 0 & r = R_C + 1, \ldots, R. \end{cases} \tag{2.10}$$

This implies that for $r > R_C,$ the replicate weights $w_i^{*(r)} = w_{1i}^{*(r)}$ in (2.8), i.e. the unadjusted control totals are used to calibrate the replicate weights. While the a_r are written with the first R_C values non-zero, this is for notational convenience only. The assignment of the replicates from the control survey to those of the primary survey should be randomized, to ensure that estimators and replicate estimators from both surveys remain independent regardless of the replication methods.

Using the replicate weights (2.9) with constants (2.10), the replicate variance estimator (2.1) becomes

$$\hat{V}(\hat{t}_{y,\mathrm{reg}}) = A\sum_{r=1}^{R}(\hat{t}_{y,\mathrm{reg}}^{(r)} - \hat{t}_{y,\mathrm{reg}})^2, \tag{2.11}$$

Ignoring terms of smaller order as well as those with $a_r = 0,$ this is approximately equal to

$$\begin{aligned} \hat{V}(\hat{t}_{y,\mathrm{reg}}) \approx\ & A\sum_{r=1}^{R}(\hat{t}_{\hat{e}}^{(r)} - \hat{t}_{\hat{e}})^2 + \hat{\boldsymbol{\beta}}^T A_C \sum_{r=1}^{R_C}(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}} \\ & + A\sum_{r=1}^{R} a_r (\hat{t}_{\hat{e}}^{(r)} - \hat{t}_{\hat{e}})(\hat{\mathbf{t}}_{Cx}^{(r)} - \hat{\mathbf{t}}_{Cx})^T \hat{\boldsymbol{\beta}}. \end{aligned}$$

The cross-term is likewise of smaller order because of the independence of the two surveys and the fact that $\sum_{r=1}^{R} \hat{t}_{\hat{e}}^{(r)} / R \approx \hat{t}_{\hat{e}}$ and $\sum_{r=1}^{R_C} \hat{\mathbf{t}}_{Cx}^{(r)} / R_C \approx \hat{\mathbf{t}}_{Cx}.$ Hence, the replicate variance estimator (2.11) inherits the design consistency of the original replication methods for both surveys and is design consistent for the linearized variance (2.5).

Finally, we discuss the case when $R_C > R.$ The above approach is readily extended to this case by repeating the R replicates of the primary survey K times, such that $R_C \le KR$ with K the smallest positive integer for which this inequality is satisfied. The resulting replicate variance estimator is of the same form as (2.1) but with R replaced by KR and A is replaced by $A/K.$ Then, the method discussed above applies directly to this new replicate variance estimator for the primary survey. For instance, if $R = 120$ and $R_C = 150,$ each replicate in the primary survey will be repeated $K = 2$ times, leading to 240 replicates for the primary survey of which 150 will be modified.

3. Sample-based raking calibration

In the application to the two 2016 FHWAR surveys, we used raking rather than regression estimation for calibration. As for regression estimation, the goal is to create new raking controls for each of the replicates, so that the replicate variance estimator for the primary survey accounts for the variability of the control totals from the secondary survey. The above results for regression estimation do not apply directly, but we can apply the same reasoning as in Deville and Särndal (1992) to show that they continue to hold for raking. Instead of relying on this general result, we will derive it here explicitly to show how to obtain the adjusted control totals for the replicates.

For simplicity, we describe here the case in which we are controlling for the marginal counts in domains defined by the levels of 2 categorical variables, denoted a and b, having K and L levels, respectively. In the primary survey, the estimated counts in the domains defined by the intersections of the two variables are $\hat{N}_{a_k b_l} = \sum_s w_i \delta_i (a_k, b_l)$, with $\delta_i (a_k, b_l) = 1$ if element i is in the domain defined by the intersection of a_k and b_l and 0 otherwise. The marginal estimated counts are defined analogously, $\hat{N}_{a_k} = \sum_s w_i \delta_i (a_k, \cdot)$ and $\hat{N}_{b_l} = \sum_s w_i \delta_i (\cdot, b_l)$. We write $\boldsymbol{\delta}_i = \left(\delta_i (a_1), \ldots, \delta_i (a_K), \delta_i (b_1), \ldots, \delta_i (b_L)\right)^T$ for the vector of indicators for the marginal domains for element i. The estimated marginal counts in the primary survey are $\hat{\mathbf{N}} = \sum_s w_i \boldsymbol{\delta}_i = (\hat{N}_{a_1}, \ldots, \hat{N}_{a_K}, \hat{N}_{b_1}, \ldots, \hat{N}_{b_L})^T$ and the corresponding control totals from the secondary survey are $\hat{\mathbf{N}}_C = \sum_{s_C} w_{Ci}\, \boldsymbol{\delta}_i = (\hat{N}_{Ca_1}, \ldots, \hat{N}_{Ca_K}, \hat{N}_{Cb_1}, \ldots, \hat{N}_{Cb_L})^T$. Using the classical raking ratio algorithm of Deming and Stephan (1940) until convergence, the raked weights for the primary survey can be written as

$$\begin{aligned} w_i^* &= w_i \exp(\hat{u}_{a_k} + \hat{u}_{b_l}) \quad \text{for} \quad \delta_i (a_k, b_l) = 1 \\ &= w_i \exp(\hat{\mathbf{u}}^T \boldsymbol{\delta}_i) \end{aligned} \tag{3.1}$$

where $\hat{\mathbf{u}} = (u_{a_1}, \ldots, u_{a_K}, u_{b_1}, \ldots, u_{b_L})^T$ is a solution to the system of $K + L$ equations

$$\begin{aligned} \sum_{l=1}^{L} \hat{N}_{a_k b_l} \exp(u_{a_k} + u_{b_l}) &= \hat{N}_{Ca_k} \quad (k = 1, \ldots, K) \\ \sum_{k=1}^{K} \hat{N}_{a_k b_l} \exp(u_{a_k} + u_{b_l}) &= \hat{N}_{Cb_l} \quad (l = 1, \ldots, L). \end{aligned} \tag{3.2}$$

The solution to these equations is not unique, so one of the unknowns can be set to 0 and an equation removed. This does not affect the values of $\exp(u_{a_k} + u_{b_l})$, and we will set $v_{b_L} = 0$ and remove the last equation in what follows.

There is no explicit expression for the solution to (3.2), but it can be approximated by using a linearization argument. Under the usual survey asymptotic framework that ensures design consistency of Horvitz-Thompson estimators, the $\hat{u}_{a_k}$ and $\hat{u}_{b_l}$ converge to 0 as the sample sizes of the two surveys increase, so that expansion around 0 is valid. Doing so for the equations in (3.2), we approximate the reduced set of $K + L - 1$ equations by

$$\begin{aligned} \sum_{l=1}^{L} \hat{N}_{a_k b_l} (1 + u_{a_k} + u_{b_l} + o_p(1)) &= \hat{N}_{Ca_k} \quad (k = 1, \ldots, K) \\ \sum_{k=1}^{K} \hat{N}_{a_k b_l} (1 + u_{a_k} + u_{b_l} + o_p(1)) &= \hat{N}_{Cb_l} \quad (l = 1, \ldots, L - 1), \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \hat{N}_{a_k} u_{a_k} + \sum_{l=1}^{L} \hat{N}_{a_k b_l} u_{b_l} &= (\hat{N}_{Ca_k} - \hat{N}_{a_k})(1+o_p(1)) \quad (k=1,\ldots,K) \\ \hat{N}_{b_l} u_{b_l} + \sum_{k=1}^{K} \hat{N}_{a_k b_l} u_{a_k} &= (\hat{N}_{Cb_l} - \hat{N}_{b_l})(1+o_p(1)) \quad (l=1,\ldots,L-1). \end{aligned} \tag{3.3}$$

Ignoring the smaller order remainders, the solution to this system of linear equations can be written in the form $\hat{\mathbf{u}} = \hat{\mathbf{J}}^{-1}(\hat{\mathbf{N}}_C - \hat{\mathbf{N}}),$ where $\hat{\mathbf{J}}$ is a symmetric $(K+L-1)\times(K+L-1)$ matrix containing estimated domain counts readily obtained from the left-hand side of (3.3). Note that the resulting $\hat{\mathbf{u}}$ is not a linear estimator, because in the linearization we conditioned on $\hat{\mathbf{N}}.$ Finally, plugging $\hat{\mathbf{u}}$ into the expression (3.1) and linearizing again, we obtain

$$w_i^* \approx w_i \left(1 + (\hat{\mathbf{N}}_C - \hat{\mathbf{N}})^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_i\right) \tag{3.4}$$

and the estimator after raking is

$$\hat{t}_{y,\text{rak}} = \sum_s w_i^* y_i \approx \sum_s w_i y_i + (\hat{\mathbf{N}}_C - \hat{\mathbf{N}})^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s. \tag{3.5}$$

Note that the size of the control variable $\boldsymbol{\delta}_i$ and associated estimates is now $K+L-1,$ but we maintain the prior notation for simplicity. In (3.5), $\hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s$ corresponds to $\hat{\boldsymbol{\beta}}$ in the regression estimator (2.2). This asymptotic equivalence between the raking estimator and the regression estimator with the same control totals was established by Deville and Särndal (1992). In particular, they provide sufficient conditions under which the asymptotic variance of the equivalent regression estimator can also be used for inference for the raking estimator. Hence, a variance estimator of the form (2.6) can be constructed, with $\hat{t}_{\hat{e}} = \sum_s w_i (y_i - \boldsymbol{\delta}_i^T \hat{\mathbf{J}}^{-1} \boldsymbol{\delta}_s^T \mathbf{W}_s \mathbf{Y}_s).$

We now consider the construction of replicate weights for the primary survey that estimate the asymptotic variance of the raking estimator. As before, we construct new replicate control totals $\hat{\mathbf{N}}_C + a_r (\hat{\mathbf{N}}_C^{(r)} - \hat{\mathbf{N}}_C)$ using the replicate estimates $\hat{\mathbf{N}}_C^{(r)}$ from the secondary survey. Each of the sets of replicate weights $w_i^{(r)}$ of the primary survey are adjusted by raking to its corresponding set of control totals, to obtain the $w_i^{*(r)}$ and the raked replicate estimates $\hat{t}_{y,\text{rak}}^{(r)}.$ Using the approximation in (3.5) for each replicate, we obtain that the resulting replication variance estimator is consistent for the asymptotic variance of the raking estimator.

4. Application

We return to the calibration problem encountered while bridging the two 2016 FHWAR surveys. For both surveys, the population is defined as individuals of ages 16 and older, living in U.S. households. The main data sources for this application are record-level data files, containing weights and replicate weights for both surveys. Using these datasets, we conducted an initial analysis and identified discrepancies in the demographics, which we adjusted by sample-based raking. Estimated population totals constructed using the record-level data from the National survey were considered as random controls, for the crosstabs of census divisions (nine categories) and each of the following demographic variables:

- residency: two categories corresponding to urban and rural classification,
- age: eight categories corresponding to age ranges 16-17; 18-24; 25-34; 35-44; 45-54; 55-64; 65-74, and 75+,
- sex: two categories corresponding to male and female,
- race-ethnicity: four categories corresponding to Hispanic, non-Hispanic White, non-Hispanic African American, and non-Hispanic all other,
- annual income: nine categories corresponding to income ranges -\$20,000; \$20,000-\$29,999; \$30,000-\$39,999; \$40,000-\$49,999; \$50,000-\$74,999; \$75,000-\$99,999; \$100,000-\$149,999; \$150,000+, and not reported.

For the application in this article, we use the 50-State survey public use file, which does not contain information on income. Therefore, we illustrate the proposed method in a slightly simplified setting here, using the crosstabs of census divisions and residency, age, sex, and race-ethnicity as the raking dimensions. We implemented both the Fuller (1998) method and the proposed calibration method described in Section 2 using the public-use data files available for both surveys. For comparison, we also show the results of calibrating without adjusting the variance estimates for the random controls, referred to below as the "naive" method because it ignores the variability of the controls in the variance estimates. To compare the variance estimation methods for survey variables that are not control variables, we will also show estimates for domains defined by crosstabs of residency and sex.

While the replication methods of the two FHWAR surveys are different, they both use $R = R_C = 160$ replicates. Referring to expression (2.1), the replication constant for the DAGJK method of the 50-State survey is $A = 159/160$ and the corresponding constant for the SDR method of the National survey is $A_C = 4/160,$ both available from their respective survey documentation. Hence, the replication adjustment constants a_r in (2.10) are equal to $2/\sqrt{159}$ for $r = 1, \ldots, R.$

The estimates we will consider are all estimated domain counts, so we define the target variable $y_i = I_{\{i \in U_d\}}$ for a domain of interest $U_d.$ For the 144 domains defined by the raking dimensions, we write the estimated domain counts as $\hat{t}_k = \sum_s w_i I_{\{i \in U_k\}}, k = 1, \ldots, 144.$ Likewise, the control totals are estimated domain counts from the National survey, so the auxiliary variable vector is $\mathbf{x}_i = \mathbf{I}_i,$ a vector of length 144 containing the indicators for inclusion of respondent i in the control domains $U_k, k = 1, \ldots, 144,$ and let $\hat{t}_{C,k} = \sum_{s_C} w_{Ci} I_{\{i \in U_k\}}.$ We denote the vector of control totals as $\hat{\mathbf{t}}_C = (\hat{t}_{C,1}, \ldots, \hat{t}_{C,144})^T$ and the adjusted replicate control totals are $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C + 2/\sqrt{159}(\hat{\mathbf{t}}_C^{(r)} - \hat{\mathbf{t}}_C).$

In order to implement the Fuller (1998) method, we estimated the variance-covariance matrix of the control totals $\hat{V}_C(\hat{\mathbf{t}}_C)$ using the National survey replicate weights. The spectral decomposition of this matrix resulted in a set of 144 eigenvectors $\mathbf{q}_i$ and associated eigenvalues λ_i, for $i = 1, \ldots, 144$. Following Fuller (1998), we obtain a set of 144 vectors $\mathbf{v}_i$ satisfying

$$\hat{V}_C(\hat{t}_C) = \sum_{i=1}^{144} \mathbf{v}_i' \mathbf{v}_i,$$

where $\mathbf{v}_i = \sqrt{\lambda_i}\, \mathbf{q}_i$, for $i = 1, \ldots, 144$. Finally, the adjusted replicate controls are $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C + \frac{\sqrt{160}}{2} \mathbf{v}_r$ for $r = 1, \ldots, 144$ and $\hat{\mathbf{t}}_C^{*(r)} = \hat{\mathbf{t}}_C$ for $r = 145, \ldots, 160$. This points to a drawback of the Fuller (1998) method: while our approach perturbs the control totals of all 160 replicates, this method only perturbs a fraction of them in this case. In addition, 30 of the 144 eigenvalues were nearly zero, 18 of which less than zero due to floating point error. We truncated the 18 negative eigenvalues to zero, and left the rest unchanged. Hence, to the extent that not all replicates contribute to variance estimates for some survey estimates (e.g., domain totals), there is a risk that the sample-based calibration will be imperfectly reflected in the variance estimates. In general, we expect that a larger number of replicates will be perturbed using our approach, since the estimated variance-covariance matrix of the control totals can only be reliably estimated if its dimension is suitably smaller than R_C.

Tables 4.1 and 4.2 show the estimates and standard errors, respectively, for domains defined by residency and sex, before and after calibration. The first four rows contain the results for marginal totals for raking variables, which are exactly calibrated, while the last four are totals that correspond to the intersection of raking dimensions and are therefore not exactly calibrated.

Both surveys are representative of the same target population, but the estimates and associated standard errors differ, reflecting both sampling variability as well as different calibration approaches applied by the two survey organizations. As Table 4.1 confirms, after the 50-state survey is raked to the National survey, the estimated totals for domains defined as exact calibration domains indeed match exactly between both surveys. For the domains defined by the crosstabulation of residency and sex, the raked estimates for the 50-State survey are close but not identical to those of the National survey.

Table 4.1
Population estimates before and after calibration, rounded to the nearest integer, after scaling by 10^3

Domain		**Before Calibration**		**After Calibration**
		50-State	**National**	
Residency:	Urban	203,445	208,695	208,695
	Rural	51,511	45,991	45,991
Sex:	Male	128,276	121,775	121,775
	Female	126,680	132,911	132,911
Rural:	Male	99,547	98,511	98,089
	Female	103,898	110,184	110,607
Urban:	Male	28,729	23,264	23,686
	Female	22,782	22,727	22,305

Table 4.2 shows the standard errors obtained by the two replication methods with adjusted control totals and by the naive method, which does not account for the randomness of the control totals. By construction, the proposed replication-based adjustment method and the Fuller (1998) method lead to identical variance estimates for domains that are used in the calibration. These variance estimates are also equal to those from the control survey in this case. This reflects the fact that the variance component corresponding to the first term in (2.5) is set to 0 for the control totals, while the variance component for the second term is exactly equal to the control survey variance estimate in the case of raking. Because that variance component is ignored in the naive method, the variance estimates are equal to zero. For the estimated totals for domains defined as the crosstabulation of residency and sex, the variance estimates of the two methods are not identical but close (within 8% of each other), reflecting the fact that both are consistent for the asymptotic variance (2.5). The variance estimates under the naive method are smaller than the variance estimates under the other two calibration methods, leading to an obviously incorrect result due to not accounting for the variance in the random control totals. For other variables, the variance is still expected to be underestimated under the naive method, due to the fact that the second term in the asymptotic variance (2.5) is ignored.

Table 4.2
Standard errors of population estimates before and after calibration, rounded to the nearest integer, after scaling by 10^3

Domain		**Before Calibration**		**After Calibration**		
		50-State	**National**	**Naive**	**Fuller**	**Proposed**
Residency:	Urban	1,922	2,664	0	2,664	2,664
	Rural	1,922	2,598	0	2,598	2,598
Sex:	Male	2,117	1,074	0	1,074	1,074
	Female	2,117	1,112	0	1,112	1,112
Rural:	Male	2,118	1,399	853	1,658	1,533
	Female	2,514	1,797	853	1,964	1,970
Urban:	Male	1,595	1,449	853	1,709	1,641
	Female	979	1,271	853	1,470	1,547

5. Conclusions

We have proposed an approach to account for sample-based calibration in the variance estimates. The approach applies to situations in which both the survey being calibrated and the survey providing the calibration controls use replicate variance estimation, as is often the case in large-scale government surveys. The replication methods in each are arbitrary, as long as they are both valid for their specific surveys. We described the approach for the cases of calibration by regression estimation (including post-stratification) and raking, two commonly used methods in practice, and we anticipate it would work similarly for other types, such as the general calibration estimators of Deville and Särndal (1992).

The main alternative to the proposed method is that of Fuller (1998). Relative to that method, an important advantage of our approach is that it does not require computation of the estimated variance-covariance matrix of the control totals, so that it is very straightforward to implement. In the typical application in which the number of control totals is smaller than the number of replicates, another potential advantage of the proposed method is that the perturbations will be applied across a larger fraction of the replicates. This reduces the risk of computing replicate variance estimates that do not fully reflect the variability of the control totals. For instance, this can occur when only a subset of the replicates contributes to the variance estimate of a domain mean. If these replicates are mostly unperturbed, the resulting variance estimate can underestimate the variance. Further investigation of the performance of the proposed method when the number of replicates of the two surveys are different appears warranted.

References

Deming, W., and Stephan, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

Dever, J.A., and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36, 1, 45-56. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11251-eng.pdf.

Dever, J., and Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4, 289-318.

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 Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 154-159.

Fish and Wildlife Service and Census Bureau (2018). *2016 National Survey of Fishing, Hunting, and Wildlife-Related Recreation*. Methodology Report.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

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

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Kott, P.S. (2005). Using the delete-a-group jackknife variance estimator in an economic survey of US farms. In *Proceedings of the 55th Session of the World Statistics Congress*, Sydney. International Statistical Institute.

Lahiri, P. (2020). Preface to special issue on statistical data integration. *Statistics in Transition*, 21, III-VI.

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

Lohr, S., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

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

Nadimpalli, V., Judkins, D. and Chu, A. (2004). Survey calibration to CPS household statistics. In *Proceedings of the Section on Survey Research Method*s, American Statistical Association, Alexandria, VA, 4090-4094.

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

Rockville Institute (2018). *2016 50-State Survey of Fishing, Hunting, and Wildlife-Related Recreation*. Methodology Report.

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

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2 Ed.). New York: Springer-Verlag Inc.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 279-297
Statistics Canada, Catalogue No. 12-001-X

Two local diagnostics to evaluate the efficiency of the empirical best predictor under the Fay-Herriot model

Éric Lesage, Jean-François Beaumont and Cynthia Bocci[1]

Abstract

The Fay-Herriot model is often used to produce small area estimates. These estimates are generally more efficient than standard direct estimates. In order to evaluate the efficiency gains obtained by small area estimation methods, model mean square error estimates are usually produced. However, these estimates do not reflect all the peculiarities of a given domain (or area) because model mean square errors integrate out the local effects. An alternative is to estimate the design mean square error of small area estimators, which is often more attractive from a user point of view. However, it is known that design mean square error estimates can be very unstable, especially for domains with few sampled units. In this paper, we propose two local diagnostics that aim to choose between the empirical best predictor and the direct estimator for a particular domain. We first find an interval for the local effect such that the best predictor is more efficient under the design than the direct estimator. Then, we consider two different approaches to assess whether it is plausible that the local effect falls in this interval. We evaluate our diagnostics using a simulation study. Our preliminary results indicate that our diagnostics are effective for choosing between the empirical best predictor and the direct estimator.

Key Words: Empirical best predictor; Design mean square error; Model mean square error; Local diagnostic; Local effect; Fay-Herriot model.

1. Introduction

Governments need socioeconomic information at increasingly fine levels of detail. National statistical offices are therefore required to produce statistics for sub-populations that were not identified or could not be taken into account when the survey's precision objectives were determined. As a result, the number of sampled units for these sub-populations may be too small to ensure good precision of standard design-based direct estimators such as the Horvitz-Thompson estimator or calibration estimators. This type of sub-population, where the sample size is insufficient, is called a small domain (or small area). To remedy the lack of precision of direct estimators for small domains, indirect estimators, or small area estimators, can be used. These small area estimators usually rely on a model such as the Fay-Herriot model (Fay and Herriot, 1979). The Empirical Best (EB) predictor, also called the Empirical Bayes predictor or EB estimator, is a small area estimator frequently used in practice.

Small area estimation methods use statistical models to leverage information from the survey and from auxiliary data sources. The Fay-Herriot model is a linear model that breaks down the parameter of interest of a domain into two terms: the first term is the effect explained by the model and the second term is the model error that can be interpreted as an unexplained and unknown local effect.

Classical statistical tools, such as graphs of model residuals, can be used to assess the validity of the Fay-Herriot model. However, these tools give little indication of the efficiency of an indirect estimate for a particular domain. The model Mean Square Error (MSE) of an indirect estimator can be viewed as a

---

[1] Éric Lesage, Insee, Regional Directorate, 35, place du Colombier - CS 94439 - 35044 Rennes Cedex. E-mail: eric.lesage@insee.fr; Jean-François Beaumont, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca; Cynthia Bocci, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: cynthia.bocci@statcan.gc.ca.

local quality indicator since it varies across domains. The model MSE accounts for the effect explained by the model, but integrates out the unexplained local effect.

The design MSE is an alternative to the model MSE that does not integrate out the unexplained local effect. However, unbiased design MSE estimates of small area estimators tend to be very unstable, particularly for domains with few sampled units (Rivest and Belmonte, 2000; Rao, Rubin-Bleuer and Estevao, 2018; and Pfeffermann and Ben-Hur, 2019). To circumvent this problem, literature suggests taking an average over several domains of the design MSE (Rao and Molina, 2015; and Pfeffermann and Ben-Hur, 2019) as a quality measure. However, many public statistics users are only concerned with their specific domain and do not buy into an overall quality criterion to assess the efficiency of estimates for their domain of interest. This is especially the case when they are convinced that their domain is very specific and that this specificity is not found in the explanatory term of the model, but rather in the error term, i.e., the unexplained local effect.

To address the problem of the instability of unbiased estimators of the design MSE, Rao, Rubin-Bleuer and Estevao (2018) proposed a composite estimator that they evaluated in a simulation study. Their composite estimator consists of taking a weighted average of a model MSE estimator and a design MSE estimator. They achieve greater stability at the cost of an increase in bias. Pfeffermann and Ben-Hur (2019) also proposed a method for estimating the design MSE of a small area estimator. The method is rather complex and relies mainly on the choice of an appropriate model. It is therefore not entirely design-based. Apart from these attempts to estimate the design MSE, to the best of our knowledge, there is no local diagnostic in the literature that can be used to determine whether small area estimation is preferable to direct estimation for a specific domain.

In this paper, a different approach is proposed to compare the efficiency under the design of the EB and direct estimators. We proceed in two steps. First, we determine the unexplained local effect interval that ensures the design MSE of the Best (B) predictor, also called Bayes predictor or B estimator, is smaller than the design MSE of the direct estimator. The second step is to assess whether it is plausible that the unexplained local effect lies within this interval. To this end, two diagnostics are proposed: one based on the conditional distribution of the unexplained local effect given the direct estimate, and a second based on a hypothesis test on the unexplained local effect carried out with respect to the sampling design. We found that, depending on the magnitude of the standardized model residual and a factor associated with the precision of the direct estimate, it is possible to detect whether the B or EB estimators are likely to have a smaller design MSE than that of the direct estimator.

Section 2 presents the Fay-Herriot model and describes how the best predictor (B estimator) of the population parameter of interest is constructed. In Section 3, the model and design MSEs of the direct estimator and best predictor are derived. Section 4 describes the two proposed diagnostics. Section 5 explains how to estimate the model parameters and obtain the empirical best predictor (EB estimator) and the estimators of the diagnostics. Section 6 presents the results of a simulation study using real auxiliary data. A brief conclusion is provided in Section 7.

2. The Fay-Herriot model and the best predictor

We consider a finite population U of size N and a sample s of size n drawn from U according to a sampling design $p(s)$. The population U is partitioned into m domains that do not overlap. The domains are identified by the subscript i taking values from 1 to m. The population of domain i, with a size of N_i, is denoted as U_i. The sample of domain i is denoted as s_i and its size is n_i. We are interested in estimating m finite population parameters, $\theta_i, i = 1, \ldots, m$, associated with the m domains. The parameter θ_i is usually a total, an average or a ratio for domain i. Auxiliary information is available in the form of vectors, $\mathbf{z}_i$, available for all domains $i = 1, \ldots, m$. The set containing the m auxiliary vectors is denoted by $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1,\ldots,m}$. Furthermore, we denote by Ω, the set of all variables used to make inferences excluding the inclusion indicators in the sample s; Ω includes $\mathcal{Z}$ and $\theta_i, i = 1, \ldots, m$, among others. The design expectation of a random variable, say A, will thus be denoted by $\mathbb{E}(A \mid \Omega)$.

We consider a linking model that breaks down the parameters of interest θ_i as follows:

$$\theta_i = \boldsymbol{\beta}^{\top} \mathbf{z}_i + b_i v_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a vector of model parameters of the same dimension as $\mathbf{z}_i$, b_i are fixed factors that can be used to account for heterosedasticity in the model and v_i are error terms that follow the normal distribution: $v_i \mid \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$, where σ_v^2 is a model parameter. In practice, $b_i = 1$ is a common choice but it may be more natural to choose $b_i = N_i$ when θ_i is a total. The term $\boldsymbol{\beta}^{\top} \mathbf{z}_i$ is the known effect or effect explained by the model of the finite population parameter θ_i, while $b_i v_i$ is the unknown or unexplained effect that is called the unexplained local effect of θ_i or simply the local effect of θ_i.

The direct estimator of θ_i is denoted by $\hat{\theta}_i$. It is usually obtained by assigning a survey weight to each unit of the sample s_i. The survey weight of a unit can simply be the inverse of its probability of selection in the sample s or a calibration weight. The sampling error is defined as:

$$e_i = \hat{\theta}_i - \theta_i. \tag{2.2}$$

In what follows, the direct estimator will be assumed to be design-unbiased, i.e. $\mathbb{E}(\hat{\theta}_i \mid \Omega) = \theta_i$ or $\mathbb{E}(e_i \mid \Omega) = 0$. This assumption is not always satisfied in practice, for example when using calibration weights, but we will make the usual assumption that the bias remains negligible. We will also assume that the direct estimator $\hat{\theta}_i$, and thus the error e_i, follows a normal distribution. As discussed in Rao and Molina (2015, page 77), the normality assumption of the errors e_i is possibly weaker than the normality assumption of the errors v_i because of the effect of the central limit theorem on $\hat{\theta}_i$. Of course, this effect is less pronounced for smaller domains. Under these assumptions, we have: $e_i \mid \Omega \sim \mathcal{N}(0, \psi_i)$, where $\psi_i = \mathbb{V}(\hat{\theta}_i \mid \Omega)$ is the design variance of $\hat{\theta}_i$. The sample size n_i can be very small, which can lead to poor precision of the direct estimator $\hat{\theta}_i$. This problem has been at the origin of small area estimation research.

By combining the model (2.1) and the expression (2.2), we obtain the combined model, also called the Fay-Herriot model:

$$\hat{\theta}_i = \boldsymbol{\beta}^{\top} \mathbf{z}_i + b_i\, v_i + e_i. \tag{2.3}$$

Noting that v_i is fixed under the sampling design, it can easily be shown that $\mathbb{V}(b_i v_i + e_i \mid \mathcal{Z}) = b_i^2 \sigma_v^2 + \tilde{\psi}_i$, where $\tilde{\psi}_i = \mathbb{E}(\psi_i \mid \mathcal{Z})$ is the smoothed variance (see the remark at the end of this section). The standardized error of the combined model is given by:

$$\varepsilon_i = \frac{\hat{\theta}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}{\sqrt{b_i^2\sigma_v^2 + \tilde{\psi}_i}}. \tag{2.4}$$

The direct estimate $\hat{\theta}_i$ provides information about θ_i. Rao and Molina (2015, Chapter 9, pages 271-272) give the conditional distribution of θ_i:

$$\theta_i \mid \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i + \gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i), (1-\gamma_i)b_i^2\sigma_v^2\}, \tag{2.5}$$

where $\gamma_i = \frac{b_i^2\sigma_v^2}{b_i^2\sigma_v^2 + \tilde{\psi}_i}$.

The best predictor of θ_i, conditionally on $\hat{\theta}_i$ (Rao and Molina, 2015), is then given by:

$$\hat{\theta}_i^B = \mathbb{E}(\theta_i \mid \mathcal{Z}, \hat{\theta}_i) = \gamma_i\, \hat{\theta}_i + (1-\gamma_i)\, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i. \tag{2.6}$$

In the remainder of this paper, the best predictor $\hat{\theta}_i^B$ will be called the B estimator.

In Sections 3 and 4, the theory is developed assuming that $\boldsymbol{\beta}$, σ_v^2 and $\tilde{\psi}_i$ are known. In Section 5, the estimation of these three quantities is discussed, which allows us to obtain an empirical version of the best predictor and our diagnostics.

Remark: In the literature on small area estimation, the theory is usually developed under the assumption that ψ_i is fixed. Therefore, it is implicitly assumed that $\tilde{\psi}_i = \psi_i$. When making inferences under the Fay-Herriot model, ψ_i cannot be expected to be fixed. For example, consider the case where θ_i is a proportion in the domain i and a stratified simple random sampling with replacement design is used with strata that coincide with domains. The direct estimator $\hat{\theta}_i$ is simply the sample proportion in the domain i and it is well known that its variance is given by $\psi_i = n_i^{-1}\theta_i(1-\theta_i)$. In this case, it is obvious that ψ_i is random since it depends on θ_i. It is also easy to show that $\tilde{\psi}_i = n_i^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i(1-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i) - b_i^2\sigma_v^2) \neq \psi_i$ unless $v_i = \sigma_v = 0$. In the rest of this paper, the entire theory is developed under the usual assumption that $\tilde{\psi}_i = \psi_i$. In practice, these two variances are unknown and have to be estimated. Section 5 discusses the estimation of $\tilde{\psi}_i$ using a smoothing model. It can easily be shown that if a model-unbiased estimator, $\hat{\tilde{\psi}}_i$, is available, that is $\mathbb{E}(\hat{\tilde{\psi}}_i \mid \mathcal{Z}) = \tilde{\psi}_i$, then this estimator is also model-unbiased for ψ_i, that is $\mathbb{E}(\hat{\tilde{\psi}}_i - \psi_i \mid \mathcal{Z}) = 0$. The reverse is also true: a model-unbiased estimator for ψ_i will also be model-unbiased for $\tilde{\psi}_i$. Therefore, although $\tilde{\psi}_i \neq \psi_i$, both variances can be estimated by the same estimator. This suggests that the assumption $\tilde{\psi}_i = \psi_i$ may not be so critical in practice.

3. The mean square errors of the direct and B estimators

A mean square error criterion is often chosen to assess the efficiency of the B estimator given in equation (2.6). There are two natural possibilities: either consider the design MSE, or consider the model MSE (MSE with respect to the combined model 2.3).

The model MSE of the direct estimator $\hat{\theta}_i$ is:

$$\mathrm{MSE}_m(\hat{\theta}_i) = \mathbb{E}\{(\hat{\theta}_i - \theta_i)^2 \mid \mathcal{Z}\} = \tilde{\psi}_i$$

and the model MSE of the B estimator is:

$$\mathrm{MSE}_m(\hat{\theta}_i^B) = \mathbb{E}\{(\hat{\theta}_i^B - \theta_i)^2 \mid \mathcal{Z}\} = \gamma_i\tilde{\psi}_i.$$

The B estimator is thus always more efficient than the direct estimator with model-based inferences. This property is the result of the actual construction of the B estimator. On the other hand, and this is a legitimate question, is the B estimator always more efficient than the direct estimator under design-based inferences?

The design mean square errors of the direct and B estimators for the domain i are:

$$\begin{aligned}\mathrm{MSE}_p(\hat{\theta}_i) = \mathbb{E}\{(\hat{\theta}_i - \theta_i)^2 \mid \Omega\} &= \psi_i \\ &= \tilde{\psi}_i \end{aligned} \tag{3.1}$$

and

$$\begin{aligned}\mathrm{MSE}_p(\hat{\theta}_i^B) = \mathbb{E}\{(\hat{\theta}_i^B - \theta_i)^2 \mid \Omega\} &= \gamma_i^2 \psi_i + (1-\gamma_i)^2\, b_i^2 v_i^2 \\ &= \gamma_i \tilde{\psi}_i + (1-\gamma_i)^2 b_i^2 (v_i^2 - \sigma_v^2). \end{aligned} \tag{3.2}$$

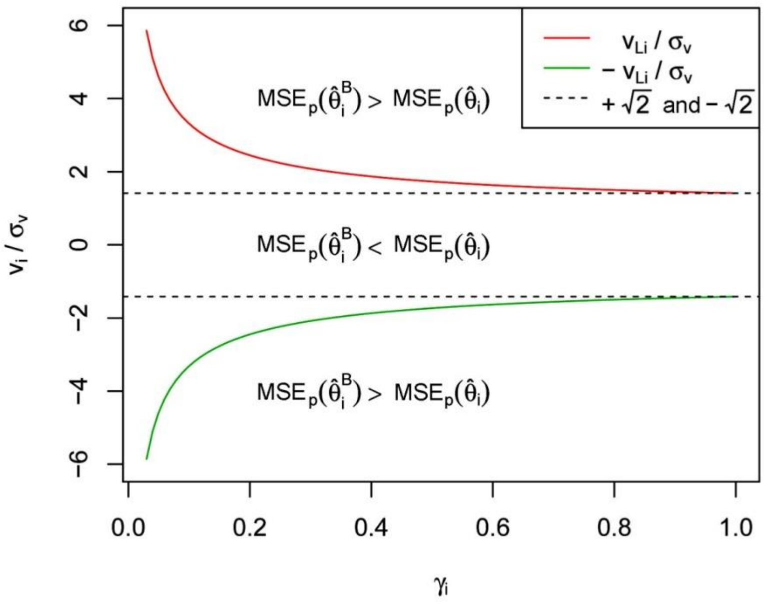
Note that the second equality of (3.1) and (3.2) results from the assumption $\tilde{\psi}_i = \psi_i$. We observe that $\mathrm{MSE}_p(\hat{\theta}_i^B)$ can be very different from $\mathrm{MSE}_m(\hat{\theta}_i^B)$ when the unknown value v_i^2 is far from σ_v^2. Therefore, for a domain with a large value of v_i^2, $\mathrm{MSE}_m(\hat{\theta}_i^B)$ could be significantly smaller than $\mathrm{MSE}_p(\hat{\theta}_i^B)$ and lead to an inaccurate conclusion about the relative efficiency of the direct and B estimators.

By noticing that $\gamma_i \tilde{\psi}_i = (1-\gamma_i) b_i^2 \sigma_v^2$, we can show that $\mathrm{MSE}_p(\hat{\theta}_i^B) \le \mathrm{MSE}_p(\hat{\theta}_i)$ if and only if

$$v_i \in [-v_{L,i}; v_{L,i}],$$

where $v_{L,i} = \sigma_v \sqrt{\frac{1+\gamma_i}{\gamma_i}}$. Figure 3.1 shows the limit values $v_{L,i} / \sigma_v$ and $-v_{L,i} / \sigma_v$ as a function of γ_i. We note that when $|v_i| \le \sigma_v \sqrt{2}$, $\mathrm{MSE}_p(\hat{\theta}_i^B) \le \mathrm{MSE}_p(\hat{\theta}_i)$ for every value of γ_i. We also note that the direct estimator may become more efficient than the B estimator for domains where the local effect is large, especially when γ_i is not small. But how does one know if the local effect is large or not for a given domain i? This is the purpose of the following section where we present two diagnostics.

Figure 3.1 Limit values of the local effect standardized by σ_v versus γ_i.



4. Two diagnostics to evaluate the local performance of the B estimator

4.1 An approach conditional on $\hat{\theta}_i$

From expression (2.5) in Section 2 and noting that $\gamma_i\left(\hat{\theta}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i\right) = b_i\sigma_v\sqrt{\gamma_i}\varepsilon_i,$ we obtain the conditional distribution of v_i:

$$v_i \,\middle|\, \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\left(\sigma_v\sqrt{\gamma_i}\varepsilon_i,\ (1-\gamma_i)\,\sigma_v^2\right).$$

Conditioning on $\hat{\theta}_i$ gives a better idea of the possible values v_i can take. In particular, when the value of γ_i is strictly greater than 0, the conditional distribution of v_i may deviate significantly from its unconditional distribution: $v_i \,|\, \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$.

The first diagnostic is defined as the conditional probability:

$$\begin{aligned} D_{1i} &= \operatorname{Prob}\left(\operatorname{MSE}_p(\hat{\theta}_i^B) \le \operatorname{MSE}_p(\hat{\theta}_i) \,\middle|\, \mathcal{Z}, \hat{\theta}_i\right) \\ &= \operatorname{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \,\middle|\, \mathcal{Z}, \hat{\theta}_i\right). \end{aligned} \tag{4.1}$$

This diagnostic can be written as a function of γ_i and the standardized error (2.4):

$$\begin{aligned} D_{1i} = D_{1i}\left(\gamma_i, |\varepsilon_i|\right) &= \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} \\ &\quad - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i| - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}, \end{aligned} \tag{4.2}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. The proof of result (4.2) is given in Appendix A.

When this diagnostic takes values close to 0, we may conclude that $|v_i|$ is most likely larger than $v_{L,i}$ and that the direct estimator is preferable to the B estimator. To obtain a decision rule associated with this diagnostic, it is necessary to choose a threshold below which we decide to choose the direct estimator and above which the B estimator is chosen. A 50% threshold seems quite natural. Another idea is to apply an empirical approach and identify a break in the distribution of the values of diagnostic D_{1i} for the m domains.

This diagnostic is not entirely design-based because it involves the conditional distribution $v_i \,|\, \mathcal{Z}, \hat{\theta}_i$. It is therefore necessary to validate carefully the Fay-Herriot model before using it. Unfortunately, it is not possible to validate the assumptions on both v_i and e_i because the values of the parameters $\theta_i, i = 1, \ldots, m$ are not observed. However, the combined Fay-Herriot model (2.3) can be validated using model residuals (see, for example, Hidiroglou, Beaumont and Yung, 2019). These residuals are obtained by replacing the unknown quantities in the standardized error (2.4) with their estimates (see Section 5). A graph of residuals versus model predicted values is often suggested to validate the linearity assumption of the model. The normality assumption of the error $b_i v_i + e_i$ can be verified by a Q-Q plot of the residuals or normality tests such as the Shapiro-Wilk test. In case the model is not completely satisfactory, a conservative threshold of 75% may be appropriate.

The diagnostic in the following section is entirely design-based. It is therefore not dependent on the validity of the linking model. In this sense, it is considered more robust than the diagnostic (4.2). However, it relies on assumptions about the sampling errors e_i, discussed in Section 2, including the normality assumption of e_i.

4.2 Use of a design-based hypothesis test on the parameter v_i

In the design-based approach to inference, v_i is fixed and the standardized error (2.4) follows the distribution:

$$\varepsilon_i \mid \Omega \sim \mathcal{N}\left(v_i \frac{\sqrt{\gamma_i}}{\sigma_v},\ (1-\gamma_i)\right). \tag{4.3}$$

We have a unique observation of this random variable. We use it to test if $|v_i|$ is larger than $v_{L,i}$. We consider the test:

$$\mathbf{H_0}: |v_i| = v_{L,i} \quad \text{versus} \quad \mathbf{H_1}: |v_i| > v_{L,i}.$$

We use $|\varepsilon_i|$ as our test statistic. We expect that $|\varepsilon_i|$ will have smaller values under H_0 than under H_1. Let $\varepsilon_{\text{obs},i}$ be the observed value of the statistic ε_i and $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \mid \Omega; v_i\right)$. The p-value of the test is defined as the probability that the statistic $|\varepsilon_i|$ is greater than the observed value $|\varepsilon_{\text{obs},i}|$ under the null hypothesis. Appendix B shows that the p-value is:

$$P_i(v_{L,i}) = P_i(-v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\right),$$

where

$$\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}.$$

Since the second term is often negligible compared to the first term, especially when $\tau_i > 0$ or γ_i is large, our second diagnostic is:

$$D_{2i} = D_{2i}\left(\gamma_i, |\varepsilon_{\text{obs},i}|\right) = \Phi\left(\frac{\sqrt{1+\gamma_i} - |\varepsilon_{\text{obs},i}|}{\sqrt{1-\gamma_i}}\right). \tag{4.4}$$

This second diagnostic can be interpreted as follows: When D_{2i} is small, we can assume that $|v_i|$ is likely to be larger than $v_{L,i}$ and the direct estimator is then preferred to the B estimator. For the choice of a decision threshold, values typically used as levels for hypothesis testing (e.g., 5% or 10%) can be used as a guide. With these small values, the B estimator is favoured. As with the previous diagnostic, the threshold can be determined by locating a break in the distribution of the values of diagnostic D_{2i} for the m domains.

4.3 Some properties of diagnostics 1 and 2

In this section, we study the behaviour of the functions $D_{1i}\left(\gamma_i, |\varepsilon_i|\right)$ and $D_{2i}\left(\gamma_i, |\varepsilon_i|\right)$ for limiting cases of γ_i and $|\varepsilon_i|$ and note their similarities and differences.

Case 1: $0 < \gamma_i < 1$ is fixed and $|\varepsilon_i| \to \infty$.

From equations (4.2) and (4.4) it can be shown that, for $|\varepsilon_i| > 0,$ the two functions $D_{1i}\left(\gamma_i, |\varepsilon_i|\right)$ and $D_{2i}\left(\gamma_i, |\varepsilon_i|\right)$ decrease as $|\varepsilon_i|$ increases. In other words, the derivative of these functions with respect to $|\varepsilon_i|$ is negative. In addition, the limit when $|\varepsilon_i| \to \infty$ of these two functions tends toward 0. For a sufficiently large value of $|\varepsilon_i|$, the two diagnostics will therefore favour the direct estimator.

Case 2: $0 < \gamma_i < 1$ is fixed and $|\varepsilon_i| = 0.$

From equation (4.2), we observe that

$$D_{1i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right) - \Phi\left(-\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right).$$

We can show that $D_{1i}\left(\gamma_i, 0\right)$ is minimized when $\gamma_i = -1 + \sqrt{2}.$ Therefore, $D_{1i}\left(\gamma_i, 0\right) \geq D_{1i}\left(-1+\sqrt{2}, 0\right) =$ 0.98. Since this value is close to 1, diagnostic 1 leads to choosing the B estimator in this case if a threshold of 0.50 or even 0.75 is chosen.

From equation (4.4) we obtain:

$$D_{2i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{1-\gamma_i}}\right).$$

We can show that, for $0 \leq \gamma_i < 1,$ the function $D_{2i}\left(\gamma_i, 0\right)$ is minimized when $\gamma_i = 0.$ Hence, $D_{2i}\left(\gamma_i, 0\right) \geq D_{2i}\left(0, 0\right) =$ 0.84. With a threshold smaller than 0.50, diagnostic 2 leads to the same decision as diagnostic 1 in this case, i.e. to choose the B estimator.

Case 3: $|\varepsilon_i| < \sqrt{2}$ is fixed and $\gamma_i \to 1.$

The two functions $D_{1i}\left(\gamma_i, |\varepsilon_i|\right)$ and $D_{2i}\left(\gamma_i, |\varepsilon_i|\right)$ tend toward 1 in this case. Therefore, diagnostics 1 and 2 lead to choosing the B estimator.

Case 4: $|\varepsilon_i| > \sqrt{2}$ is fixed and $\gamma_i \to 1.$

The two functions $D_{1i}\left(\gamma_i, |\varepsilon_i|\right)$ and $D_{2i}\left(\gamma_i, |\varepsilon_i|\right)$ tend toward 0 in this case. Diagnostics 1 and 2 lead here to choosing the direct estimator.

Case 5: $|\varepsilon_i|$ is fixed and $\gamma_i \to 0.$

The function $D_{1i}\left(\gamma_i, |\varepsilon_i|\right)$ tends toward 1 for any fixed value of $|\varepsilon_i|$. Therefore, diagnostic 1 favours the B estimator for small values of γ_i.

We note that $D_{2i}\left(0, \left| \varepsilon_i \right|\right) = \Phi\left(1 - \left| \varepsilon_i \right|\right)$. Therefore, contrary to Diagnostic 1, Diagnostic 2 will lead to choosing the direct estimator if $\left| \varepsilon_i \right|$ is sufficiently large even when γ_i is infinitely close to 0. For example, with a decision threshold at 0.05 and $\gamma_i = 0$, Diagnostic 2 favours the direct estimator when $\left| \varepsilon_i \right| > 1 - \Phi^{-1}\left(0.05\right) =$ 2.64.

In the first four cases above, both diagnostics lead to the same decision. There is a difference only in Case 5 where $\gamma_i \to 0$. We therefore expect that Diagnostic 2 will choose the direct estimator more often than Diagnostic 1 for small values of γ_i. Consider, for example, a threshold of 0.5 for Diagnostic 1 and of 0.05 for Diagnostic 2. For a threshold of 0.5, we can show that Diagnostic 1 leads to choosing the direct estimator as soon as $\left| \varepsilon_i \right|$ is larger than a value approximately equal to $\frac{\sqrt{1+\gamma_i}}{\gamma_i}$, i.e. as soon as $\left| \varepsilon_i \right| \gtrsim \frac{\sqrt{1+\gamma_i}}{\gamma_i}$. As for Diagnostic 2, for a threshold of 0.05, it leads to choosing the direct estimator as soon as $\left| \varepsilon_i \right| > \sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\,\Phi^{-1}\left(0.05\right)$. For $\gamma_i =$ 0.01, Diagnostic 1 thus leads to choosing the direct estimator when $\left| \varepsilon_i \right| \gtrsim$ 100.5, while Diagnostic 2 leads to choosing the direct estimator when $\left| \varepsilon_i \right| >$ 2.64. The gap narrows as γ_i increases. For example, for $\gamma_i =$ 0.2, Diagnostic 1 chooses the direct estimator when $\left| \varepsilon_i \right| \gtrsim$ 5.48 and Diagnostic 2 chooses the direct estimator when $\left| \varepsilon_i \right| >$ 2.57. The above discussion seems to suggest that Diagnostic 2 leads to choosing the direct estimator more often than Diagnostic 1. However, there are cases where Diagnostic 1 chooses the direct estimator contrary to Diagnostic 2. These cases generally occur for fairly large values of γ_i. For example, for $\gamma_i =$ 0.8, Diagnostic 1 chooses the direct estimator when $\left| \varepsilon_i \right| \gtrsim$ 1.68, while Diagnostic 2 chooses the direct estimator only when $\left| \varepsilon_i \right| >$ 2.08.

5. Empirical version of the B estimator and diagnostics

The theory has been developed assuming the parameters $\boldsymbol{\beta}$, σ_v^2 and $\tilde{\psi}_i$ are known. In practice, these quantities are unknown and the best predictor $\hat{\theta}_i^B$ cannot be used. They can be replaced by estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ and $\hat{\tilde{\psi}}_i$ to obtain the empirical best predictor (EB estimator):

$$\hat{\theta}_i^{\mathrm{EB}} = \hat{\gamma}_i\,\hat{\theta}_i + \left(1 - \hat{\gamma}_i\right)\hat{\boldsymbol{\beta}}^{\top}\mathbf{z}_i,$$

where $\hat{\gamma}_i = \frac{b_i^2 \hat{\sigma}_v^2}{b_i^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}$.

In what follows, we first discuss the estimation of $\boldsymbol{\beta}$ assuming σ_v^2 and $\tilde{\boldsymbol{\psi}} = \left(\tilde{\psi}_1, \ldots, \tilde{\psi}_m\right)^{\top}$ are known. This yields the estimator $\tilde{\boldsymbol{\beta}}\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right)$ of $\boldsymbol{\beta}$. Next, the estimation of σ_v^2 is discussed assuming that $\tilde{\boldsymbol{\psi}}$ is known and we obtain the estimator $\tilde{\sigma}_v^2\left(\tilde{\boldsymbol{\psi}}\right)$ of σ_v^2. Finally, the estimation of the smoothed variances $\tilde{\psi}_i, i = 1, \ldots, m,$ is discussed. We denote the resulting estimators by $\hat{\tilde{\psi}}_i, i = 1, \ldots, m,$ and we let $\hat{\tilde{\boldsymbol{\psi}}} = \left(\hat{\tilde{\psi}}_1, \ldots, \hat{\tilde{\psi}}_m\right)^{\top}$. In practice, the smoothed variances must first be estimated and then successively we compute $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right)$, the estimates of σ_v^2 and $\boldsymbol{\beta}$.

Assuming σ_v^2 and $\tilde{\boldsymbol{\psi}} = \left(\tilde{\psi}_1, \ldots, \tilde{\psi}_m\right)^{\top}$ are known, the estimation of $\boldsymbol{\beta}$ can be done using the generalized least squares method, which is equivalent to the maximum likelihood estimation method under the assumption of independence and normality of the errors $b_i v_i + e_i$. We obtain:

$$\tilde{\boldsymbol{\beta}}\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = \left(\sum_{i=1}^{m} \frac{\mathbf{z}_i \mathbf{z}_i^{\top}}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}\right)^{-1} \sum_{i=1}^{m} \frac{\mathbf{z}_i \hat{\theta}_i}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}.$$

Different methods exist for estimating σ_v^2. For example, the method of moments of Fay and Herriot (1979), the maximum likelihood or restricted maximum likelihood method can be used. The latter is more common in practice. All these methods consist of iteratively solving an estimation equation of the form $g\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = 0,$ where the function g depends on the method. The resulting estimator is denoted by $\tilde{\sigma}_v^2\left(\tilde{\boldsymbol{\psi}}\right)$. Rao and Molina (2015, Chapters 5 and 6) provide more details on the estimation of $\boldsymbol{\beta}$ and σ_v^2 and on the properties of estimators such as model consistency.

Before estimating σ_v^2 and $\boldsymbol{\beta}$ by $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right),$ it is first necessary to estimate the smoothed variance $\tilde{\psi}_i = \mathbb{E}(\psi_i \,|\mathcal{Z}), i = 1, \ldots, m.$ We suppose that a design-unbiased estimator, $\hat{\psi}_i,$ is available, i.e. $\mathbb{E}(\hat{\psi}_i \,|\Omega) = \psi_i$. Under this assumption, we observe that $\mathbb{E}(\hat{\psi}_i \,|\mathcal{Z}) = \tilde{\psi}_i$. The estimator $\hat{\psi}_i$ is therefore unbiased for the smoothed variance $\tilde{\psi}_i$ but can be very unstable when n_i is small. In general, it is preferable to model $\hat{\psi}_i$ given $\mathbf{z}_i$ to increase stability. The following smoothing model is frequently used in practice:

$$\log(\hat{\psi}_i) = \boldsymbol{\alpha}^{\top} \mathbf{x}_i + \eta_i,$$

where $\mathbf{x}_i$ is a function of $\mathbf{z}_i$, $\boldsymbol{\alpha}$ is a vector of model parameters and $\eta_i, i = 1, \ldots, m,$ are independent and identically distributed errors with a mean equal to 0 and a variance equal to σ_η^2. It can easily be shown that

$$\tilde{\psi}_i = \mathbb{E}(\hat{\psi}_i \,|\mathcal{Z}) = \exp(\boldsymbol{\alpha}^{\top} \mathbf{x}_i)\,\Delta,$$

where $\Delta = \mathbb{E}\{\exp(\eta)\}$ and η is a random variable that follows the same distribution as the error term in the above smoothing model. A model-consistent estimator of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, is obtained using the least squares method. Hidiroglou, Beaumont and Yung (2019) suggest estimating Δ by a model-consistent estimator, $\hat{\Delta}$, using a method of moments. The smoothed variance estimator is written as follows:

$$\hat{\tilde{\psi}}_i = \exp(\hat{\boldsymbol{\alpha}}^{\top} \mathbf{x}_i)\,\hat{\Delta},$$

where

$$\hat{\Delta} = \frac{\sum_{i=1}^{m} \hat{\psi}_i}{\sum_{i=1}^{m} \exp(\hat{\boldsymbol{\alpha}}^{\top} \mathbf{x}_i)}.$$

It can be expected that the design MSE of the EB estimator,

$$\text{MSE}_p(\hat{\theta}_i^{\text{EB}}) = \mathbb{E}\left\{(\hat{\theta}_i^{\text{EB}} - \theta_i)^2 \,\middle|\, \Omega\right\},$$

is greater than the design MSE of the B estimator given in equation (3.2). As mentioned above, the estimators of the parameters $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ and Δ are model-consistent, as m increases, provided certain regularity conditions hold. Note also that that the design mean square error of the B estimator (see equation 3.2) does not depend on m. Therefore, the increase in the mean square error resulting from the estimation of these parameters can be expected to be modest when the number of domains is large. This suggests that, if m is large, the derivation of the bound $v_{L,i}$ will be little affected by the estimation of $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ and Δ. Thus, our two diagnostics (4.2) and (4.4) should remain relevant even if the EB estimator is used instead of the B estimator. However, γ_i must be replaced by $\hat{\gamma}_i$ and ε_i by

$$\hat{\varepsilon}_i = \frac{\hat{\theta}_i - \hat{\boldsymbol{\beta}}^{\top} \mathbf{z}_i}{\sqrt{b_i^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

in expressions (4.2) and (4.4) to be able to calculate these diagnostics with real data. As a result, we obtain $\hat{D}_{1i}$, the estimator of D_{1i}, and $\hat{D}_{2i}$, the estimator of D_{2i}.

6. Simulation study

A simulation study was conducted to evaluate the effectiveness of $\hat{D}_{1i}$ and $\hat{D}_{2i}$ in detecting which of the direct and EB estimators is preferable. We considered $m =$ 140 domains representing Canadian cities. In this simulation study, the vector of auxiliary variables is: $\mathbf{z}_i^{\top} = (1, z_{1i})$. The auxiliary variable z_{1i} is obtained from administrative files and is defined as the ratio of the number of employment insurance beneficiaries in city i to the number of people over 15 years of age in city i. The sample size in city i, n_i, was obtained from the Canadian Labour Force Survey (LFS). Of the 140 cities, 2 have a sample size smaller than 10, 10 have a sample size smaller than 30, 40 have a sample size smaller than 60, and 68 have a sample size smaller than 100, representing almost 50% of the cities. For these 68 cities, the estimated coefficients of variation of the LFS unemployment rates are in most cases too large to publish direct estimates of the unemployment rate; as a result, small area estimation techniques are required for these domains. In contrast, there are also 17 of the 140 cities with a sample size larger than 1,000 for which the direct estimate of the unemployment rate is reliable.

The population parameter θ_i was simulated for the m domains using the actual values of n_i and $\mathbf{z}_i$. It can be interpreted as the proportion of unemployed people in city i. The parameter θ_i was generated using the beta distribution with mean $\boldsymbol{\beta}^{\top}\mathbf{z}_i$ and variance σ_v^2, where $\sigma_v^2 = 7.58 \times 10^{-5}$ and $\boldsymbol{\beta}^{\top}$ = (0.0484, 0.95). These values of $\boldsymbol{\beta}$ and σ_v^2 were chosen from real data. We set $b_i = 1, i = 1, \ldots, m$. Then, we manually changed the values of θ_i for four domains (cities) in order to have a local effect v_i equal to $5\sigma_v$. Cities with different sample sizes were chosen: 10, 100, 501 and 3,773. In the rest of this section, the smallest of these four cities is identified by City 1 $(n_i = 10)$, the second smallest by City 2 $(n_i = 100)$, the second largest by City 3 $(n_i = 501)$ and the largest by City 4 ($n_i =$ 3,773).

A stratified simple random sampling with replacement design was considered where strata coincide with domains. The direct estimator $\hat{\theta}_i$ of θ_i is simply the proportion of sampled people in area i who have the characteristic of interest (e.g., being unemployed). Under such a simple design, it is easy to see that the direct estimator can be generated as follows: $\hat{\theta}_i = n_i^{-1}\,\text{Binomial}(n_i, \theta_i)$. It is therefore not necessary to create the population of people in domain i to generate $\hat{\theta}_i$. We proceeded in this way in the simulation. The design variance of $\hat{\theta}_i$ is given by $\psi_i = n_i^{-1}\theta_i\left(1-\theta_i\right)$ and its estimator by $\hat{\psi}_i = \left(n_i - 1\right)^{-1}\hat{\theta}_i\left(1-\hat{\theta}_i\right)$. The smoothed variance $\tilde{\psi}_i$ is estimated using the smoothing model in Section 5 with $\mathbf{x}_i^{\top} = \left(1, \log\left(z_{1i}\right), \log\left(1-z_{1i}\right), \log\left(n_i\right)\right)$.

In order to simulate a realistic scenario, the underlying assumptions of the Fay-Herriot model are not entirely satisfied in our simulation. For example, the errors v_i and e_i do not exactly follow normal distributions. We used a beta distribution to generate θ_i; the normality assumption of v_i is therefore not satisfied although the deviation from the normal distribution is not severe in our simulation. The estimates $\hat{\theta}_i$ were generated from a binomial distribution,which can be approximated by a normal distribution for domains with a large value of n_i. The relationship between the simulated estimates $\hat{\theta}_i$ and the auxiliary vectors $\mathbf{z}_i$ is similar to the one observed with the real LFS estimates. Moreover, our simulation scenario is such that the assumption $\tilde{\psi}_i = \psi_i$ is not satisfied since, for this simple design,

$$\tilde{\psi}_i = n_i^{-1}\left(\boldsymbol{\beta}^{\top}\mathbf{z}_i(1-\boldsymbol{\beta}^{\top}\mathbf{z}_i) - b_i^2\sigma_v^2\right)$$

(see remark in Section 2). However, we note that the correlation coefficient between $\tilde{\psi}_i$ and ψ_i is 0.98, which indicates that the deviation from the assumption $\tilde{\psi}_i = \psi_i$ is modest. As mentioned in the previous paragraph, the smoothing model in Section 5 is used to estimate $\tilde{\psi}_i$. This allows us to remain in a realistic framework where the postulated smoothing model is different from the true model used to generate the estimates $\hat{\psi}_i$.

We conducted a design-based simulation study, i.e., the population parameters $\theta_i, i = 1, \ldots, m,$ were generated only once. We repeated sample selection K = 10,000 times. For each replicate k, $k = 1, \ldots, K$, a direct estimate $\hat{\theta}_i\left(k\right)$ was generated and a smoothed variance estimate $\hat{\tilde{\psi}}_i\left(k\right)$ was calculated as described above. The EB estimate was then calculated as:

$$\hat{\theta}_i^{\text{EB}}\left(k\right) = \hat{\gamma}_i\left(k\right)\hat{\theta}_i\left(k\right) + \left(1-\hat{\gamma}_i\left(k\right)\right)\hat{\boldsymbol{\beta}}\left(k\right)^{\top}\mathbf{z}_i,$$

where $\hat{\gamma}_i\left(k\right) = \frac{\hat{\sigma}_v^2(k)}{\hat{\sigma}_v^2(k) + \hat{\tilde{\psi}}_i(k)}$ and $\hat{\boldsymbol{\beta}}\left(k\right)$ and $\hat{\sigma}_v^2\left(k\right)$ are calculated as described in Section 5. The generalized least squares method was used to obtain $\hat{\boldsymbol{\beta}}\left(k\right)$ and the restricted maximum likelihood method was used to obtain $\hat{\sigma}_v^2\left(k\right)$. Calculations were performed using Statistics Canada's small area estimation system (Hidiroglou, Beaumont and Yung, 2019).

For each replicate, standardized residuals $\hat{\varepsilon}_i\left(k\right)$ and diagnostics $\hat{D}_{1i}\left(k\right)$ and $\hat{D}_{2i}\left(k\right)$ were also calculated for the m domains. We recorded whether the direct estimator was preferred over the EB estimator for each of the two diagnostics. Decision thresholds were used for this purpose. Below the thresholds, the direct estimator is used. For Diagnostic 1, thresholds of 50% and 75% were used and for Diagnostic 2, thresholds of 5% and 25% were used.

From the previous quantities, calculated for each of the 10,000 replicates, the Monte Carlo averages of Diagnostics 1 and 2 were calculated for the m domains: $\bar{\hat{D}}_{1i}$ and $\bar{\hat{D}}_{2i}$. The selection rate of the direct estimator was also calculated for each of the two diagnostics, i.e., the percentage of times a given diagnostic led to the selection of the direct estimator.

The Monte Carlo approximation of $\mathrm{MSE}_p(\hat{\theta}_i^{\mathrm{EB}})$ was calculated as:

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_i^{\mathrm{EB}}(k) - \theta_i\right)^2.$$

From this Monte Carlo MSE, the relative efficiency of the EB estimator was calculated as:

$$\frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) - \psi_i}{\psi_i}. \tag{6.1}$$

This ratio is positive when the EB estimator is less efficient than the direct estimator under the design. A diagnostic is potentially useful if it is negatively correlated with this ratio.

Figures 6.1 and 6.2 present the Monte Carlo averages of Diagnostic 1 and 2 respectively as a function of the relative efficiency of the EB estimator defined in equation (6.1). The four cities whose values of v_i have been changed, Cities 1 to 4, are shown in *purple*, *orange*, *green* and *red*. In the legend, the sample size of these cities has been indicated. The values of the parameter γ_i for Cities 1 to 4 are 0.01, 0.08, 0.35 and 0.81 respectively. All other cities are shown in blue.

First, we can see in Figures 6.1 and 6.2 that the EB estimator is more efficient than the direct estimator for City 1 (in purple) since this city is to the left of the vertical line (negative relative efficiency) despite the strong local effect. The explanation of this phenomenon is shown in Figure 3.1. It shows that the range of values of v_i for which the B estimator is more efficient than the direct estimator increases as γ_i decreases. Since γ_i is small for City 1 $(\gamma_i = 0.01)$, it is not surprising to observe a negative relative efficiency despite a pronounced local effect. For City 2 (in orange), the direct estimator is slightly more efficient than the EB estimator. On the other hand, for Cities 3 (in green) and 4 (in red), the direct estimator is much more efficient than the EB estimator. Note also that there are five cities for which the direct estimator is more efficient than the EB estimator: Cities 2 to 4 as well as two other cities whose values of v_i were randomly generated and not manually modified. One of these cities has the smallest value of v_i and the other has the largest value of v_i after excluding the four cities that had their value manually modified. These two cities have large values of γ_i (0.62 and 0.49).

Figures 6.1 and 6.2 indicate that our two diagnostics seem to be quite effective in detecting cases where the direct estimator is more efficient than the EB estimator except for City 2 $(n_i = 100)$ where the Monte Carlo average of Diagnostic 1 is very high at 0.97. However, this is a domain where choosing the least efficient estimator is not really problematic since there is very little difference between the efficiencies of the two estimators. Apart from this specific case, Diagnostic 1 seems to have better properties than Diagnostic 2. The Monte Carlo average of Diagnostic 1 is very close to 1 when the EB estimator is significantly more efficient than the direct estimator, decreases slowly when the efficiencies of the two estimators approach each other and becomes small when the direct estimator is significantly more efficient than the EB estimator. Not exactly the same behaviour is observed for Diagnostic 2. The Monte Carlo average of Diagnostic 2 is small when the direct estimator is significantly more efficient than the EB estimator but it is not close to 1 when the EB estimator is significantly more efficient than the direct estimator. Furthermore, it seems to increase when the efficiencies of the two estimators come closer, which is counterintuitive.

Figure 6.1 Monte Carlo average of Diagnostic 1 estimates for the 140 cities versus the relative efficiency of the EB estimator.

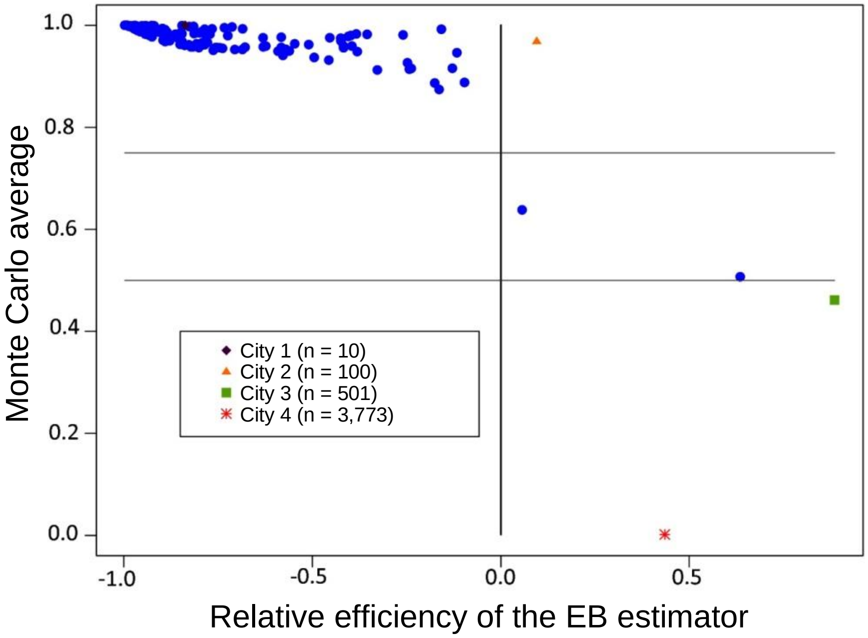
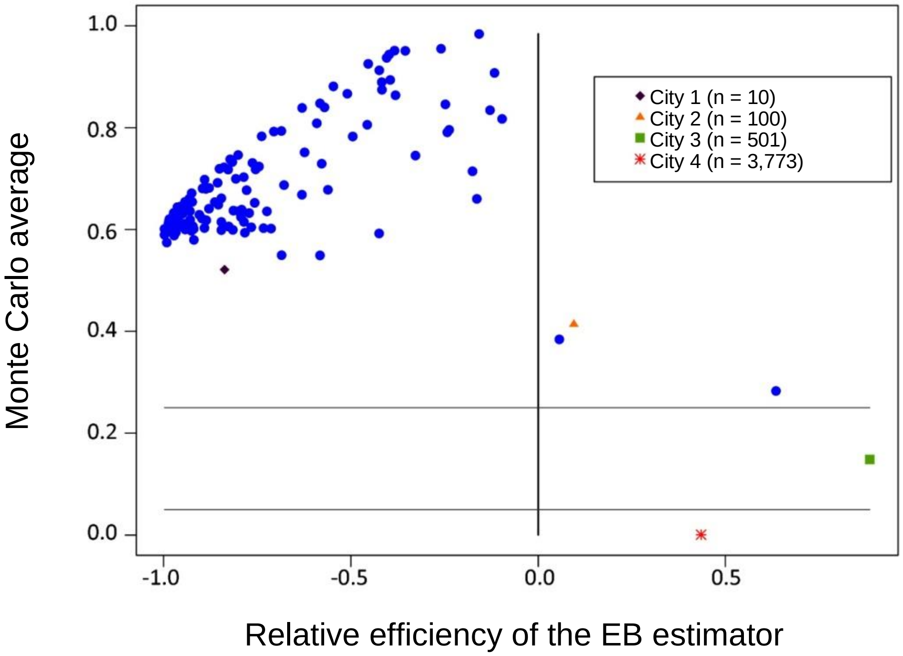


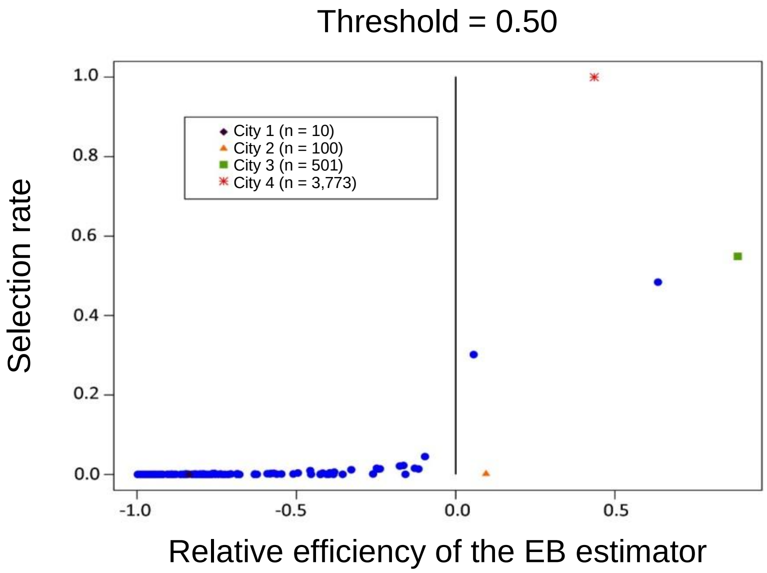
Figure 6.2 Monte Carlo average of Diagnostic 2 estimates for the 140 cities versus the relative efficiency of the EB estimator.



Figures 6.3 and 6.4 show the selection rate of the direct estimator over the 10,000 replicates for Diagnostics 1 and 2. Similar conclusions can be drawn as those obtained by analyzing Figures 6.1 and 6.2.

As expected, the thresholds of 75% for Diagnostic 1 and 25% for Diagnostic 2 allow better detection of cases where the direct estimator is more efficient than the EB estimator, but these thresholds also lead to the direct estimator being chosen a little too often when it was less efficient than the EB estimator. This is particularly notable for Diagnostic 2. This error can be dampened by decreasing the thresholds, but this also reduces the selection rate of the direct estimator when it is more efficient than the EB estimator. As noted earlier, Diagnostic 1 appears to have better properties than Diagnostic 2, regardless of the thresholds chosen, with very small selection rates of the direct estimator when it is significantly less efficient than the EB estimator. This seems to show the limitations of a fully design-based approach, such as the one presented in Section 4.2, to address the challenge of small domain sample sizes.

Figure 6.3 Direct estimator selection rate for Diagnostic 1.


Threshold = 0.50

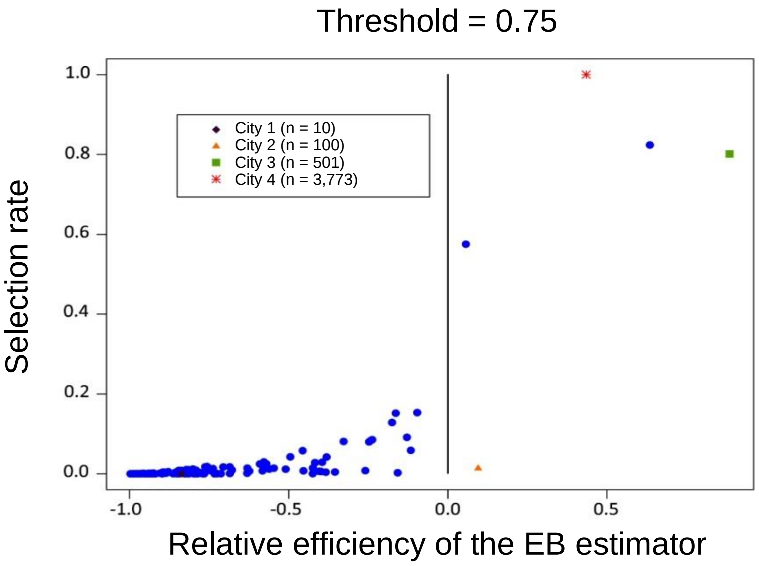
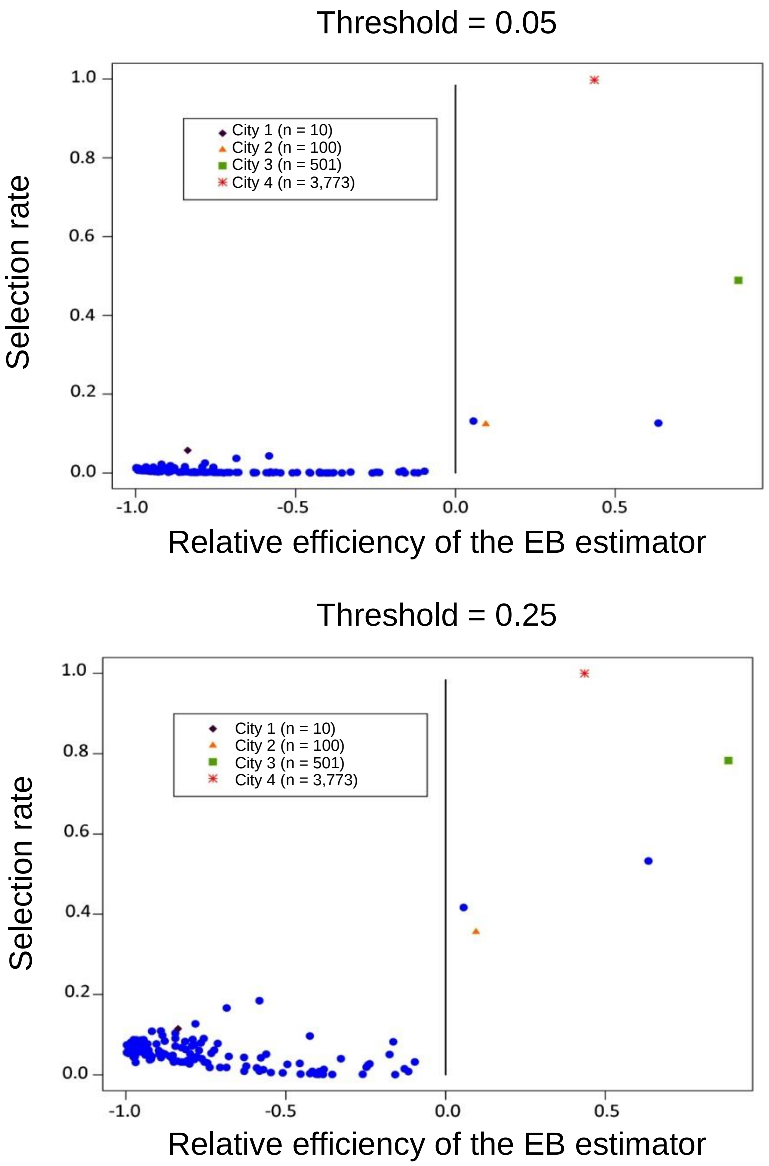

Threshold = 0.75

Figure 6.4 Direct estimator selection rate for Diagnostic 2.



7. Conclusion

Users of small area estimates are usually interested in only one domain. Therefore, they seek a quality indicator that applies to their domain and not an overall indicator. The design MSE of small area estimators is a conceptually attractive quality indicator since it conditions on the unexplained local effect. However, it is known that design-unbiased estimators of the design MSE are generally unstable when the domain sample size is small. To circumvent this problem, we proposed two diagnostics that are intended to identify domains where the design MSE of the direct estimator is smaller than that of the EB estimator. Our simulation results seem promising and allow us to envision the implementation of a useful indicator for choosing between the direct and EB estimators for a particular domain. In future research, it would be interesting to evaluate the efficiency of a hybrid estimator that would leverage these diagnostics.

Appendix

A. Proof of equivalence between equations (4.1) and (4.2)

Using equation (4.1) and the conditional distribution of v_i given in Section 4.1, we have:

$$\begin{aligned} D_{1i} &= \text{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \,\middle|\, Z, \hat{\theta}_i\right) \\ &= \text{Prob}\left(\frac{-v_{L,i} - \sigma_v \sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_{L,i} - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \,\middle|\, Z, \hat{\theta}_i\right). \end{aligned}$$

Replacing $v_{L,i}$ with $\sigma_v \sqrt{\frac{1+\gamma_i}{\gamma_i}}$ results in:

$$\begin{aligned} D_{1i} &= \text{Prob}\left(\frac{-\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \,\middle|\, Z, \hat{\theta}_i\right) \\ &= \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(-\varepsilon_i + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(-\varepsilon_i - \sqrt{\frac{1+\gamma_i}{\gamma_i}}\right)\right\}. \end{aligned}$$

Since for any value $t,$ we have $\Phi(t) = 1 - \Phi(-t)$ then

$$D_{1i} = \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(\varepsilon_i + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(\varepsilon_i - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}.$$

We notice that D_{1i} is a symmetric function of ε_i around 0, i.e. $D_{1i}(\varepsilon_i) = D_{1i}(-\varepsilon_i).$ Therefore, we can rewrite D_{1i} as in equation (4.2):

$$D_{1i} = \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\,\varepsilon_i\,| + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\,\varepsilon_i\,| - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}.$$

B. p-value associated with the test statistic $|\,\varepsilon_i\,|$

First, recall that $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \,\middle|\, \Omega; v_i\right).$ We define the p-value as the maximum of $P_i(v_{L,i})$ and $P_i(-v_{L,i}).$ Since $\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}},$ we can then write:

$$\begin{aligned} P_i(v_i) &= \text{Prob}\left(|\,\varepsilon_i\,| > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\;\tau_i \,\middle|\, \Omega; v_i\right) \\ &= \text{Prob}\left(\varepsilon_i > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\;\tau_i \,\middle|\, \Omega; v_i\right) \\ &\quad + \text{Prob}\left(\varepsilon_i < -\sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\;\tau_i \,\middle|\, \Omega; v_i\right). \end{aligned}$$

Using the standardized error distribution (4.3), we obtain:

$$\begin{aligned}
P_i(v_i) &= \operatorname{Prob}\left(\frac{\varepsilon_i - v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} > \frac{\sqrt{1+\gamma_i} - v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} + \tau_i \,\middle|\, \Omega; v_i \right) \\
&\quad + \operatorname{Prob}\left(\frac{\varepsilon_i - v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} < \frac{-\sqrt{1+\gamma_i} - v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i \,\middle|\, \Omega; v_i \right) \\
&= \Phi\left(\frac{-\sqrt{1+\gamma_i} + v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i \right) \\
&\quad + \Phi\left(\frac{-\sqrt{1+\gamma_i} - v_i \sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i \right).
\end{aligned}$$

Using the expression $v_{L,i} = \sigma_v \sqrt{\frac{1+\gamma_i}{\gamma_i}}$, we have:

$$\begin{aligned}
P_i(v_i) &= \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \left[\frac{v_i}{v_{L,i}} - 1 \right] \right) \\
&\quad + \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \left[-\frac{v_i}{v_{L,i}} - 1 \right] \right).
\end{aligned}$$

Under the null hypothesis H_0, $v_i = v_{L,i}$ or $v_i = -v_{L,i}$ and in both cases the above equation reduces to:

$$P_i(v_{L,i}) = P_i(-v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \right).$$

We will now show that if we reject H_0 (with a threshold smaller than 0.5 such as 0.1) then we would reject even more strongly the null hypothesis $H_0^*{:}\,|v_i| = v_i^*$ for any value $0 \le v_i^* < v_{L,i}$. First, if $\tau_i \le 0$, i.e., $|\varepsilon_{\text{obs},i}| \le \sqrt{(1+\gamma_i)}$, we observe that $P_i(v_{L,i}) = P_i(-v_{L,i}) \ge 0.5$ and we never reject the null hypothesis H_0. Second, if $\tau_i > 0$, we can easily show that the function $P_i(v_i)$ is increasing in v_i over the interval $[0, v_{L,i}]$. We also note that it is a function of v_i that is symmetrical around $v_i = 0$ since $P_i(v_i) = P_i(-v_i)$. Consequently, $P_i(v_i)$ is decreasingon the interval $[-v_{L,i}, 0]$, is minimum when $v_i = 0$ and maximum when $v_i = v_{L,i}$ and $v_i = -v_{L,i}$. Therefore, when $|v_i| < v_{L,i}$, we have:

$$P_i(v_i) < P_i(v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \right).$$

References

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

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

Pfeffermann, D., and Ben-Hur, D. (2019). Estimation of randomisation mean square error in small area estimation. *International Statistical Review*, 87, S1, S31-S49.

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

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, 44, 2, 151-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54958-eng.pdf.

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.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 299-311
Statistics Canada, Catalogue No. 12-001-X

Estimating the false negatives due to blocking in record linkage

Abel Dasylva and Arthur Goussanou[1]

Abstract

When linking massive data sets, blocking is used to select a manageable subset of record pairs at the expense of losing a few matched pairs. This loss is an important component of the overall linkage error, because blocking decisions are made early on in the linkage process, with no way to revise them in subsequent steps. Yet, measuring this contribution is still a major challenge because of the need to model all the pairs in the Cartesian product of the sources, not just those satisfying the blocking criteria. Unfortunately, previous error models are of little use because they typically do not meet this requirement. This paper addresses the issue with a new finite mixture model, which dispenses with clerical reviews, training data, or the assumption that the linkage variables are conditionally independent. It applies when applying a standard blocking procedure for the linkage of a file to a register or a census with complete coverage, where both sources are free of duplicate records.

Key Words: Indexing; Massive data sets; Entity resolution; Data integration; Machine learning; Classification.

1. Introduction

Record linkage aims at finding records from the same individual in one or many files (Fellegi and Sunter, 1969; Christen, 2012; Statistics Canada, 2017a). It is different from statistical matching; an imputation method that looks for records from similar individuals (D'Orazio, Di Zio and Scanu, 2006). It has become an important data integration method that includes blocking as an important step. To block is to select a manageable subset of record pairs, which contains most matched pairs, i.e., the pairs with records that come from the same individual. Fellegi and Sunter (1969, Section 3.4) abstractly define blocking as the selection of a subset of the Cartesian product of the two data sources. Herzog, Scheuren and Winkler (2007, page 123, second paragraph) provide a similar definition when they write that "Blocking is a scheme that reduces the number of pairs of records that needs be examined." Christen (2012, page 28, third paragraph) rather uses the term indexing with the same meaning, when he writes that "To reduce the possibly very large number of pairs of records that need to be compared, indexing techniques are commonly applied… These techniques filter out record pairs that are very unlikely to correspond to matches." In this work, the term blocking is used to denote this process that is essential when linking massive data sets that are comprised of millions of records. Indeed the Cartesian product is simply too large. The purpose of blocking is to enforce a trade-off between the computational and memory resources on one hand and the loss of a few matched pairs on the other hand. These matched pairs correspond to false negatives and are an important part of the overall linkage error, if only because the blocking decisions are usually made early on in the linkage process, with no opportunity to change them later. Yet the empirical evidence has been scarce because these false negatives are never reported with few

---

[1] Abel Dasylva, Statistics Canada, 100 Tunney's Pasture, Ottawa, Canada, K1A 0T6. E-mail: abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistics Canada, 100 Tunney's Pasture, Ottawa, Canada, K1A 0T6. E-mail: arthur.goussanou@statcan.gc.ca.

exceptions, which include the identification of duplicate records in a sampling frame as described by Herzog et al. (2007, Section 12.3). In that rare instance, where the frame comprised of 176,000 business records, the number of matched pairs was estimated at 3,219 of which 3,050 were estimated to be selected by the blocking criteria, i.e. a false negative rate of (3,219 -3,050) / 3,219 = 5.25%, which is not negligible when comparing to the false negative rates reported in various linkage studies reviewed by Bohensky (2016). Nowadays, it is tempting to minimize the blocking false negatives by relaxing the blocking criteria as much as the computing resources permit. After all, these resources are already considerable and ever growing in this age of big data. Yet this may lead to the undesirable situation where the parameters of a probabilistic linkage cannot be estimated because the proportion of matched pairs is too small (Winkler, 2016, Section 2.2.3.2). Thus the issue of the blocking false negatives remains relevant regardless of the available computing resources. However estimating them has been a challenge because of the need to consider all the pairs in the Cartesian product of the two sources, and not just those satisfying the blocking criteria. In that regard, most previous error models are of little use because they do not meet this requirement, including Fellegi and Sunter (1969), Armstrong and Mayda (1993), Thibaudeau (1993), Winkler (1993), Belin and Rubin (1995), Sariyar, Borg and Pommerening (2011), Daggy, Xu, Hui, Gamache and Grannis (2013), and Chipperfield, Hansen and Rossiter (2018). Herzog et al. (2007, Chapter 12.5) have described a capture-recapture technique that does not have this drawback but is impractical because it requires clerical reviews and the conditional independence of some blocking variables.

In this work, a new solution is described, which requires neither. It is based on an extension of the model by Blakely and Salmond (2002) for situations where the records are heterogeneous and the underlying finite population is large. The solution is first developed in the ideal setting where two duplicate-free sources are linked, including a file and a register or a census with complete coverage, such that the decision to keep a pair in the blocks solely depends on its two records, as with standard blocking procedures (see Christen, 2012, Chapter 4.1). Yet, it is of interest in practical settings where both sources have few duplicate records and the census has near complete coverage, such as the linkage of tax records to the Canadian Census to replace income questions (Statistics Canada, 2017b), or a cohort study where mortality records are linked to a census (Blakely and Salmond, 2002).

The following sections are organized as follows. Section 2 presents the assumptions, notations and terminology. Section 3 explains why the distribution of neighbours provides important error information. Section 4 describes the proposed mixture model. Section 5 presents the expectation-maximization procedure. Section 6 describes the empirical study. Section 7 presents the conclusions and future work.

2. Definitions, notations and assumptions

Matched records: In record linkage, like in other automated classification problems, a clear distinction must be made between the nature of the entities to classify (whether two records are actually from the same entity) and the decisions made (whether the records are *deemed* from the same entity) according to the observations on these entities (the level of agreement between the records). However there is no consensus on the terms used to refer to these key concepts because record linkage is a multidisciplinary field, at the intersection of statistics, epidemiology and computer science. Indeed, in the first paragraph of their abstract, Fellegi and Sunter (1969) writes that “A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be matched).” Thus they refer to whether two given records belong to the same entity. In their book, Herzog, Scheuren and Winkler (2007, page 83, last paragraph) use the term “true match” for the same concept. Yet in the computer science literature, the word “matched” has an entirely different meaning. It refers to the classification decision; the best example being given by Christen (2012) in his book entitled “Data matching”. In his book, Newcombe (1988, page 105, second paragraph) also laments the lack of consensus on the meaning of the word “matched” when he writes that “This word is variously used in the literature on record linkage. In this book, however, it is given no special technical meaning and merely implies a pairing of records on the basis of some stated similarity (or dissimilarity).”

In what follows, the term “matched” is used according to the definition given by Fellegi and Sunter (1969) to refer to records from the same entity that may be a person, business, household, etc. It is also applied to a pair with the meaning that the constituent records are matched. Two records are called *unmatched* if they come from different entities.

Finite population and data sources: For the problem at hand consider a large finite population that comprises of N individuals and a recording process such that records from different individuals are mutually independent with independent recording errors. Let m denote the file size, which is assumed to be a random variable such that $m \leq N$ and $m \rightarrow \infty$ when $N \rightarrow \infty$ (e.g. $m = O(N)$). Let V denote the set of possible record values in either data source, and let v_i denote record i from the file where $v_i \in V$ by definition. For simplicity V is assumed to be finite even if it is usually very large. To further simplify, assume that the two data sources are actually free of duplicate records and that the register has no undercoverage. In other words, each record from the file corresponds to exactly one record from the same individual in the register. Each record is also assumed complete, i.e. without missing values.

Blocking strategies: When linking two large data sources, blocking is used to eliminate the vast majority of pairs with records from different individuals, while keeping all the other pairs and expanding few computing resources. Yet some pairs with records from the same individual are inevitably lost in the process. Christen (2012, Chapter 4.4) has reviewed a variety of blocking procedures including the simplest strategy, where a pair is selected if the records agree perfectly on a single key. Such a procedure is often assumed in the published literature on the analysis of linked data (Chambers and Kim, 2016; Han and Lahiri, 2018). It selects a subset of pairs based on the union of Cartesian products across disjoint post-strata that are also called blocks. In practice, a refinement of this approach is used where a pair is kept if the records agree perfectly on at least one key among many. As a result, the subset of selected pairs is no longer the union of Cartesian products across disjoint post-strata. In what follows we shall not be concerned with such details but with our ability to accurately estimate the loss resulting from the blocking procedure, when linking a file to a register or census, where both sources have few duplicate records and the register or census has little undercoverage. Perfect examples of such studies are provided by the linkage of tax records to the Canadian Census (Statistics Canada, 2017b) or by a cohort study with mortality records linked to a census (Blakely and Salmond, 2002).

In what follows, it is assumed that the decision to keep a pair only depends on its constituent records. i.e. the blocking decision is equivalent to a mathematical map from $V \times V$ into $\{0, 1\}$. This includes a large class of blocking procedures, including standard blocking procedures (Christen, 2012, Section 4.4). Yet it excludes blocking strategies that use some form of clustering such as canopy clustering (Christen, 2012, Section 4.8).

Errors: When applying blocking criteria, two kinds of errors may arise including *false negatives* and *false positives*. A false negative occurs if a matched pair is rejected by the blocking criteria. A false positive occurs if an unmatched pair is accepted by the blocking criteria. These errors are measured by the *false negative rate* (FNR) and the *false positive rate* (FPR), where the former is the proportion of matched pairs that are rejected, and the latter is the proportion of unmatched pairs that are accepted.

When designing the blocking criteria one may minimize the false positive rate while keeping the false negative rate below a threshold (e.g. 1%). Since there are usually many more unmatched pairs than matched pairs in the blocks, this roughly corresponds to minimizing the number of pairs in the blocks while keeping the proportion of lost matched pairs below the said threshold. Of course, the implementation of such a strategy requires the accurate estimation of both error rates. The false positive rate is often much easier to estimate than the false negative rate. Indeed, let B denote the total number of pairs accepted by the blocking criteria. Since the false positive rate isno less than $(B-m)/m(N-1)$ and no more than $B/m(N-1)$, it is well approximated by B/mN if $B >> m$. This estimator is related to the *reduction ratio* that is defined as $1-B/mN$ (Christen, 2012, Chapter 7.3). Estimating the false negatives is a much harder problem. Fortunately the concept of *neighbour* provides valuable insights.

3. Neighbours and errors

When examining the potential errors, it helps to look at how many register records form an accepted (by the blocking criteria) pair with a given file record. In what follows, these records are called *neighbours* of the file record, and their number is denoted by n_i for record i on the file. The empirical n_i distribution provides much information about the errors because in the current setting, each file record would have exactly one neighbour if the blocking strategy were error-free, i.e. no false negatives or false positives. Note that the satisfaction of this condition does not imply the absence of errors. As an example, consider the situation shown in Figure 3.1, where the two sources are registers of a population with $N = 6$ individuals, such that individual i is associated with record i in each register, i.e. the record pair (i, i) is matched for $i = 1, \ldots, 6.$ Before looking at the n_i's, it is known that the number of false negatives is either

0 or 1, while the number of false positives is between 0 and 5, for each record in the first register. However the n_i's provide additional error information. Indeed, when $n_i = 0$ (e.g. with record 2), it is known with certainty that there is a false negative but no false positives. When $n_i = 1$, one of two cases may occur, including a first case (as with record 5) where the neighbour is the matched record such that there are no errors, and a second case (as with record 3) where the neighbour is an unmatched record such that there are two errors including a false negative and a false positive. In summary, when $n_i = 1$, the number of false negatives is 0 or 1, while the number of false positives is also 0 or 1. Thus there is no additional information about the false negatives since it was known to be 0 or 1 prior to looking at n_i. However, much information is gained about the false positives, since it was known to be in a wider interval (0 to 5) before looking at n_i. This observation confirms the relative ease with which the false positives may be estimated.

Figure 3.1 Two registers with six individuals.

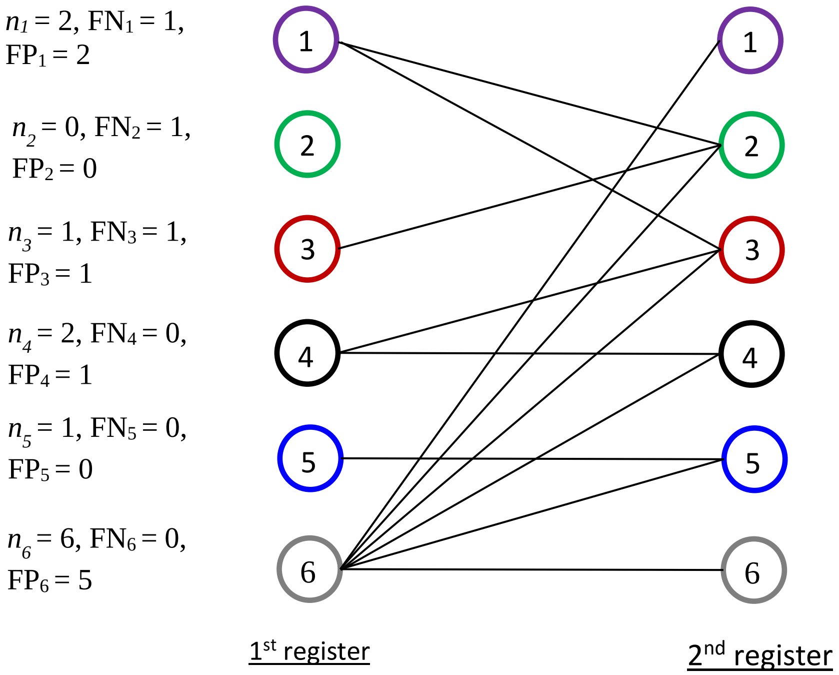


Table 3.1 summarizes the general connection between the number of neighbours and the linkage errors at a given record, in the current setting where each file record is matched with exactly one register record.

Table 3.1
Neighbours and errors

Neighbours (n_i)	False negatives	False positives	Full error information (yes/no)
0	1	0	Yes
1	0 or 1	0 or 1	No
$[2, N-1]$	0 or 1	$n_i - 1$ or n_i	No
N	0	$N - 1$	Yes

The above table clearly demonstrates that the number of neighbours provides much error information, including in the case $n_i = 0$ and the more unlikely case $n_i = N,$ where this information is complete. When the blocking decision about two records depends on no other record, with a high probability when m and N are large, some file records are bound to have no neighbour. In theory one could design the blocking criteria to ensure a positive n_i for each file record, but this would violate the assumption made in Section 2 that the blocking decision about two records depends on no other record. Under this assumption, the number of neighbours does provide valuable error information, but some uncertainty remains in the case $1 \leq n_i \leq N-1,$ where a statistical model is needed to predict the errors based on $n_i.$

4. Finite mixture model

The linkage of the two sources is of interest when it is a viable option even N is very large. To capture the essence of such situations, the following two regularity conditions are assumed.

a. Two matched records are neighbours with a probability that is bounded away from 0 regardless of $N.$
b. Two unmatched records are *accidental* neighbours with a probability of $O(1/N).$

These assumptions imply that each record has a bounded expected number of neighbours and that $O(N)$ pairs (instead of $O(mN)$ pairs and even $O(N^2)$ pairs if $m = O(N)$) are selected by the blocking criteria. They also imply that there is enough linkage information to identify matched records with a success probability, which is bounded away from zero, regardless of the population size. The above assumptions further imply a particular limiting distribution for the number of neighbours $n_i.$ Indeed, let $n_i = n_{i|M} + n_{i|U},$ where $n_{i|M}$ is the number of matched neighbours and $n_{i|U}$ is the number of unmatched neighbours. Note that these latter variables are not directly observed expect when $n_i = 0$ or $n_i = N$ (see Table 3.1). They are also conditionally independent given v_i and such that $n_{i|M} \big| v_i \sim \text{Bernoulli}\,(p\,(v_i)),$ $n_{i|U} \big| v_i \sim \text{Binomial}\,(N-1, \lambda\,(v_i)\,/\,(N-1)),$ if an unmatched record is a neighbour with the probability $\lambda(v_i)\,/\,(N-1)$ independently of the other unmatched records. When the functions $p(.)$ and $\lambda(.)$ do not depend on N and N is large, we have $n_{i|U} \big| v_i \mathrel{\dot\sim} \text{Poisson}\,(\lambda(v_i))$ (Billingsley, 1995), where $\dot\sim$ means approximately distributed as. Hence, $n_i \,\big|\, v_i \mathrel{\dot\sim} \text{Bernoulli}\,(p\,(v_i)) * \text{Poisson}\,(\lambda\,(v_i)),$ where $*$ is the convolution operator. Note that, in general, the functions $p(.)$ and $\lambda(.)$ are unknown high-dimensional parameters. To simplify, further assume that $(p(.), \lambda(.))$ is (well approximated by) a piecewise constant function with G levels, such that we have the finite mixture model $n_i \sim \sum_{g=1}^{G} \alpha_g\,(\text{Bernoulli}\,(p_g) * \text{Poisson}\,(\lambda_g))$ holds approximately. When G is fixed, the unknown model parameters are given by the vector $\psi = \left[(\alpha_g, p_g, \lambda_g)\right]_{1 \leq g \leq G}$ that may be estimated with the Expectation-Maximization (EM) procedure in the next section.

The connection between the error rates and model parameters is made by first noting that the FNR and FPR definitions imply

$$\begin{aligned} \text{FNR} &= \frac{1}{m}\sum_{i=1}^{m}(1-n_{i\,|M}), \\ (N-1)\ \text{FPR} &= \frac{1}{m}\sum_{i=1}^{m} n_{i\,|U}. \end{aligned} \tag{4.1}$$

When $m = N$ almost surely, the above equations imply that

$$\begin{aligned} E[\text{FNR}] &= 1 - E[n_{i\,|M}], \\ &= 1 - E[p(v_i)], \\ (N-1)\ E[\text{FPR}] &= E[n_{i\,|U}] \\ &= E[\lambda(v_i)], \end{aligned} \tag{4.2}$$

where $E[p(v_i)] = \sum_{g=1}^{G} \alpha_g p_g$ and $E[\lambda(v_i)] = \sum_{g=1}^{G} \alpha_g \lambda_g$ with the finite mixture model. When m is random and such that

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m} n_{i\,|M} &\overset{p}{\rightarrow} E[n_{i\,|M}], \\ \frac{1}{m}\sum_{i=1}^{m} n_{i\,|U} &\overset{p}{\rightarrow} E[n_{i\,|U}], \end{aligned} \tag{4.3}$$

as $N \rightarrow \infty,$ the error rates and the model parameters are related as follows

$$\begin{aligned} \text{FNR} &\overset{p}{\rightarrow} 1 - E[p(v_i)], \\ (N-1)\ \text{FPR} &\overset{p}{\rightarrow} E[\lambda(v_i)]. \end{aligned} \tag{4.4}$$

5. Estimation procedure

The model parameters may be estimated by maximizing the *composite* likelihood (Varin, Reid and Firth, 2011) of the sample $n_1, \ldots, n_m$. For brevity, this composite likelihood is subsequently called likelihood. To develop the EM procedure (Dempster, Laird and Rubin, 1977) it is convenient to first derive the maximum likelihood (ML) equations for the *complete data*, which are comprised of the latent variables $n_{i\,|\,M}$, $n_{i\,|\,U}$ and $(c_{i1}, \ldots, c_{iG})$ for each i; c_{ig} being the indicator that record i is from class g.

After some algebra, the ML equations for the complete data are as follows.

$$\begin{aligned} \hat{p}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i\,|M}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i\,|U}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} c_{ig}. \end{aligned} \tag{5.1}$$

Consequently the ML equations for the observed data (the n_i's) are as follows.

$$\begin{aligned}\hat{p}_g &= \frac{\sum_{i=1}^m E[c_{ig} n_{i|M} \mid n_i; \psi]}{\sum_{i=1}^m E[c_{ig} \mid n_i; \psi]} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^m E[c_{ig} n_{i|U} \mid n_i; \psi]}{\sum_{i=1}^m E[c_{ig} \mid n_i; \psi]} \\ \hat{\alpha}_g &= \frac{1}{m} \sum_{i=1}^m E[c_{ig} \mid n_i; \psi].\end{aligned} \tag{5.2}$$

The EM procedure alternates between the M-step given by Equation (5.2) and the E-step equations in Appendix A.

The above procedure may produce consistent point estimators even if it treats the sample $n_1, \ldots, n_m$ as if it were independent and identically distributed. However this is likely to generate some bias when estimating the variance and the critical levels of hypothesis tests.

6. Empirical study

The empirical study is based on staffing data from the Public Service Resourcing System (PSRS), which is used by applicants to the federal public service in Canada. A given user may open many accounts and apply to many jobs using the same account; each account being associated with a distinct email address. To fulfill its mandate, the Public Service Commission needs to identify all accounts from a given applicant. However this is a challenge because there is no unique identifier except for a minority of applicants. Instead, for most records, the linkage must be based on the given name, the surname and the partial birthdate, which are available for all records. The partial birthdate is comprised of the day and month of birth along with the last digit of the birth year.

The empirical study is based on a subset of 126,330 records selected from the PSRS data since 2006. The selection is based on the following criteria.

- A nonmissing unique identifier.
- Nonmissing given name, surname and partial birthdate.
- Two records for each selected value of the unique identifier.

The selected records represent 63,155 distinct values of the identifier and so many distinct individuals, with two matched records per individual. These records are split into two complete and duplicate-free registers that are linked with the following blocking criteria, and without the unique identifier. A pair is selected if the partial birthdate is the same and the SOUNDEX code (Herzog et al., 2007, Chapter 11) is the same for the given name or the surname. The expected error rates are estimated with the model and compared with the actual values based on the unique identifier.

In Figure 6.1, the histogram shows that the vast majority or records have exactly one neighbour. However 1,659 records have no neighbour, while five records have five neighbours; the maximum number of neighbours of any record.

Figure 6.1 Histogram of the number of neighbours.

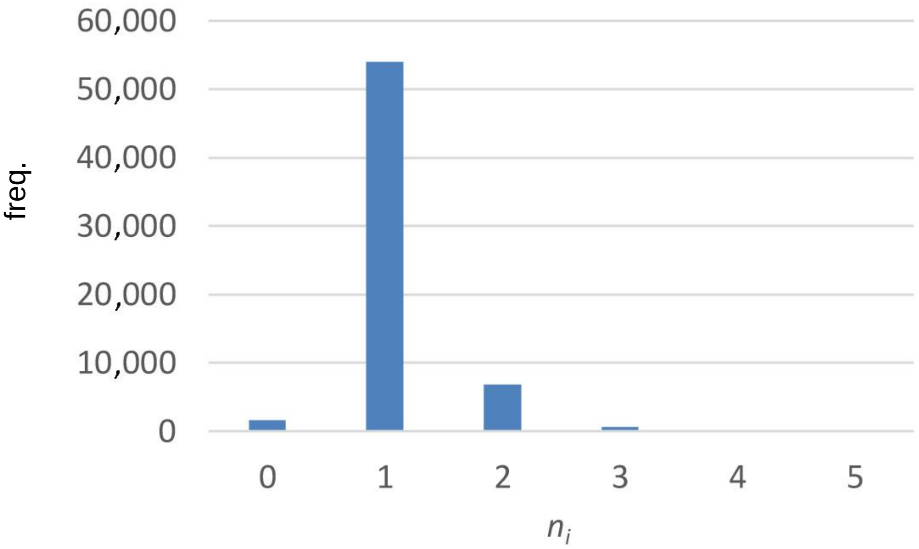


Table 6.1 cross-classifies the records by their number of neighbours and linkage errors, in agreement with Table 3.1.

Table 6.1
Number of neighbours and errors

Neighbours (n_i)	False negatives	False positives	Freq.
0	1	0	1,659
1	1	1	116
1	0	0	53,835
2	1	2	8
2	0	1	6,867
3	1	3	1
3	0	2	602
4	0	3	62
5	0	4	5

The confusion matrix is as follows.

Table 6.2
Confusion matrix

	Link	Non-link	Total
Matched	61,371	1,784	63,155
Unmatched	8,412	3.99E9	3.99E9
Total	69,783	3.988E9	3.989E9

From this matrix, $\mathrm{FNR} = 1{,}784 / 63{,}155 = 0.0282$ and $\mathrm{FPR} = 8{,}412 / 3.99\mathrm{E}9 = 2.11\mathrm{E} - 6.$ Both measures may be viewed as the estimators $\hat{E}[\mathrm{FNR}]$ and $\hat{E}[\mathrm{FPR}]$ of their respective expectations. Since the false negative rate is the summation of independent and identically distributed random variables, its variance may be estimated by

$$\hat{\mathrm{var}}(\mathrm{FNR}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (1 - n_{i|M} - \mathrm{FNR})^2,$$

based on the latent variables $n_{1|M}, \ldots, n_{m|M}$, which are not directly observed in practice. As a result, the estimated FNR variance is $\hat{\mathrm{var}}(\mathrm{FNR}) = 4.35\mathrm{E} - 7.$ This means the estimated standard error $\hat{\mathrm{SE}}(\hat{E}[\mathrm{FNR}]) = 6.6\mathrm{E} - 4$ for the estimator $\hat{E}[\mathrm{FNR}],$ and the 95% normal confidence interval $\hat{E}[\mathrm{FNR}] \mp z_{\alpha/2} \hat{\mathrm{SE}}(\hat{E}[\mathrm{FNR}]) = (2.82\mathrm{E} - 2 \mp 1.3\mathrm{E} - 3)$ for the expected FNR, where $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$. The corresponding 99% confidence interval is $(2.82\mathrm{E} - 2 \mp 1.71\mathrm{E} - 3)$. Estimating the FPR variance is more difficult because the FPR involves a second order U statistic (Hoeffding, 1948; Lee, 1990). As a matter of fact, Table 6.1 does not give enough information to estimate this statistic. Estimating the variance of the model-based estimators is also challenging because the n_i's are correlated. All the point estimates are given in Table 6.3, where the first row gives the actual FNR and FPR.

Table 6.3
Point estimates

		$\hat{E}$[FNR]	$\hat{E}$[FPR]
Unique id		0.0282	2.11E-6
Model	G = 1	0.0301	2.14E-6
	G = 2	0.0298	2.13E-6
	G = 3	0.0303	2.14E-6

The results show that the model based estimates are very close to the actual FNR and FPR when using one, two or three classes. For the false negative rate, the relative error is $100 \times |0.0303 - 0.0282| / 0.0282 = 7.45\%,$ while this relative error is $100 \times |2.11 - 2.14| / 2.11 = 1.42\%$ for the false positive rate. The small relative errors are encouraging regarding the accuracy of the proposed estimators, even if the model estimates of the expected FNR lie slightly outside the 95% confidence interval. However, the estimate belongs to the 99% confidence interval when using two classes. Choosing two classes seems optimal because the resulting estimate has the smallest relative error with respect to the actual FNR.

7. Conclusions and future work

A new finite mixture has been proposed for estimating the false negatives due to a standard blocking procedure, when linking a file to a register or a census with complete coverage, when both sources are free of duplicate records. An empirical study with social data gives encouraging results. Yet future work must address the issues of variance estimation and statistical inference about the number of classes. Extensions are also required to account for undercoverage and duplicate records.

Disclaimer

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that might not reflect those currently implemented by the Agency.

Acknowledgements

The authors express their gratitude towards Dr. Jonnagada Rao for his insight and towards the Public Service Commission for access to the data.

Appendix A

For the E-step, the equations are as follows.

$$
\begin{aligned}
P\,(n_i \mid c_{ig} = 1) &= I\,(n_i = 0)\,(1 - p_g)\,e^{-\lambda_g} + I\,(n_i > 0)\left(p_g + (1 - p_g)\frac{\lambda_g}{n_i}\right)\frac{e^{-\lambda_g}\,\lambda_g^{n_i - 1}}{(n_i - 1)!} \\
P\,(c_{ig} = 1 \mid n_i) &= \frac{\alpha_g P\,(n_i \mid c_{ig} = 1)}{\sum_{g' = 1}^{G} \alpha_{g'} P\,(n_i \mid c_{ig'} = 1)} \\
P\,(n_{i\,|M} = 1 \mid n_i, c_{ig} = 1) &= \frac{p_g n_i}{p_g n_i + (1 - p_g)\,\lambda_g} \\
P\,(n_{i\,|U} = n_i \mid n_i, c_{ig} = 1) &= I\,(n_i = 0) + I\,(n_i > 0)\frac{(1 - p_g)\,\lambda_g}{p_g n_i + (1 - p_g)\,\lambda_g} \\
P\,(n_{i\,|U} = n_i - 1 \mid n_i, c_{ig} = 1) &= \frac{p_g n_i}{p_g n_i + (1 - p_g)\,\lambda_g}
\end{aligned}
$$

and

$$
\begin{aligned}
E\,[c_{ig} n_{i\,|M} \mid n_i] &= P\,(c_{ig} = 1 \mid n_i)\,P\,(n_{i\,|M} = 1 \mid n_i, c_{ig} = 1) \\
E\,[c_{ig} n_{i\,|U} \mid n_i] &= P\,(c_{ig} = 1 \mid n_i)\,E\,[n_{i\,|U} \mid n_i, c_{ig} = 1] \\
E\,[n_{i\,|U} \mid n_i, c_{ig} = 1] &= \left(\frac{p_g\,(n_i - 1) + (1 - p_g)\,\lambda_g}{p_g n_i + (1 - p_g)\,\lambda_g}\right) n_i.
\end{aligned}
$$

References

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

Belin, T., and Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Billingsley, P. (1995). *Probability and Measure*, Wiley.

Blakely, T., and Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *Journal of Epidemiology*, 31, 1246-1252.

Bohensky, M. (2016). Bias in data linkage studies. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 63-82.

Chambers, R., and Kim, G. (2016). Secondary analysis of linked data. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 83-108.

Chipperfield, J., Hansen, N. and Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *International Statistical Review*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*, Springer.

Daggy, J., Xu, H., Hui, S., Gamache, R. and Grannis, S. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39, 1-38.

D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching, Theory and Practice*, Wiley.

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

Han, Y., and Lahiri, P. (2018). Statistical analysis with linked data. *International Statistical Review*, 87, special issue, 139-157, 1C19 doi:10.1111/insr.12295.

Herzog, T., Scheuren, F. and Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, Springer.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19, 293-325.

Lee, A.J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker.

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Sariyar, M., Borg, A. and Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Statistics Canada (2017a). *Record Linkage Project Process Model*, Catalogue No. 12-605-X.

Statistics Canada (2017b). *Income Reference Guide, Census of Population, 2016*, Catalogue No. 98-500-X2016004.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology*, 19, 1, 31-38. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993001/article/14477-eng.pdf.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Winkler, W. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the 1993 Joint Statistical Meetings*, American Statistical Association, 274-279.

Winkler, W.E. (2016). Probabilistic linkage. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 7-35.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 313-347
Statistics Canada, Catalogue No. 12-001-X

With-replacement bootstrap variance estimation for household surveys Principles, examples and implementation

Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet and Cédric Garcia[1]

Abstract

Variance estimation is a challenging problem in surveys because there are several nontrivial factors contributing to the total survey error, including sampling and unit non-response. Initially devised to capture the variance of non-trivial statistics based on independent and identically distributed data, the bootstrap method has since been adapted in various ways to address survey-specific elements/factors. In this paper we look into one of those variants, the with-replacement bootstrap. We consider household surveys, with or without sub-sampling of individuals. We make explicit the benchmark variance estimators that the with-replacement bootstrap aims at reproducing. We explain how the bootstrap can be used to account for the impact sampling, treatment of non-response and calibration have on total survey error. For clarity, the proposed methods are illustrated on a running example. They are evaluated through a simulation study, and applied to a French Panel for Urban Policy. Two SAS macros to perform the bootstrap methods are also developed.

Key Words: Bootstrap; Calibration; Variance estimation; Unit non-response.

1. Introduction

Variance estimation is a challenging problem in surveys. The final weights used at the estimation stage include several statistical treatments, including correction of unit non-response and calibration, and their impact on the variance is to be assessed. Bootstrap is a useful tool, leading to the creation of so-called bootstrap weights released with the survey data set. These weights can be used to compute repeatedly the bootstrap version of the parameter of interest, leading to a simulation-based variance estimator or confidence interval. The interest for practitioners is that no information other than the bootstrap weights is needed for variance estimation. In particular, a comprehensive description of the original sampling design and estimation process is not required, which would be the case under an analytic approach where the variance estimator needs to be worked out. And thus the same set of bootstrap weights is to be used to obtain a variance estimate regardless of whether the parameters of interest are totals, medians or regression coefficients. Even when a comprehensive description of the sampling design and estimation process is available, the analytic approach poses issues for important parameters for which linearization variance estimation is not straightforward; see for example Shao (1994) for L-statistics, and Shao and Rao (1993) for low income proportions.

There is an extensive literature on bootstrap in survey sampling, see for example Rao and Wu (1988), Rao, Wu and Yue (1992), Shao and Tu (1995, Chapter 6), Davison and Hinkley (1997, Section 3.7), Davison and Sardy (2007), Chauvet (2007) and Mashreghi, Haziza and Léger (2016) for detailed reviews.

---

[1] Pascal Bessonneau, Ined; Gwennaëlle Brilhaut, Ined; Guillaume Chauvet, Ensai (Irmar), Campus de Ker Lann, Bruz - France. E-mail: guillaume.chauvet@ensai.fr; Cédric Garcia, Université Gustave Eiffel.

One of these techniques is the so-called rescaled bootstrap proposed by Rao and Wu (1988), which may be summarized as follows. First, inside each first-stage sample S_h of size n_h selected in stratum h, a with-replacement simple random sample of size m_h is selected, leading to the initial bootstrap weights. Then, these weights may be rescaled so as to reproduce an unbiased variance estimator for the estimation of a total (linear case). As explained by Rao and Wu (1988), the rescaled bootstrap may be applied to a variety of sampling designs including two-stage sampling and with/without-replacement sampling at the first stage. However, it is not straightforward to account for some practical features of a survey such as the treatment of unit non-response. This is considered in Yeo, Mantel and Liu (1999) and Girard (2009). A related topic is treated in Kim, Navarro and Fuller (2006), who consider replication variance estimation for two-phase sampling.

Applying the Rao-Wu bootstrap in the particular case when the resample sizes are $m_h = n_h - 1$ leads to the so-called bootstrap of Primary Sampling Units (PSUs) or with-replacement bootstrap (McCarthy and Snowden, 1985). The with-replacement bootstrap is fairly simple to implement; in particular, it requires to resample the primary sampling units only, and not the final units. Accounting for treatment of non-response and calibration is fairly natural, as explained in this paper. An important property of a bootstrap method is to match (at least, approximately) a known variance estimator in the linear case, which we call the benchmark variance estimator. For with-replacement bootstrap, it is possible to state precisely this benchmark variance estimator at any step of the method, which is helpful in understanding how the method works to assess the total survey error. The with-replacement bootstrap leads to conservative variance estimation, in the sense that the first-stage sampling variance is overestimated if the sampling designs used inside strata at first-stage are more efficient than multinomial sampling, which we assume to hold true in this paper. This is therefore a prudent approach in producing confidence intervals. The positive bias of the bootstrap variance estimator is expected to be negligible when the first-stage sampling rates inside strata are negligible, which is often the case in phone surveys. Also, if the survey is repeated over time, the contribution of the first-stage sampling variance is likely to fade while the variance due to attrition and unit non-response grows bigger.

Our paper, which examines the with-replacement bootstrap, is intended to be user-oriented. In particular, we do not propose particular modifications of the with-replacement bootstrap. Rather, we explain how this bootstrap method may be applied to account for sampling, treatment of non-response and calibration, and in so doing, what is the variance estimator that we aim at reproducing when estimating a total. We give some running examples to illustrate how bootstrap weights are computed in simple cases. Two SAS macros implementing the proposed bootstrap methods are presented, evaluated through a simulation study, and illustrated on a real survey dataset from the Panel for Urban Policy.

For simplicity of presentation, our terminology is that of household surveys, which is our original motivation for this paper. We consider two cases: first, when a sample of households only is selected; secondly, when a subsample of individuals is selected inside the selected households. Despite this specific terminology, our approach is general and may be applied to any other situation when a survey is performed by one-stage sampling (first case) or by two-stage sampling (second case).

We are in particular interested in household phone surveys, which have been extensively used at the French National Institute for Demographic Studies (INED) over the last decades. Originally, a sample of phone numbers was selected from a register of fixed-line numbers, and more recently the phone numbers used in the survey are randomly generated to account for households not covered in the registers (unlisted or cell numbers). In a second step, individuals are selected within the households, using classic selection methods (e.g., Kish individual). Phone surveys have proved to be efficient, specifically for sensitive subjects like sexuality, violence or addictions. Some examples of surveys performed by INED include the national survey on violence against women in France in 2000 (ENVEFF), the national survey on violence and gender exchange in 2015 and 2018 (VIRAGE and VIRAGE overseas, respectively), or the national survey on the context of sexuality in France in 2006. The same protocol is likely to be used in a near future for surveys on similar subjects, like the one on young adults' sexuality or the one on birth control, to begin between 2021 and 2023.

The paper is organized as follows. In Section 2, our main notations are defined, and we consider the estimation of a total by accounting for sampling, unit non-response and calibration. We treat in Section 2.1 the situation when a sample of households only is selected (one-stage case), and in Section 2.2 the case when individuals are sub-sampled within households (two-stage case). The basic bootstrap method is described in Section 3: the one-stage case is considered in Sections 3.1 and 3.2, and the two-stage case is considered in Sections 3.3 and 3.4. We explain in Section 3.5 how the basic bootstrap procedure may be applied to obtain an estimator of variance or a confidence interval. The proposed bootstrap methods are evaluated in Section 4 through a simulation study. We present in Section 5 an illustration on a sample of households and individuals from the French Panel for Urban Policy. We conclude in Section 6. The benchmark variance estimators for the sample of individuals are presented in Appendix A. The SAS program used to perform bootstrap variance estimation are presented in Appendices B and C. These SAS programs are available upon request to the corresponding author.

2. Notation and estimation

In this section, we define our main notations, and we describe the sampling and estimation process. We first consider in Section 2.1 the case when a sample of households only is selected, and we describe the estimation process which includes treatment of unit non-response and calibration. We indicate in each case what is the benchmark variance estimator considered, i.e. the variance estimator that we aim at reproducing for the estimation of a total with the bootstrap method proposed in Section 3. The case when individuals are sub-sampled inside households is covered in Section 2.2. The benchmark variance estimators for this second case are given in Appendix A.

2.1 Case of a sample of households only

We consider estimation for a population U_{hh} of households. We let y_k denote the value taken by some variable of interest for the household k. We are interested in the estimation of the total

$$Y_{hh} = \sum_{k \in U_{hh}} y_k. \tag{2.1}$$

2.1.1 Sampling design

We suppose that a sample S_{hh} is selected in U_{hh} by means of a stratified one-stage sampling design. The population U_{hh} is partitioned into H strata $U_{hh}^1, \ldots, U_{hh}^H,$ the samples $S_{hh}^1, \ldots, S_{hh}^H$ are selected inside independently, and the sample S_{hh} is the union of these samples. We let π_k denote the inclusion probability of a given household k. The design weight is

$$d_k = \frac{1}{\pi_k}. \tag{2.2}$$

In case of full response, the estimator of Y_{hh} is

$$\hat{Y}_{hh} = \sum_{k \in S_{hh}} d_k y_k. \tag{2.3}$$

We consider as a benchmark variance estimator

$$v_{\text{mult}}(\hat{Y}_{hh}) = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k y_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} y_{k'} \right)^2 \right], \tag{2.4}$$

with n_h the size of the sample S_{hh}^h. This variance estimator is unbiased if the samples are selected inside strata by multinomial sampling (Tillé, 2011, Section 5.4), a.k.a. sampling with replacement. It is conservative if the sampling designs used inside strata are more efficient than multinomial sampling (Särndal, Swensson and Wretman, 1992, Section 4.6), which we assume to hold true in the rest of the paper. The positive bias of this variance estimator is expected to be negligible when the sampling rates inside strata are themselves negligible, which is often the case in phone surveys. This is illustrated by the results of our simulation study, see Section 4.

2.1.2 Treatment of non-response

In practice, the sample S_{hh} is prone to unit non-response, which leads to the observation of a sub-sample of respondents $S_{r,hh}$ only. We let r_k denote the response indicator of a household k, and p_k denote the response probability of the household k. We suppose that the households respond independently of one another. Also, we suppose that unit non-response is handled through the method of Response Homogeneity Groups (RHGs), which is popular in practice (e.g. Brick, 2013; Juillard and Chauvet, 2018). Under this framework, it is assumed that the sample S_{hh} may be partitioned into C RHGs denoted as $S_{1,hh}, \ldots, S_{C,hh}$ such that the response probability p_k is constant inside a RHG.

For $c = 1, \ldots, C,$ we let p_c denote the common response probability inside the RHG $S_{c,hh}$. It is estimated by

$$\hat{p}_c = \frac{\sum_{k \in S_{c,hh}} \theta_k r_k}{\sum_{k \in S_{c,hh}} \theta_k}, \tag{2.5}$$

with θ_k some weight attached to the household k. The choice $\theta_k = 1$ leads to estimating p_c by the unweighted response rate inside the RHG. The choice $\theta_k = d_k$ leads to estimating p_c by the response rate inside the RHG, weighted by the sampling weights (e.g. Kott, 2012).

Accounting for the estimated response probabilities leads to the weights corrected for non-response

$$d_{rk} = \frac{d_k}{\hat{p}_{c(k)}}, \tag{2.6}$$

with $c(k)$ the RHG of the household k. The estimator of Y_{hh} adjusted for non-response is

$$\hat{Y}_{r,hh} = \sum_{k \in S_{r,hh}} d_{rk} y_k. \tag{2.7}$$

Building on the multinomial variance estimator in (2.4) and on linearization for estimators reweighted for unit-non-response (Kim and Kim, 2007, Section 2), our benchmark variance estimator is

$$v_{\text{mult}}(\hat{Y}_{r,hh}) = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k u_{1k} - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} u_{1k'} \right)^2 \right], \tag{2.8}$$

with

$$u_{1k} = \theta_k \pi_k \bar{y}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}} \left\{ y_k - \theta_k \pi_k \bar{y}_{rc(k)} \right\},$$

and

$$\bar{y}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k y_k}{\sum_{k \in S_{c,hh}} \theta_k r_k}.$$

This is a conservative estimator for the asymptotic variance of $\hat{Y}_{r,hh}$. A key assumption for this to hold is that the response indicators r_k are mutually independent.

2.1.3 Calibration

Lastly, the weights adjusted for non-response are calibrated on auxiliary totals known on the population. For simplicity, we describe only the Generalized REGression estimator (GREG, Särndal et al., 1992, Chapter 6). Let x_k denote the vector of calibration variables at the household level, and X_{hh} the total on the population U_{hh}. For the sample $S_{r,hh}$, this leads to the linear calibrated weights

$$w_k = d_{rk} \left(1 + x_k^\top \lambda_{hh} \right),$$

with

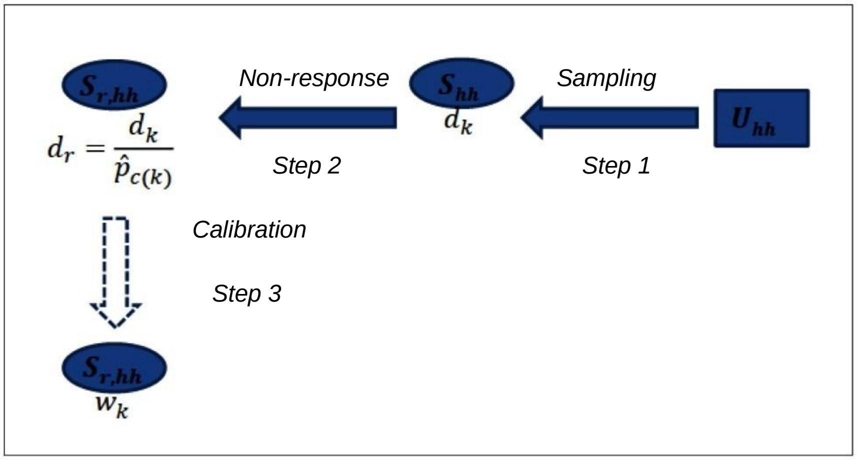
$$\lambda_{hh} = \left(\sum_{k \in S_{r,hh}} d_{rk} x_k x_k^{\top} \right)^{-1} \left(X_{hh} - \hat{X}_{r,hh} \right), \tag{2.9}$$

and where $\hat{X}_{r,hh}$ is the estimator of X_{hh}, obtained by plugging x_k into (2.7). The calibrated estimator is

$$\hat{Y}_{\text{cal},hh} = \sum_{k \in S_{r,hh}} w_k y_k. \tag{2.10}$$

The sampling and estimation steps are summarized in Figure 2.1.

Figure 2.1 Sampling and estimation steps for a household sample.



Using linearization for estimators reweighted for unit-non-response and calibrated (Kim and Kim, 2007, Section 5), our benchmark variance estimator is

$$v_{\text{mult}}(\hat{Y}_{\text{cal},hh}) = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k u_{2k} - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} u_{2k'} \right)^2 \right], \tag{2.11}$$

with

$$u_{2k} = \theta_k \pi_k \bar{e}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}} \left\{ e_k - \theta_k \pi_k \bar{e}_{rc(k)} \right\},$$

and

$$\bar{e}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k e_k}{\sum_{k \in S_{c,hh}} \theta_k r_k},$$

where we let

$$e_k = y_k - \hat{B}_{r,hh}^{\top} x_k \quad \text{with} \quad \hat{B}_{r,hh} = \left(\sum_{k \in S_{r,hh}} d_{rk} x_k x_k^{\top} \right)^{-1} \sum_{k \in S_{r,hh}} d_{rk} x_k y_k \tag{2.12}$$

denote the estimated regression residuals of the variable of interest on the calibration variables. This is a conservative estimator for the asymptotic variance of $\hat{Y}_{\text{cal},hh}$.

2.1.4 Computation of household weights on an example

To fix ideas, we describe a small example. We consider a population U_{hh} of $N_{hh} = 100$ households. We suppose without loss of generality that a single stratum is used, and that a sample of $n_{hh} = 10$ households is selected.

The sample is $S = \{A, B, \dots, J\}$. The inclusion probabilities of the selected units are (say)

$$\pi_A = \pi_B = \pi_C = \pi_D = \frac{1}{4} \quad \text{and} \quad \pi_E = \pi_F = \pi_G = \pi_H = \pi_I = \pi_J = \frac{1}{16}, \tag{2.13}$$

resulting in the design weights

$$d_A = d_B = d_C = d_D = 4 \quad \text{and} \quad d_E = d_F = d_G = d_H = d_I = d_J = 16. \tag{2.14}$$

Among the 10 selected households, 7 only are surveyed due to non-response. It is accounted for by using the method of RHGs, with two groups: the units $A,$ $B,$ F and J in the first one, and the units $C,$ $D,$ $E,$ $G,$ $H,$ and I in the second one. The units $B,$ C and G are non-respondents. Inside each RHG, we compute estimated response probabilities, weighted by the design weights $(\theta_k = d_k).$ This leads to

$$\begin{aligned} \hat{p}_1 &= \frac{\sum_{k \in S_{1,hh}} d_k r_k}{\sum_{k \in S_{1,hh}} d_k} = \frac{d_A + d_F + d_J}{d_A + d_B + d_F + d_J} = \frac{9}{10}, \\ \hat{p}_2 &= \frac{d_D + d_E + d_H + d_I}{d_C + d_D + d_E + d_G + d_H + d_I} = \frac{13}{18}. \end{aligned} \tag{2.15}$$

The weights accounting for non-response are obtained for the respondents by dividing the sampling weights by the estimated response probabilities. This leads to the weights

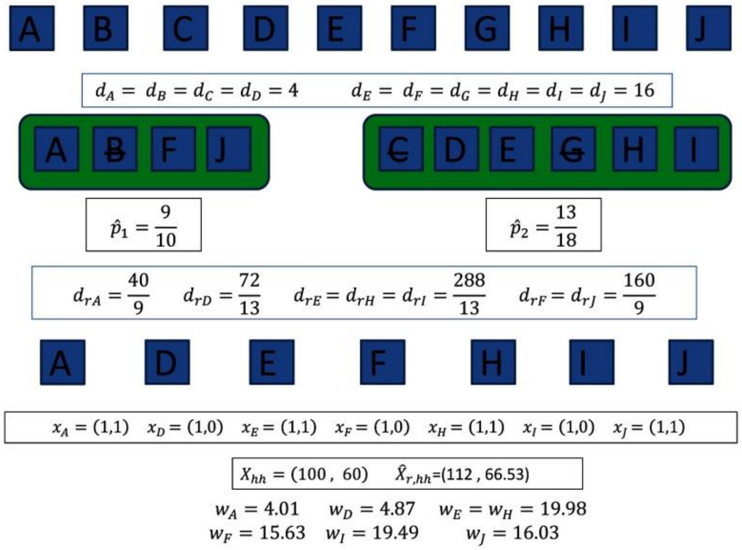
$$d_{rA} = \frac{40}{9} \quad d_{rD} = \frac{72}{13} \quad d_{rE} = d_{rH} = d_{rI} = \frac{288}{13} \quad d_{rF} = d_{rJ} = \frac{160}{9}. \tag{2.16}$$

Finally, the weights are calibrated to match exactly the population size $N_{hh} = 100$ and an auxiliary total $X_{1,hh} = 60.$ Note that, using the sample of respondents, we obtain $\hat{N}_{r,hh} = 112$ and $\hat{X}_{1r,hh} = 66.53.$ The calibrated weights are

$$\begin{aligned} &w_A = 4.01, \quad w_D = 4.87, \quad w_E = w_H = 19.98, \\ &w_F = 15.63, \quad w_I = 19.49, \quad w_J = 16.03. \end{aligned} \tag{2.17}$$

The sampling and estimation steps are summarized in Figure 2.2.

Figure 2.2 Estimation steps for the weighting of households.


A B C D E F G H I J
$d_A = d_B = d_C = d_D = 4$ $d_E = d_F = d_G = d_H = d_I = d_J = 16$
A ~~B~~ F J
~~C~~ D E ~~G~~ H I
$\hat{p}_1 = \frac{9}{10}$
$\hat{p}_2 = \frac{13}{18}$
$d_{rA} = \frac{40}{9}$ $d_{rD} = \frac{72}{13}$ $d_{rE} = d_{rH} = d_{rI} = \frac{288}{13}$ $d_{rF} = d_{rJ} = \frac{160}{9}$
A D E F H I J
$x_A = (1,1)$ $x_D = (1,0)$ $x_E = (1,1)$ $x_F = (1,0)$ $x_H = (1,1)$ $x_I = (1,0)$ $x_J = (1,1)$
$X_{hh} = (100, 60)$ $\hat{X}_{r,hh} = (112, 66.53)$
$w_A = 4.01$ $w_D = 4.87$ $w_E = w_H = 19.98$
$w_F = 15.63$ $w_I = 19.49$ $w_J = 16.03$

2.2 Case of a sample of households and individuals

We are interested in the population U_{ind} of individuals associated to the population U_{hh} of households considered in Section 2.1. If we let y_l denote the value taken by some variable of interest for the individual l, the parameter of interest is

$$Y_{\text{ind}} = \sum_{l \in U_{\text{ind}}} y_l. \tag{2.18}$$

2.2.1 Sampling design

Within any sampled household $k \in S_{hh}$, a subsample $S_{\text{ind},k}$ of individuals is selected, and the sample S_{ind} is the union of these samples. We let $\pi_{l|k}$ denote the conditional inclusion probability of the individual l inside the household k. The conditional design weight of l is

$$d_{l|k} = \frac{1}{\pi_{l|k}} \quad \text{for any} \quad l \in k, \tag{2.19}$$

and the non-conditional design weight is

$$d_l = d_{l|k} \times d_k \quad \text{for any} \quad l \in k. \tag{2.20}$$

In case of full response, the estimator of Y_{ind} is

$$\hat{Y}_{\text{ind}} = \sum_{k \in S_{hh}} d_k \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l = \sum_{k \in S_{\text{ind}}} d_l y_l. \tag{2.21}$$

The benchmark variance estimator for $\hat{Y}_{\text{ind}}$ is obtained from (2.4), by replacing y_k with

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l. \tag{2.22}$$

2.2.2 Treatment of non-response

The weights of individuals accounting for the non-response of households are

$$d_{rl} = d_{rk(l)} d_{l|k(l)} \quad \text{with } k(l) \text{ the household containing } l, \tag{2.23}$$

with d_{rk} the weight of household k corrected for unit non-response (see equation (2.6)), and $d_{l|k}$ the conditional sampling weight of individual l inside the household k (see equation (2.19)). We let

$$S_{r,\text{ind}} = \bigcup_{k \in S_{r,hh}} S_{\text{ind},k} \tag{2.24}$$

denote the set of all sampled individuals inside the responding households.

The individuals in $S_{r,\text{ind}}$ are themselves prone to non-response, though it is usually expected to be to a smaller extent. This leads to the observation of a sub-sample of respondents $S_{rr,\text{ind}}$ only. We let r_l denote the response indicator and p_l denote the response probability of the individual l. We suppose that the individuals respond independently of one another. Also, we suppose that this non-response is handled through the method of RHGs: the sample $S_{r,\text{ind}}$ may be partitioned into D RHGs denoted as $S_{r1,\text{ind}}, \ldots, S_{rD,\text{ind}}$ such that the response probability p_l is constant inside a RHG.

For $d = 1, \ldots, D,$ we let p_d denote the common response probability inside the RHG $S_{rd,\text{ind}}$. It is estimated by

$$\hat{p}_d = \frac{\sum_{l \in S_{rd,\text{ind}}} \theta_l r_l}{\sum_{l \in S_{rd,\text{ind}}} \theta_l}, \tag{2.25}$$

with θ_l some weight attached to the individual l. The choice $\theta_l = 1$ leads to estimating p_d by the unweighted response rate inside the RHG. The choice $\theta_l = d_l$ leads to estimating p_d by the response rate inside the RHG, weighted by the individual sampling weights. The choice $\theta_l = d_{rl}$ leads to estimating p_d by the response rate inside the RHG, weighted by the individual sampling weights corrected of household unit non-response. We compare these different choices in the simulation study performed in Section 4.

Accounting for the estimated response probabilities leads to the individual weights corrected for household/individual non-response

$$d_{rrl} = \frac{d_{rl}}{\hat{p}_{d(l)}} \quad \text{with } d(l) \text{ the household containing } l. \tag{2.26}$$

The estimator of Y_{ind} adjusted for household/individual non-response is

$$\hat{Y}_{rr,\text{ind}} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl} y_l. \tag{2.27}$$

2.2.3 Calibration

We let z_l denote the vector of calibration variables at the individual level, and Z_{ind} denote the total on the population U_{ind}. For the sample $S_{rr,\text{ind}},$ this leads to the linear calibrated weights

$$w_l = d_{rrl}\left(1 + z_l^{\mathrm{T}} \lambda_{\text{ind}}\right),$$

with

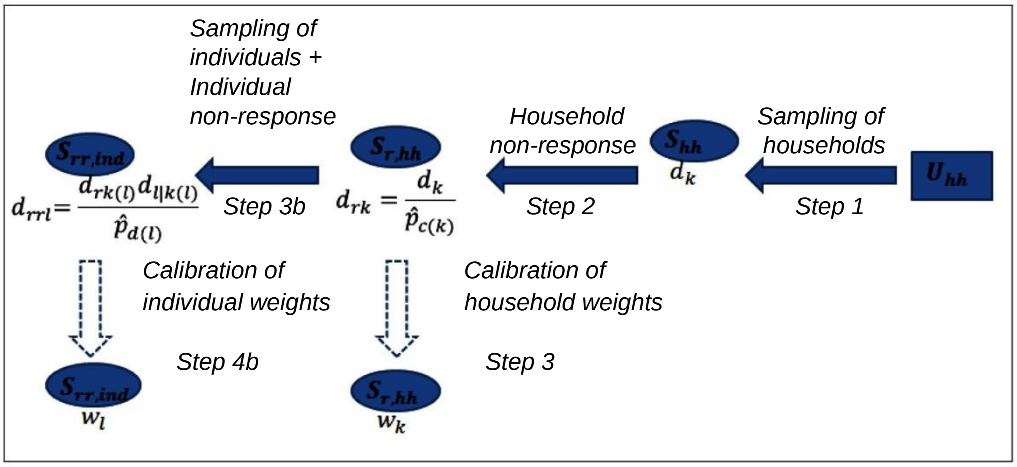
$$\lambda_{\text{ind}} = \left(\sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l z_l^{\mathrm{T}} \right)^{-1} \left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}} \right), \tag{2.28}$$

and where $\hat{Z}_{rr,\text{ind}}$ is the estimator of Z_{ind}, obtained by plugging z_l into (2.27). The calibrated estimator is

$$\hat{Y}_{\text{cal,ind}} = \sum_{l \in S_{rr,\text{ind}}} w_l y_l. \tag{2.29}$$

The sampling and estimation steps are summarized in Figure 2.3.

Figure 2.3 Sampling and estimation steps for a household sample with sub-sampling of individuals.


Sampling of individuals + Individual non-response
Household non-response
Sampling of households
$S_{rr,ind}$
$d_{rrl} = \frac{d_{rk(l)} d_{l|k(l)}}{\hat{p}_{d(l)}}$
Step 3b
$S_{r,hh}$
$d_{rk} = \frac{d_k}{\hat{p}_{c(k)}}$
Step 2
S_{hh}
d_k
Step 1
U_{hh}
Calibration of individual weights
Calibration of household weights
Step 4b
Step 3
$S_{rr,ind}$
w_l
$S_{r,hh}$
w_k

2.2.4 Computation of individual weights on an example

We continue the example initiated in Section 2.1.4. Recall that the sample of responding households is $S_{r,hh} = \{A, D, E, F, H, I, J\}$. The set of all individuals inside the responding households is as follows (say):

$$\underbrace{(i_1, i_2, i_3)}_{A} \quad \underbrace{(i_4)}_{D} \quad \underbrace{(i_5, i_6)}_{E} \quad \underbrace{(i_7, i_8, i_9)}_{F} \quad \underbrace{(i_{10}, i_{11})}_{H} \quad \underbrace{(i_{12})}_{I} \quad \underbrace{(i_{13})}_{J}. \tag{2.30}$$

We suppose that the sampling design consists in selecting one individual exactly inside each household. The set $S_{r,\text{ind}}$ of all sampled individuals inside the responding households is

$$S_{r,\text{ind}} = \{i_1, i_4, i_6, i_8, i_{11}, i_{12}, i_{13}\}. \tag{2.31}$$

From equations (2.23) and (2.16), the individual weights corrected for household non-response are therefore

$$d_{r1} = \frac{40}{3},\ d_{r4} = \frac{72}{13},\ d_{r6} = \frac{576}{13},\ d_{r8} = \frac{160}{3},\ d_{r11} = \frac{576}{13},\ d_{r12} = \frac{288}{13},\quad d_{r13} = \frac{160}{9}. \tag{2.32}$$

Among these 7 selected individuals, 4 only are surveyed due to non-response, accounted for by using the method of Response Homogeneity Groups (RHGs). We suppose that there are two RHGs: the units i_1, i_6, i_8 and i_{11} in the first one, and the units i_4, i_{12} and i_{13} in the second one. The units i_4, i_{11} and i_{13} are non-respondents. Inside each RHG, we compute unweighted estimated response probabilities $(\theta_l = 1)$. This leads to

$$\hat{p}_1 = \frac{\sum_{l \in S_{r1,\text{ind}}} r_l}{\sum_{l \in S_{r1,\text{ind}}} 1} = \frac{3}{4},$$
$$\hat{p}_2 = \frac{\sum_{l \in S_{r2,\text{ind}}} r_l}{\sum_{l \in S_{r2,\text{ind}}} 1} = \frac{1}{3}. \tag{2.33}$$

The weights accounting for household/individual non-response are obtained for the respondents by dividing the weights in (2.32) by the estimated response probabilities. This leads to the weights

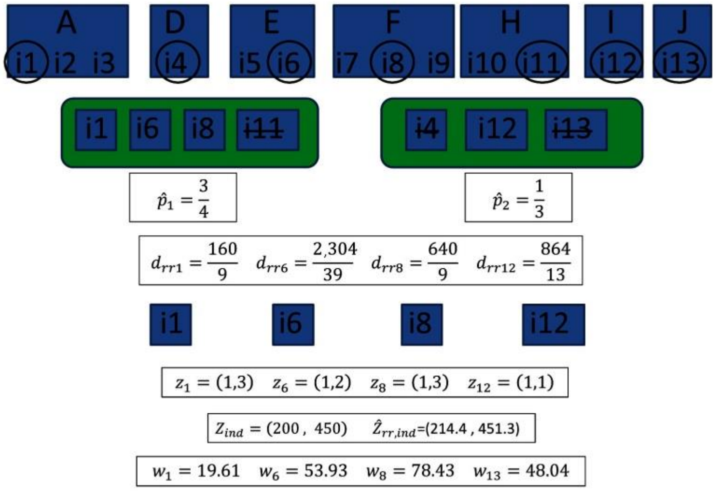
$$d_{rr1} = \frac{160}{9}, \quad d_{rr6} = \frac{2{,}304}{39}, \quad d_{rr8} = \frac{640}{9}, \quad d_{rr12} = \frac{864}{13}. \tag{2.34}$$

Finally, the weights are calibrated to match the population size $N_{\text{ind}} = 200$ and an auxiliary total $Z_{1,\text{ind}} = 450$. Note that, using the sample of respondents, we obtain $\hat{N}_{r,\text{ind}} = 214.4$ and $\hat{Z}_{1r,\text{ind}} = 451.3$. The calibrated weights are

$$w_1 = 19.61, \quad w_6 = 53.93, \quad w_8 = 78.43, \quad w_{13} = 48.04. \tag{2.35}$$

The sampling and estimation steps are summarized in Figure 2.4.

Figure 2.4 Estimation steps for the weighting of individuals.


A
i1 i2 i3
D
i4
E
i5 i6
F
i7 i8 i9
H
i10 i11
I
i12
J
i13
i1 i6 i8 ~~i11~~
~~i4~~ i12 ~~i13~~
$\hat{p}_1 = \frac{3}{4}$
$\hat{p}_2 = \frac{1}{3}$
$d_{rr1} = \frac{160}{9}$ $d_{rr6} = \frac{2,304}{39}$ $d_{rr8} = \frac{640}{9}$ $d_{rr12} = \frac{864}{13}$
i1 i6 i8 i12
$z_1 = (1,3)$ $z_6 = (1,2)$ $z_8 = (1,3)$ $z_{12} = (1,1)$
$Z_{ind} = (200, 450)$ $\hat{Z}_{rr,ind} = (214.4, 451.3)$
$w_1 = 19.61$ $w_6 = 53.93$ $w_8 = 78.43$ $w_{13} = 48.04$

3. Bootstrap variance estimation

We begin in Section 3.1 with the description of the basic step of the bootstrap method when a sample of households only is selected. An illustration is given in Section 3.2 on the example initiated in

Section 2.1.4. The bootstrap method when individuals are sampled inside the households is described in Section 3.3, and an illustration is given in Section 3.4. In Section 3.5, we explain how the basic step of the proposed bootstrap method is used to perform variance estimation and to produce confidence intervals.

3.1 Basic step of the bootstrap for households

Using the with-replacement bootstrap, we first draw inside the original sample S_{hh}^h selected in the stratum U_{hh}^h a with-replacement resample S_{hh*}^h of $n_h - 1$ households, with equal probabilities. Note that the resampling is performed on the sampling unit (a household) rather than on the final unit of observation (an individual), which is key to correctly capture the sampling variance. In particular, this bootstrap method enables to capture the variance due to the second-stage sampling (selection of individuals) without resampling the final units in the bootstrap process. For any $k \in S_{hh}^h,$ we define the reweighting adjustment factor

$$G_k = \frac{n_h}{n_h - 1} \times m_k, \tag{3.1}$$

with m_k the number of times the household k is selected in the resample $S_{hh*}^h,$ a.k.a. the multiplicity. Note that some unit $k \in S_{hh}^h$ may not appear in the resample, in which case this unit has multiplicity zero; see Section 3.2 for an example. The reweighting adjustment factors G_k are used to obtain the bootstrap weights accounting for the sampling design, for unit non-response and for the calibration, as described in Algorithm 1. The steps refer to Figure 2.1. The resampling presented in Algorithm 1 is then repeated B times independently for variance estimation and/or to produce a confidence interval, see Algorithm 3 in Section 3.5.

Algorithm 1. Computation of bootstrap household weights accounting for non-response and calibration

- Step 1: we account for the sampling of households by computing, for any $k \in S_{hh},$ the bootstrap sampling weight

  $$d_{k*} = G_k d_k. \tag{3.2}$$

  The bootstrap version of the full-response estimator given in (2.3) is

  $$\hat{Y}_{hh*} = \sum_{k \in S_{hh}} d_{k*} y_k. \tag{3.3}$$

- Step 2: we account for household unit non-response by computing the bootstrap estimated probabilities inside the RHGs

  $$\hat{p}_{c*} = \frac{\sum_{k \in S_{c,hh}} G_k \theta_k r_k}{\sum_{k \in S_{c,hh}} G_k \theta_k}, \tag{3.4}$$

and we compute the bootstrap weights corrected for non-response

$$d_{rk*} = \frac{d_{k*}}{\hat{p}_{c(k)*}}, \tag{3.5}$$

with $c(k)$ the RHG containing the household k. The bootstrap version of the estimator corrected for unit non-response given in (2.7) is

$$\hat{Y}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} y_k. \tag{3.6}$$

- Step 3: we account for the calibration by calibrating the weights d_{rk*} on the totals X_{hh}. This leads to the bootstrap calibrated weights

  $$w_{k*} = d_{rk*}\left(1 + x_k^{\top} \lambda_{hh*}\right), \tag{3.7}$$

  with

  $$\lambda_{hh*} = \left(\sum_{k \in S_{r,hh}} d_{rk*} x_k x_k^{\top}\right)^{-1} \left(X_{hh} - \hat{X}_{r,hh*}\right)$$

  and

  $$\hat{X}_{r,hh*} = \sum_{k \in S_{r,hh}} d_{rk*} x_k.$$

  The bootstrap version of the calibrated estimator given in (2.10) is

  $$\hat{Y}_{\text{cal},hh*} = \sum_{k \in S_{r,hh}} w_{k*} y_k. \tag{3.8}$$

The treatment of unit non-response in the bootstrap process deserves some explanations. Firstly, our approach is conditional on the response indicators r_k. Contrarily to the sample membership indicators which are bootstrapped at Step 1 of Algorithm 1, the response indicators remain fixed in the bootstrap process. This is due to the fact that we aim at reproducing a variance estimator which considers the sample S_{hh} as selected with replacement, and that in such case bootstrapping the r_k's is not needed. Secondly, accounting for unit non-response at Step 2 of Algorithm 1 is performed conditionally on the RHGs: we do not bootstrap the process leading to the building of the RHGs (e.g., Girard, 2009; Haziza and Beaumont, 2017). Finally, bootstrapping the response probabilities as described in equation (3.4) accounts for the estimation of the response probabilities p_c. In other words, we use within each resample the same RHGs identified on the basis of the sample, but the non-response adjustments inside the RHGs are based on a resample's content. This is illustrated in the example developed in Section 3.2. If we do not bootstrap the response probabilities and directly plug in equation (3.5) the original estimated probabilities $\hat{p}_c$, then the response probabilities are treated as if they were known, which usually results in an overestimation of the variance (Beaumont, 2005; Kim and Kim, 2007).

Now, we discuss bootstrap variance estimation for calibrated estimators, as considered in Step 3 of Algorithm 1 where the calibration step is performed on the true population total X_{hh}. Following the bootstrap principle which states that the sample S_{hh} is to the bootstrap sample S_{hh*} what the population U_{hh} is to the sample S_{hh}, it could seem more intuitive to rather calibrate on the estimated totals $\hat{X}_{hh}$ obtained by plugging x_k into equation (2.3). Both approaches seem valid for bootstrap variance estimation for the calibrated estimator $\hat{Y}_{\text{cal},hh}$, but the calibration variables x_k may be prone to non-response on the sample S_{hh}, making the estimator $\hat{X}_{hh}$ not possible to compute, while the total X_{hh} is known from an external source.

3.2 An example of computation of bootstrap household weights

We continue with the example initiated in Section 2.1.4. The bootstrap is performed by first selecting a resample of $n_{hh} - 1 = 9$ households, with replacement and with equal probabilities, among the original sampled households. In this example, we suppose that the household A is selected three times, that the household G is selected twice, and that the households D, E, H and I are selected once. Making use of equation (3.2), this leads to the bootstrap sampling weights

$$d_{A*} = \frac{40}{3} \quad d_{D*} = \frac{40}{9} \quad d_{E*} = d_{H*} = d_{I*} = \frac{160}{9} \quad d_{G*} = \frac{320}{9}. \tag{3.9}$$

The bootstrap sampling weights are corrected for non-response in the same way than in the original correction of non-response: using the same RHGs, and weighted estimated probabilities. In this case, the first RHG contains only the unit A which is a respondent, so that $\hat{p}_{1*} = 1$. The second RHG contains D, E, G (non-respondent), H and I. This leads to

$$\hat{p}_{2*} = \frac{d_{D*} + d_{E*} + d_{H*} + d_{I*}}{d_{D*} + d_{E*} + d_{G*} + d_{H*} + d_{I*}} = \frac{13}{21}, \tag{3.10}$$

and to the bootstrap weights corrected for non-response

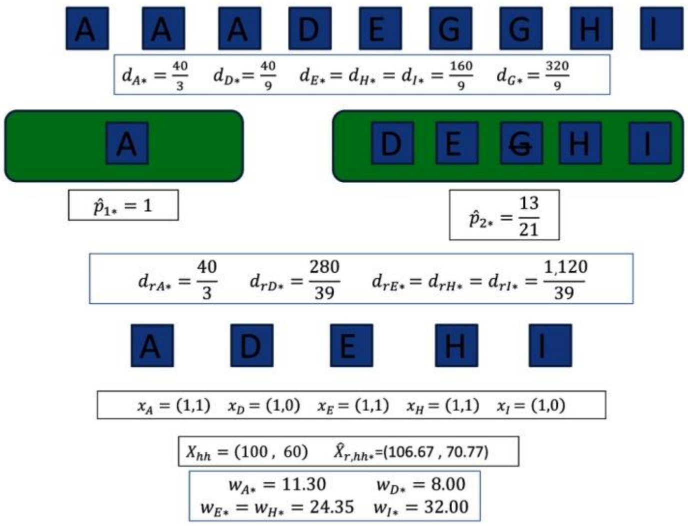
$$d_{rA*} = \frac{40}{3} \quad d_{rD*} = \frac{280}{39} \quad d_{rE*} = d_{rH*} = d_{rI*} = \frac{1{,}120}{39}. \tag{3.11}$$

Finally, the weights are calibrated to match the population size $N_{hh} = 100$ and the auxiliary total $X_{1,hh} = 60$. This leads to the bootstrap calibrated weights

$$w_{A*} = 11.30 \quad w_{D*} = 8.00 \quad w_{E*} = w_{H*} = 24.35 \quad w_{I*} = 32.00. \tag{3.12}$$

The computation of the bootstrap weights is summarized in Figure 3.1.

Figure 3.1 Computation of bootstrap household weights.


A A A D E G G H I
$d_{A*} = \frac{40}{3}$ $d_{D*} = \frac{40}{9}$ $d_{E*} = d_{H*} = d_{I*} = \frac{160}{9}$ $d_{G*} = \frac{320}{9}$
A
D E G H I
$\hat{p}_{1*} = 1$
$\hat{p}_{2*} = \frac{13}{21}$
$d_{rA*} = \frac{40}{3}$ $d_{rD*} = \frac{280}{39}$ $d_{rE*} = d_{rH*} = d_{rI*} = \frac{1{,}120}{39}$
A D E H I
$x_A = (1,1)$ $x_D = (1,0)$ $x_E = (1,1)$ $x_H = (1,1)$ $x_I = (1,0)$
$X_{hh} = (100, 60)$ $\hat{X}_{r,hh*} = (106.67, 70.77)$
$w_{A*} = 11.30$ $w_{D*} = 8.00$
$w_{E*} = w_{H*} = 24.35$ $w_{I*} = 32.00$

3.3 Computation of bootstrap weights for individuals

The computation of the bootstrap weights accounting for the sampling design, for household/individual non-response and for calibration is described in Algorithm 2. The steps refer to Figure 2.3. In addition to the bootstrap steps in Algorithm 1, note that Algorithm 2 involves bootstrapping the computation of response individual probabilities only. Note that the sub-sampling of individuals inside households does not need to be bootstrapped, as discussed in Section 3.1.

Algorithm 2. Computation of bootstrap individual weights accounting for non-response of households, for non-response of individuals and for calibration

- Perform Steps 1 and 2 of Algorithm 1. The bootstrap weights of households corrected for non-response are d_{rk}^*, as given in equation (3.5).
- Step 3b: we first account for the sampling of individuals by computing the bootstrap individual weights corrected for household unit non-response

$$d_{rl*} = d_{rk(l)*} d_{l|k(l)} \quad \text{with } k(l) \text{ the household containing } l. \tag{3.13}$$

We then account for individual unit non-response. We compute the bootstrap estimated probabilities inside the RHGs

$$\hat{p}_{d*} = \frac{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l r_l}{\sum_{l \in S_{rd,\text{ind}}} G_{k(l)} \theta_l}. \tag{3.14}$$

We compute the bootstrap weights of individuals corrected for household/individual non-response, namely

$$d_{rrl*} = \frac{d_{rl*}}{\hat{p}_{d(l)*}}, \tag{3.15}$$

with $d(l)$ the RHG containing the individual l. The bootstrap version of the estimator corrected for unit non-response given in (2.27) is

$$\hat{Y}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} y_l. \tag{3.16}$$

- Step 4b: we account for the calibration by calibrating the weights d_{rrl*} on the totals Z_{ind}. This leads to the bootstrap calibrated weights

$$w_{l*} = d_{rrl*}\left(1 + z_l^\top \lambda_{\text{ind}*}\right), \tag{3.17}$$

with

$$\lambda_{\text{ind}*} = \left(\sum_{k \in S_{rr,\text{ind}}} d_{rrl*} z_l z_l^\top \right)^{-1} \left(Z_{\text{ind}} - \hat{Z}_{rr,\text{ind}*} \right)$$

and

$$\hat{Z}_{rr,\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} d_{rrl*} z_l.$$

The bootstrap version of the calibrated estimator given in (2.29) is

$$\hat{Y}_{\text{cal},\text{ind}*} = \sum_{l \in S_{rr,\text{ind}}} w_{l*} y_l. \tag{3.18}$$

3.4 An example of computation of bootstrap individual weights

We continue with the example in Section 3.2. The bootstrap sample of households is constituted of A (three times), G (two times), and $D,$ $E,$ H and I (one time). Due to household non-response, we observe $A,$ $D,$ $E,$ H and I only. From (2.30), this results in the bootstrap sample of individuals

$$S_{r,\text{ind}*} = \{i_1, i_4, i_6, i_{11}, i_{12}\}. \tag{3.19}$$

The bootstrap weights of households corrected for unit non-response are given in equation (3.11). From equation (3.13), the bootstrap weights of individuals adjusted for household non-response are

$$d_{r1*} = 40 \quad d_{r4*} = \frac{280}{39} \quad d_{r6*} = \frac{2{,}240}{39} \quad d_{r11*} = \frac{2{,}240}{39} \quad d_{r12*} = \frac{1{,}120}{39}. \tag{3.20}$$

These bootstrap weights are corrected for individual non-response in the same way than in the original correction of individual non-response: using the same RHGs and unweighted estimated probabilities. However, we need to account in these probabilities for the multiplicity m_k and the reweighting adjustment factor $G_k,$ see equation (3.1). In our case, the first RHG contains the individuals $i_1,$ i_6 and $i_{11},$ and i_{11} is a non-respondent. The individual i_1 belongs to the household $A,$ which has been selected three times $(m_A = 3)$ in the bootstrap sample. The individual i_6 belongs to the household $E,$ and the individual i_{11} belongs to the household $H,$ which have both been selected one time in the bootstrap sample $(m_E = m_H = 1).$ The computation is similar for the second RHG, and leads to

$$\begin{aligned} \hat{p}_{1*} &= \frac{G_A + G_E}{G_A + G_E + G_H} = \frac{4}{5}, \\ \hat{p}_{2*} &= \frac{G_I}{G_D + G_I} = \frac{1}{2}, \end{aligned} \tag{3.21}$$

and to the bootstrap individuals weights corrected for household/individual non-response

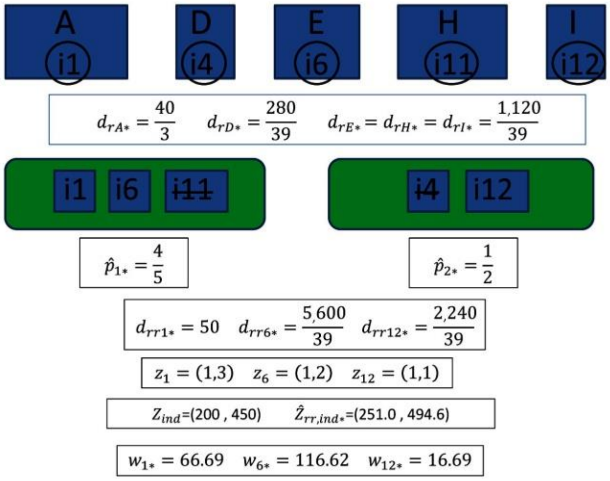
$$d_{rr1*} = 50 \quad d_{r6*} = \frac{5{,}600}{39} \quad d_{r12*} = \frac{2{,}240}{39}. \tag{3.22}$$

Finally, the weights are calibrated to match the population size $N_{\text{ind}} = 200$ and the auxiliary total $Z_{1,\text{ind}} = 450.$ This leads to the bootstrap calibrated weights

$$w_{1*} = 66.69 \quad w_{6*} = 116.62 \quad w_{12*} = 16.69. \tag{3.23}$$

The computation of bootstrap individual weights is summarized in Figure 3.2.

Figure 3.2 Computation of bootstrap individual weights.


A
i1
D
i4
E
i6
H
i11
I
i12
$d_{rA*} = \frac{40}{3}$
$d_{rD*} = \frac{280}{39}$
$d_{rE*} = d_{rH*} = d_{rI*} = \frac{1{,}120}{39}$
i1
i6
~~i11~~
~~i4~~
i12
$\hat{p}_{1*} = \frac{4}{5}$
$\hat{p}_{2*} = \frac{1}{2}$
$d_{rr1*} = 50$
$d_{rr6*} = \frac{5{,}600}{39}$
$d_{rr12*} = \frac{2{,}240}{39}$
$z_1 = (1,3)$
$z_6 = (1,2)$
$z_{12} = (1,1)$
Z_{ind}=(200 , 450)
$\hat{Z}_{rr,ind*}$=(251.0 , 494.6)
$w_{1*} = 66.69$
$w_{6*} = 116.62$
$w_{12*} = 16.69$

3.5 Bootstrap variance estimation and confidence intervals

In this section, we are interested in parameters which may be written as smooth functions of totals. We explain how the basic step of the proposed bootstrap method is used to perform variance estimation and to produce confidence intervals. For brevity, we focus on parameters defined over the population of households $U_{hh}.$ The treatment for parameters of interest in the population of individuals U_{ind} is similar.

Suppose that y_k is a q-vector of interest variables, and that we are interested in some parameter $\beta_{hh} = f(Y_{hh})$ with $f: \mathbb{R}^q \to \mathbb{R}$ a known, smooth function. In case of full response, the substitution estimator of β_{hh} is

$$\hat{\beta}_{hh} = f(\hat{Y}_{hh}), \tag{3.24}$$

see for example Deville (1999). In case of unit non-response at the household level, the estimator of β_{hh} corrected for unit non-response is

$$\hat{\beta}_{r,hh} = f(\hat{Y}_{r,hh}), \tag{3.25}$$

and the calibrated estimator of β_{hh} is

$$\hat{\beta}_{\text{cal},hh} = f(\hat{Y}_{\text{cal},hh}). \tag{3.26}$$

In each case, a bootstrap variance estimator is obtained by applying a large number of times (say B) the basic step of the bootstrap method in Algorithm 1, and then by computing the dispersion of the bootstrap estimators. This is summarized in Algorithm 3.

Algorithm 3. Bootstrap variance estimation for an estimation over the population of households

1. Repeat B times the bootstrap procedure described in Algorithm 1. Let us denote $\hat{Y}^b_{hh*}$, $\hat{Y}^b_{r,hh*}$ and $\hat{Y}^b_{\text{cal},hh*}$ for the bootstrap estimators of totals computed on the b^{th} sample. Also, let us denote $\hat{\beta}^b_{hh*}$, $\hat{\beta}^b_{r,hh*}$ and $\hat{\beta}^b_{\text{cal},hh*}$ for the associated bootstrap estimators of β_{hh}.
2. The Bootstrap variance estimator for $\hat{\beta}_{hh}$ is

$$\hat{V}_{\text{boot}}(\hat{\beta}_{hh}) = \frac{1}{B-1} \sum_{b=1}^{B} \left\{ \hat{\beta}^b_{hh*} - \frac{1}{B} \sum_{b'=1}^{B} \hat{\beta}^{b'}_{hh*} \right\}^2, \tag{3.27}$$

   and similarly for $\hat{\beta}_{r,hh}$ and $\hat{\beta}_{\text{cal},hh}$.

The bootstrap variance estimator may be used to compute a normality-based confidence interval with targeted level $1-2\alpha$. For example, the confidence interval when using the full-response estimator $\hat{\beta}_{hh}$ is

$$\text{IC}_{\text{nor}}(\beta_{hh}) = \left[\hat{\beta}_{hh} \pm u_{1-\alpha} \left\{ \hat{V}_{\text{boot}}(\hat{\beta}_{hh}) \right\}^{0.5} \right], \tag{3.28}$$

with $u_{1-\alpha}$ the quantile of order $1-\alpha$ of the standard normal distribution. This confidence interval is expected to be conservative, since the proposed bootstrap method is conservative too.

We also consider the percentile and the reverse percentile (a.k.a. basic) bootstrap confidence intervals. They can be directly computed from the bootstrap weights and are therefore attractive from a data user's perspective, unlike more computationally intensive methods like the t-bootstrap (e.g. Davison and Hinkley, 1997; Shao and Tu, 1995). For $\hat{\beta}_{hh}$, the percentile confidence interval is obtained by using the distribution of $\hat{\beta}_{hh*}$ as an approximation of the distribution of $\hat{\beta}_{hh}$. It makes use of the ordered bootstrap estimates $\hat{\beta}_{hh*}^{(1)}, \ldots, \hat{\beta}_{hh*}^{(B)}$ to form the confidence interval

$$\text{IC}_{\text{per}}(\beta_{hh}) = \left[\hat{\beta}_{hh*}^{(L)}, \hat{\beta}_{hh*}^{(U)}\right], \tag{3.29}$$

with targeted level $1 - 2\alpha$, where $L = \alpha B$ and $U = (1 - \alpha)B$. The reverse percentile confidence interval is obtained by viewing the distribution of $(\hat{\beta}_{hh*} - \hat{\beta}_{hh})$ as an approximation of the distribution of $(\hat{\beta}_{hh} - \beta_{hh})$. It leads to the confidence interval

$$\text{IC}_{\text{rev}}(\beta_{hh}) = \left[2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(U)}, 2\hat{\beta}_{hh} - \hat{\beta}_{hh*}^{(L)}\right]. \tag{3.30}$$

The properties of the bootstrap variance estimator and of the three confidence intervals are evaluated in the simulation study performed in Section 4 for the estimation of a total.

Choosing the number B of resamples is an important practical problem. Girard (2009) suggests considering several possible resample sizes (e.g., by increasing B with an increment of 100), and plotting the bootstrap variance estimators in function of B. The value for which this variance estimator starts to stabilize is then retained. This is a simple method, but which may require some compromise solution if different variables of interest lead to different stabilizing values. Beaumont and Patak (2012) suggest choosing B such that with a high probability, the length of the bootstrap confidence interval given in (3.28) is close to the length of the confidence interval obtained with an analytical variance estimator. Under the assumption that conditionally on the original sample, the normalized bootstrap estimator of the total is normally distributed, they establish that the value B may be determined from the distribution of a chi-square variable (Beaumont and Patak, 2012, equation 10). Interestingly, the value obtained does not depend on the variable of interest. Based on these results, they suggest using a value B no smaller than 750, and a larger value if the normality assumption of the bootstrap estimator may fail. We used $B =$ 1,000 in the simulation study presented in the following section. For surveys that are to serve multiple analytical needs – ranging from simple to complex population parameters and various domain sizes – selecting no fewer than 1,000 replicates is the norm given the computing resources available nowadays.

4. Simulation study

In order to evaluate the proposed bootstrap method, we conducted a simulation study on an artificial population. We first generate a population U_{hh} containing $N_{hh} =$ 100,000 households, with four auxiliary variables $x_1, \ldots, x_4$ generated from a gamma distribution with shape and scale parameters 2 and 5. Inside the population, we generate three variables of interest $y_1, \ldots, y_3$ according to the following models

$$\begin{aligned} y_{1k} &= 10 + x_{1k} + x_{2k} + \sigma_{\varepsilon}\varepsilon_k, \\ y_{2k} &= 10 + x_{1k} + x_{3k} + \sigma_{\varepsilon}\varepsilon_k, \\ y_{3k} &= 10 + x_{3k} + x_{4k} + \sigma_{\varepsilon}\varepsilon_k, \end{aligned} \tag{4.1}$$

where ε_k is generated according to a standard normal distribution. We set $\sigma_\varepsilon = 10,$ which results in a coefficient of determination of approximately 0.50 for each model. The auxiliary variables $1, x_{1k}, x_{2k}$ are used as calibration variables at the household level in this simulation study. The three variables of interest therefore correspond to cases when the calibration model is well specified $(y_1),$ partly well specified $(y_2),$ or poorly specified $(y_3).$ The population U_{hh} is randomly split into five response homogeneity groups (RHG) of equal sizes. The response probability p_c inside the RHG c is equal to 0.5 for the first group, 0.6 for the second group, ..., and 0.9 for the fifth group, resulting in an average response rate of 70% for the households.

Inside each household $k,$ we generate N_k individuals, where $N_k - 1$ is generated according to a Poisson distribution with parameter 1, which results in an average number of 2 individuals per household. Inside the corresponding population $U_{\text{ind}},$ we generate four auxiliary variables $z_1, \ldots, z_4$ with shape and scale parameters 2 and 0.5. Also, we generate three variables of interest y_4, y_5, y_6 according to the following models

$$\begin{aligned} y_{4l} &= 5 + 0.5z_{1l} + 0.5z_{2l} + \sigma_\eta \eta_l, \\ y_{5l} &= 5 + 0.5z_{1l} + 0.5z_{3l} + \sigma_\eta \eta_l, \\ y_{6l} &= 5 + 0.5z_{3l} + 0.5z_{4l} + \sigma_\eta \eta_l, \end{aligned} \tag{4.2}$$

where η_l is generated according to a standard normal distribution. We set σ_η = 0.4, which results in a coefficient of determination of approximately 0.6 for each model. The auxiliary variables $1, z_{1l}, z_{2l}$ are used as calibration variables at the individual level in this simulation study. The three variables of interest therefore correspond to a case when the calibration model is well specified $(y_4),$ partly well specified $(y_5),$ or poorly specified $(y_6).$

The population U_{ind} is split into five RHGs as follows. The individuals which are alone in their household form a separate RHG, with a response probability of 1. The rationale behind this choice is that in such case, the individual is somewhat equivalent to his/her household, and that the non-response is modeled at the household level. Among the rest of the individuals living in a household k with $N_k = 2$ individuals or more, the variables z_1 and z_2 are used to form four RHGs of approximately equal size. The response probability p_d ranges from 0.80 to 0.95 in these four remaining RHGs. This results in an overall response rate of approximately 90% for the individuals.

Inside the population $U_{hh},$ we select a sample S_{hh} of n_{hh} = 1,000 households by simple random sampling without replacement. Note that the sampling rate is small (1%), so that simple random sampling with/without replacement are not much different, and the bias of the bootstrap variance estimators is expected to be small under this set-up. The non-response is generated according to the RHG household model, which results in a sample $S_{r,hh}$ of responding households. The estimated response probabilities $\hat{p}_c$ are obtained from equation (2.5), with equal weight $\theta_k = 1.$ Inside each $k \in S_{r,hh},$ one Kish individual is randomly selected with equal probabilities, which results in the sample of individuals $S_{r,\text{ind}}.$ Inside $S_{r,\text{ind}},$ the non-response is generated according to the RHG individual model, resulting in a sample $S_{rr,\text{ind}}$ of responding individuals. The estimated response probabilities $\hat{p}_d$ are obtained from equation (2.25), in three possible ways: equal weights $\theta_l = 1,$ sampling weights $\theta_l = d_l,$ or individuals weights corrected for the household non-response $\theta_l = d_{rl}.$

The sampling and non-response steps are repeated $R =$ 1,000 times. On each sample $S_{hh},$ we compute the full-response estimator given in (2.3), and on each sample $S_{r,hh},$ we compute the estimator adjusted for non-response $\hat{Y}_{r,hh}$ given in (2.7) and the estimator $\hat{Y}_{\text{cal},hh}$ given in (2.10) with the set of calibration variables $x_k = (1, x_{1k}, x_{2k})^\top.$ On each sample $S_{rr,\text{ind}},$ we compute the estimator adjusted for non-response $\hat{Y}_{rr,\text{ind}}$ given in (2.27) and the estimator $\hat{Y}_{\text{cal,ind}}$ given in (2.29) with the set of calibration variables $z_l = (1, z_{1l}, z_{2l})^\top.$ For these five estimators, we compute the normalized root mean square error

$$\text{NRMSE}(\hat{Y}) = 100 \times \frac{\sqrt{\text{MSE}(\hat{Y})}}{Y}, \tag{4.3}$$

with $\text{MSE}(\hat{Y})$ a simulation-based approximation of the mean square error of $\hat{Y},$ obtained from an independent run of 10,000 simulations.

For these five estimators, we also compute the bootstrap variance estimators obtained by applying Algorithm 3 with $B =$ 1,000. So as to measure the bias of a variance estimator $v(\hat{Y}),$ we use the Monte Carlo Percent Relative Bias

$$\text{RB}\{v(\hat{Y})\} = 100 \times \frac{R^{-1}\sum_{c=1}^{R} v_c(\hat{Y}_c) - \text{MSE}(\hat{Y})}{\text{MSE}(\hat{Y})}, \tag{4.4}$$

where $v_c(\hat{Y}_c)$ stands for the variance estimator in the c^{th} sample. As a measure of stability of $v(\hat{Y}),$ we use the Relative Stability

$$\text{RS}\{v(\hat{Y})\} = 100 \times \frac{\left[R^{-1}\sum_{c=1}^{R}\left\{v_c(\hat{Y}_c) - \text{MSE}(\hat{Y})\right\}^2\right]^{1/2}}{\text{MSE}(\hat{Y})}. \tag{4.5}$$

Also, we compute the coverage rates of the confidence interval associated to the percentile Bootstrap, to the basic bootstrap and to the normality-based confidence interval, with nominal one-tailed error rate of 2.5% in each tail.

The results are presented in Table 4.1 for the estimation on the population of households. The normalized root mean square error of the calibrated estimator $\hat{Y}_{\text{cal},hh}$ is smaller when the calibration variables are explanatory for the variable of interest, as expected. We observe a slight positive bias of the bootstrap variance estimator for the full-response estimator $\hat{Y}_{hh},$ but almost no bias for the reweighted estimators $\hat{Y}_{r,hh}$ and $\hat{Y}_{\text{cal},hh}.$ The bootstrap variance estimator is slightly less stable with the reweighted estimators, which is likely due to the additional variability associated to the correction of unit non-response. Concerning the confidence intervals, we note that the coverage rates are well respected in all cases and for the three studied methods.

We now turn to the result on the population of individuals, which are presented in Table 4.2. We observe that the relative bias of the bootstrap variance estimator is very small in all cases. The choice of the weights θ_k used in the estimation of the response probabilities seem to have no effect on the normalized root mean square error of the estimators, but the use of the weights $\theta_l = d_{rl}$ adjusted for household non-response yields slightly more stable variance estimators for $\hat{Y}_{rr,\text{ind}}$. The coverage rates are approximately respected in all cases.

Table 4.1
Coefficient of variation of the estimator of the total, Relative Bias and Relative Stability of the Bootstrap variance estimator, and Nominal One-Tailed Error Rates of the percentile bootstrap and of the basic bootstrap for 3 variables on the population of households

					Percentile bootstrap			Basic bootstrap			Normality-based		
		NRMSE	RB	RS	L	U	L+U	L	U	L+U	L	U	L+U
$\hat{Y}_{hh}$	y_1	1.47	2.48	7.2	2.2	3.1	5.3	2.1	3.3	5.4	2.2	3.2	5.4
	y_2	1.48	0.73	6.6	2.6	3.3	5.9	2.7	3.4	6.1	2.6	3.2	5.8
	y_3	1.48	1.11	6.6	2.6	2.7	5.3	2.7	3.0	5.7	2.4	2.7	5.1
$\hat{Y}_{r,hh}$	y_1	1.82	0.42	8.7	2.4	2.4	4.8	2.3	2.7	5.0	2.3	2.6	4.9
	y_2	1.83	-0.76	8.2	2.7	2.8	5.5	2.5	3.0	5.5	2.2	2.7	4.9
	y_3	1.82	0.72	8.4	2.8	2.1	4.9	2.8	2.2	5.0	2.8	1.9	4.7
$\hat{Y}_{\text{cal},hh}$	y_1	1.29	1.27	8.3	2.4	2.7	5.1	2.8	2.8	5.6	2.8	2.7	5.5
	y_2	1.58	-0.55	8.2	2.5	3.5	6.0	2.8	3.9	6.7	2.8	3.6	6.4
	y_3	1.82	0.49	8.4	2.9	1.8	4.7	3.0	2.2	5.2	2.9	2.0	4.9

Table 4.2
Coefficient of variation of the estimator of the total, Relative Bias and Relative Stability of the Bootstrap variance estimator, and Nominal One-Tailed Error Rates of the percentile bootstrap and of the basic bootstrap for 3 variables on the population of individuals

					Percentile bootstrap			Basic bootstrap			Normality-based		
		NRMSE	RB	RS	L	U	L+U	L	U	L+U	L	U	L+U
		Equal weights $\theta_l = 1$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2.01	0.31	9.6	2.0	3.2	5.2	1.9	3.3	5.2	1.9	3.0	4.9
	y_5	2.02	-0.17	9.6	2.4	3.4	5.8	2.2	3.7	5.9	2.3	3.5	5.8
	y_6	2.02	-0.24	9.6	2.2	3.3	5.5	2.0	3.7	5.7	2.0	3.2	5.2
$\hat{Y}_{\text{cal,ind}}$	y_4	0.29	1.72	10.8	2.1	2.4	4.5	2.1	2.3	4.4	2.1	2.2	4.3
	y_5	0.39	1.04	11.3	2.3	2.5	4.8	2.3	2.5	4.8	2.2	2.4	4.6
	y_6	0.47	1.90	11.2	2.8	2.1	4.9	2.2	2.5	4.7	2.3	2.0	4.3
		Sampling weights $\theta_l = d_l$											
$\hat{Y}_{rr,\text{ind}}$	y_4	2.00	-0.08	9.5	1.8	3.8	5.6	1.7	3.8	5.5	1.7	3.4	5.1
	y_5	2.00	0.14	9.4	1.9	3.3	5.2	2.2	3.5	5.7	1.8	3.5	5.3
	y_6	1.99	0.61	9.3	1.7	3.2	4.9	1.7	3.4	5.1	1.7	3.2	4.9
$\hat{Y}_{\text{cal,ind}}$	y_4	0.29	-0.57	10.3	2.9	2.4	5.3	3.3	2.2	5.5	3.0	2.3	5.3
	y_5	0.39	0.40	11.6	2.4	3.2	5.6	2.7	3.3	6.0	2.3	3.2	5.5
	y_6	0.47	-0.05	11.2	2.3	2.2	4.5	1.8	2.3	4.1	1.8	2.3	4.1
		Weights adjusted for household non-response $\theta_l = d_{rl}$											
$\hat{Y}_{rr,\text{ind}}$	y_4	1.99	-0.71	8.9	2.5	2.3	4.8	2.6	2.7	5.3	2.5	2.4	4.9
	y_5	1.99	-0.82	8.9	3.1	2.2	5.3	2.9	2.5	5.4	2.5	2.2	4.7
	y_6	1.99	-0.26	9.1	3.1	2.3	5.4	3.0	3.0	6.0	2.9	2.5	5.4
$\hat{Y}_{\text{cal,ind}}$	y_4	0.29	1.70	10.6	2.7	3.4	6.1	2.6	3.3	5.9	2.5	3.3	5.8
	y_5	0.39	1.38	11.3	2.1	2.7	4.8	2.2	3.0	5.2	1.7	3.0	4.7
	y_6	0.47	0.61	10.9	2.5	2.8	5.3	2.3	3.0	5.3	2.3	2.8	5.1

5. Application to the French panel for urban policy

In this section, we present an illustration of the proposed methodology on a French panel for urban policy. The sampling design and the estimation steps for the sample of households are briefly described in Section 5.1, and three possible bootstrap confidence intervals are computed. The SAS macro developed to implement the proposed methodology for one-stage sampling is given in Appendix B, along with a small example. The additional sampling and estimation steps for the sample of individuals are described in Section 5.2, and three possible bootstrap confidence intervals are computed. The SAS macro developed to implement the proposed methodology for two-stage sampling is given in Appendix C, along with a small example.

5.1 Sample of households

The Panel for Urban Policy (PUP) is a survey in four waves, conducted between 2011 and 2014 by the French General Secretariat of the Inter-ministerial Committee for Cities (SGCIV). The survey aims at collecting information about security, employment, precariousness, schooling and health, for people living in the Sensitive Urban Zones (ZUS). We are only interested in the 2011 wave of the survey. A sample of households is selected, and all the individuals living in the selected households are theoretically surveyed.

The sample of households is obtained by two-stage sampling, see for example Chauvet (2015); Chauvet and Vallée (2018). Firstly, the population of districts is partitioned into 4 strata, and a global sample of $n_I = 40$ districts is selected by means of probability proportional to size sampling inside strata. A sample of households is then selected at the second-stage inside each selected district by means of simple random sampling, in such a way that the final inclusion probabilities of households are approximately equal inside strata (self-weighted sampling design). For the purpose of illustration, the two-stage selection of the households is not considered here, and the sample of households is viewed as directly selected by means of stratified simple random sampling.

The sample contains 2,971 households, but due to unit non-response only 1,256 households are observed. Non-response is accounted for by using Response Homogeneity Groups, defined with respect to five auxiliary variables: housing construction period, type of dwelling (apartment/house), number of rooms, low-income housing (yes/no), region. By using a logistic regression and the score method (e.g. Haziza and Beaumont, 2007), we obtain 8 response homogeneity groups. The five auxiliary variables used in the definition of the RHGs are also used for calibration.

We are interested in four categorical variables related to security, town planning and residential mobility. The variable y_1 gives the perceived reputation of the district (good, fair, poor, no opinion). The variable y_2 indicates if a member of the household has witnessed trafficking (never, rarely, sometimes, no opinion). The variable y_3 indicates if some significant roadworks have been done in the neighborhood in the twelve last months (yes, no, no opinion). The variable y_4 indicates if the household intends to leave the district during the next twelve months (certainly/probably, certainly not, probably not, no opinion). For each category g of each variable $y,$ we are interested in the proportion

$$\beta_{g,hh} = \frac{\sum_{k\in U_{hh}} 1(y_k = g)}{N_{hh}}, \tag{5.1}$$

with N_{hh} the total number of households. The estimator of β_g adjusted for non-response is

$$\hat{\beta}_{\text{gr},hh} = \frac{\sum_{k\in S_{r,hh}} d_{rk} 1(y_k = g)}{\sum_{k\in S_{r,hh}} d_{rk}}, \tag{5.2}$$

see equation (2.7). The calibrated estimator of β_g is

$$\hat{\beta}_{\text{gcal},hh} = \frac{\sum_{k\in S_{r,hh}} w_k 1(y_k = g)}{\sum_{k\in S_{r,hh}} w_k}, \tag{5.3}$$

see equation (2.10).

For each proportion, we give the normality-based confidence interval making use of the bootstrap variance estimator, the percentile bootstrap and the basic bootstrap confidence intervals, see Section 3.5. We use the with-replacement Bootstrap presented in Algorithm 1 with $B = 1{,}000$ resamples. The results with a nominal one-tailed error rate of 2.5% are presented in Table 5.1. The three confidence intervals are very similar in all cases.

Table 5.1
Estimation of the marginal proportions with three confidence intervals for four variables on interest

	Perceived reputation of district status							
	Estimator adj. for non-response				Calibration estimator			
	Good	Fair	Poor	No opinion	Good	Fair	Poor	No opinion
Estim.	0.217	0.225	0.531	0.027	0.217	0.224	0.532	0.027
Norm. CI	[0.194,0.241]	[0.201,0.249]	[0.503,0.559]	[0.018,0.036]	[0.193,0.240]	[0.200,0.248]	[0.504,0.560]	[0.018,0.036]
Perc. CI	[0.195,0.241]	[0.201,0.251]	[0.504,0.558]	[0.019,0.036]	[0.193,0.240]	[0.201,0.251]	[0.505,0.560]	[0.019,0.036]
Basic CI	[0.193,0.240]	[0.200,0.249]	[0.503,0.557]	[0.018,0.035]	[0.193,0.240]	[0.198,0.248]	[0.504,0.559]	[0.018,0.035]
	Witnessed trafficking							
	Estimator adj. for non-response				Calibration estimator			
	Never	Rarely	Sometimes	No opinion	Never	Rarely	Sometimes	No opinion
Estim.	0.599	0.065	0.155	0.181	0.606	0.065	0.156	0.173
Norm. CI	[0.571,0.627]	[0.050,0.079]	[0.135,0.175]	[0.161,0.201]	[0.581,0.632]	[0.050,0.079]	[0.135,0.176]	[0.159,0.188]
Perc. CI	[0.572,0.628]	[0.050,0.080]	[0.134,0.175]	[0.161,0.201]	[0.582,0.633]	[0.051,0.080]	[0.134,0.175]	[0.160,0.188]
Basic CI	[0.570,0.626]	[0.049,0.078]	[0.136,0.176]	[0.161,0.201]	[0.579,0.630]	[0.049,0.078]	[0.136,0.177]	[0.159,0.187]
	Roadworks in neighborhood							
	Estimator adj. for non-response				Calibration estimator			
	Yes	No	No opinion		Yes	No	No opinion	
Estim.	0.471	0.495	0.034		0.470	0.496	0.034	
Norm. CI	[0.444,0.498]	[0.468,0.523]	[0.024,0.044]		[0.443,0.496]	[0.469,0.523]	[0.024,0.045]	
Perc. CI	[0.442,0.496]	[0.469,0.524]	[0.025,0.045]		[0.440,0.495]	[0.470,0.524]	[0.025,0.045]	
Basic CI	[0.445,0.500]	[0.466,0.522]	[0.023,0.043]		[0.444,0.499]	[0.468,0.522]	[0.024,0.044]	
	Intention to leave the district							
	Estimator adj. for non-response				Calibration estimator			
	Cert./Prob.	Prob. not	Cert. not	No opinion	Cert./Prob.	Prob. not	Cert. not	No opinion
Estim.	0.286	0.130	0.548	0.036	0.287	0.131	0.546	0.036
Norm. CI	[0.260,0.312]	[0.111,0.149]	[0.520,0.576]	[0.025,0.047]	[0.261,0.313]	[0.112,0.150]	[0.518,0.573]	[0.025,0.047]
Perc. CI	[0.260,0.313]	[0.111,0.149]	[0.521,0.576]	[0.026,0.047]	[0.261,0.313]	[0.113,0.151]	[0.520,0.574]	[0.026,0.048]
Basic CI	[0.259,0.312]	[0.111,0.149]	[0.520,0.575]	[0.025,0.046]	[0.261,0.313]	[0.111,0.149]	[0.517,0.572]	[0.025,0.047]

5.2 Sample of individuals

The sample of responding households contains 3,098 individuals who are theoretically surveyed, but due to unit non-response we observe a subset of 2,804 individual respondents only. Non-response is accounted for by using Response Homogeneity Groups, defined with respect to eight auxiliary variables: three at the individual level (sex, age, nationality), and five at the dwelling level (housing construction period, type of dwelling, number of rooms, low-income housing or not, region). By using a logistic regression and the score method, we obtain 8 response homogeneity groups. The three individual auxiliary variables used in the definition of the RHGs are also used for calibration.

We are interested in three variables of interest. The variable y_5 is quantitative, and gives the number of children. The variable y_6 indicates whether the individual has one or several jobs (one, several, none, no answer). The variable y_7 indicates whether the individual benefits from a complementary full medical cover (yes, no, no answer). For the variable $y_5,$ we compute the estimator of the total adjusted for non-reponse and the calibrated estimator given in equations (2.27) and (2.29), respectively. For the two other variables of interest and for each category $g,$ we are interested in the proportion

$$\beta_{g,\text{ind}} = \frac{\sum_{l \in U_{\text{ind}}} 1(y_k = g)}{N_{\text{ind}}}, \tag{5.4}$$

with N_{ind} the total number of individuals. The estimator of $\beta_{g,\text{ind}}$ adjusted for non-response is

$$\hat{\beta}_{\text{grr},\text{ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} d_{rrl} 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} d_{rrl}}, \tag{5.5}$$

see equation (2.27). The calibrated estimator of $\beta_{g,\text{ind}}$ is

$$\hat{\beta}_{\text{gcal},\text{ind}} = \frac{\sum_{l \in S_{rr,\text{ind}}} w_l 1(y_l = g)}{\sum_{l \in S_{rr,\text{ind}}} w_l}, \tag{5.6}$$

see equation (2.29).

For each parameter, we give the normality-based confidence interval making use of the bootstrap variance estimator, the percentile bootstrap and the basic bootstrap confidence intervals. We use the with-replacement Bootstrap presented in Algorithm 2 with $B = 1{,}000$ resamples. The results with a nominal one-tailed error rate of 2.5% are presented in Table 5.2. The three confidence intervals are very similar in all cases.

Table 5.2
Estimation of the marginal proportions with three confidence intervals for four variables on interest

	Number of children							
	Estimator adj. for non-response				Calibration estimator			
Estim. ($\times 10^6$)	4.40				4.39			
Norm. CI	[4.15,4.64]				[4.21,4.58]			
Perc. CI	[4.16,4.65]				[4.21,4.58]			
Basic CI	[4.14,4.63]				[4.20,4.57]			
	Does the individual have several jobs?							
	Estimator adj. for non-response				Calibration estimator			
	One	Several	None	No answer	One	Several	None	No answer
Estim.	0.304	0.016	0.372	0.308	0.305	0.016	0.372	0.307
Norm. CI	[0.286,0.323]	[0.011,0.021]	[0.352,0.392]	[0.290,0.326]	[0.285,0.325]	[0.011,0.021]	[0.350,0.394]	[0.283,0.332]
Perc. CI	[0.287,0.323]	[0.011,0.021]	[0.351,0.393]	[0.289,0.326]	[0.284,0.325]	[0.011,0.020]	[0.351,0.393]	[0.284,0.333]
Basic CI	[0.286,0.322]	[0.011,0.020]	[0.351,0.393]	[0.289, 0.326]	[0.285,0.325]	[0.011,0.020]	[0.352,0.393]	[0.282,0.330]
	Complementary full medical cover							
	Estimator adj. for non-response				Calibration estimator			
	Yes	No	No answer		Yes	No	No answer	
Estim.	0.122	0.626	0.252		0.122	0.627	0.251	
Norm. CI	[0.106,0.137]	[0.603,0.650]	[0.234,0.270]		[0.105,0.138]	[0.604,0.650]	[0.227,0.275]	
Perc. CI	[0.105,0.137]	[0.603,0.651]	[0.235,0.269]		[0.105,0.138]	[0.604,0.650]	[0.230,0.276]	
Basic CI	[0.106,0.138]	[0.602,0.649]	[0.235,0.269]		[0.105,0.138]	[0.605,0.651]	[0.227,0.273]	

6. Conclusion and future work

In this paper, we have explained how the with-replacement bootstrap may be applied to household surveys, in order to account for the whole variability of the sampling process including sampling and non-response, and to a posteriori adjustments like calibration. The methods have been illustrated on a toy example for clarity of exposition, evaluated via a simulation study and applied to a French panel for urban policy. To make the implementation of the method easier for users, we have developed two SAS macros which are available upon request to the corresponding author.

The results in the simulation study show that both the bootstrap variance estimators and three bootstrap confidence intervals work well in case of a small sampling fraction. If the sampling fraction is larger, the bootstrap variance estimator is known to be conservative, and the normality-based confidence interval is therefore expected to be conservative as well. However, the coverage properties of the two other confidence intervals in such context remain unclear. This is an interesting matter for further research.

In this paper, we focused on applying the bootstrap for variance estimation, after the statistical adjustments (treatment of unit non-response and calibration) have been performed by the survey methodologist. Bootstrap may also be used a priori, as a diagnosis tool to evaluate the relevance of possible statistical adjustments. For example, it may be tempting to use a large number of Response Homogeneity Groups (RHGs) to correct unit non-response, so as to reduce the non-response bias. However, this may result in an increased variability of the reweighted estimators. Bootstrap may be used to evaluate several possible sets of RHGs, for example by producing histograms of the bootstrap non-response adjustments and/or of the bootstrap estimators corrected for unit non-response, to give some insight on the stability of estimation with a possible set of RHGs. This is helpful in finding a bias/variance trade-off. This approach in mentioned in Girard (2009), and is an important matter for further work.

We have considered the situation when the survey is performed at one time only. If we wish to perform longitudinal estimations, units are typically followed over time. If we are also interested in cross-sectional estimations at several times, additional samples are selected at posterior waves and mixed with the original sample. Bootstrap variance estimation in the context of longitudinal surveys is a very important matter for further investigation.

Acknowledgements

The authors are grateful to the associate editor and two referees for numerous suggestions on a previous version of the paper, which have helped us in significantly improving the article.

Appendix

A. Benchmark variance estimators for the sample of individuals

We first consider the estimator $\hat{Y}_{\text{ind}}$ in equation (2.21), that we use in case of full response. The benchmark variance estimator is

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k \hat{y}_k - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} \hat{y}_{k'} \right)^2, \tag{A.1}$$

with

$$\hat{y}_k = \sum_{l \in S_{\text{ind},k}} d_{l|k} y_l.$$

We now consider the estimator $\hat{Y}_{rr,\text{ind}}$ given in equation (2.27), which is adjusted for the non-response of both households and individuals. The benchmark variance estimator is

$$v_{\text{mult}}(\hat{Y}_{\text{ind}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in S_{hh}^h} \left(d_k v_{1k} - \frac{1}{n_h} \sum_{k' \in S_{hh}^h} d_{k'} v_{1k'} \right)^2, \tag{A.2}$$

where

$$v_{1k} = \hat{u}_{1k} + u_{3k},$$

where the first linearized variable $\hat{u}_{1k}$ is similar to that given in equation (2.8), while the second linearized variable u_{3k} accounts for the estimation of the individual response probabilities. We have for the first linearized variable

$$\hat{u}_{1k} = \theta_k \pi_k \bar{\hat{y}}_{rc(k)} + \frac{r_k}{\hat{p}_{c(k)}} \left\{ \hat{y}_{r,k} - \theta_k \pi_k \bar{\hat{y}}_{rc(k)} \right\},$$

and

$$\bar{\hat{y}}_{rc} = \frac{\sum_{k \in S_{c,hh}} d_k r_k \hat{y}_{r,k}}{\sum_{k \in S_{c,hh}} \theta_k r_k},$$

and

$$\hat{y}_{r,k} = \sum_{l \in S_{\text{ind},k}} \frac{d_{l|k} r_l}{\hat{p}_l} y_l, \tag{A.3}$$

and for the second linearized variable

$$u_{3k} = \frac{r_k}{d_k} \sum_{l \in S_{\text{ind},k}} \theta_l \left(1 - \frac{r_l}{\hat{p}_l}\right) \bar{y}_{rrd(l)}, \tag{A.4}$$

with

$$\bar{y}_{rrd} = \frac{\sum_{l \in S_{rd,\text{ind}}} d_{rl} r_l y_l}{\sum_{l \in S_{rd,\text{ind}}} \theta_l r_l}. \tag{A.5}$$

We now consider the calibrated estimator $\hat{Y}_{\text{cal,ind}}$ given in equation (2.29). The benchmark variance estimator is the same than given in equation (A.2) for $\hat{Y}_{rr,\text{ind}}$, by replacing the variable y_l with the estimated regression residuals of the variable of interest on the calibration variables, namely

$$e_l = y_l - \hat{B}_{rr,\text{ind}}^{\top} z_l \quad \text{with} \quad \hat{B}_{rr,\text{ind}} = \left(\sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l z_l^{\top}\right)^{-1} \sum_{l \in S_{rr,\text{ind}}} d_{rrl} z_l y_l. \tag{A.6}$$

B. SAS Program for one-stage sampling

In this section, we present the SAS macro developed to implement the proposed methodology for a sampling of households only (one-stage sampling). The parametrization of the SAS program for computing bootstrap weights is presented in Section B.1. For clarity, a small example is presented in Section B.2.

B.1 Program for computing bootstrap weights

The parameters related to the database are:

- `BASE`: library containing the SAS table with the list of sampled units. The default value is `BASE=WORK`.
- `ECHMEN`: SAS table containing the list of sampled units in the population. The non-respondents need also to be included in this table.

The parameters related to the bootstrap are:

- `ITBOOT`: number of bootstrap iterations. The default value is `ITBOOT=1000`.

The parameters related to the variables needed in the SAS table are:

- `IDMEN`: list of variables identifying the statistical unit. They need to be character variables.
- `STMEN`: list of variables of stratification used for the sample selection.
- `DMEN`: sampling weight.
- `RMEN`: response indicator (1 for a respondent, 0 for a non-respondent).
- `DRMEN`: sampling weight, corrected for non-response. The values are only needed for the respondents.
- `DCMEN`: calibrated weight. The values are only needed for the respondents.
- `GHRMEN`: list of variables identifying the response homogeneity groups.
- `WGHRMEN`: weighting used in the computation of the response probabilities inside RHGs.
  - With `WGHRMEN=0`, the response rates are not weighted. This is the default value.
  - With `WGHRMEN=1`, the response rates are weighted by the design weights.
- `XMENQUANT`: list of quantitative variables used in the calibration. The values are only needed for the respondents.
- `XMENQUALI`: list of qualitative variables used in the calibration. The values are only needed for the respondents.

The parameters related to the output are:

- `SORT_MEN`: SAS table containing the bootstrap sampling weights `WB_D1,...,WB_D&ITBOOT` for the whole sample.
- `SORT_RMEN`: SAS table containing the bootstrap weights `WB_N1,...,WB_N&ITBOOT` corrected for non-response, and the bootstrap weights `WB_C1,...,WB_C&ITBOOT` corrected for non-response and calibration, for the sub-sample of respondents.

B.2 A small example

We consider the example treated in Section 2.1.4. The sample is as follows:

```
data ech;
input idm$ stmen$ dmen rmen ghrmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;
```

We can obtain $B = 1{,}000$ bootstrap weights as follows. Since `WGHRMEN=1`, it is supposed that when unit non-response has been originally corrected by the method of RHGs, the response rates inside RHGs were weighted by the sampling weights.

```
%BOOTUP_1DEG(BASE=work,ECHMEN=ech,
      ITBOOT=1000,
      IDMEN=idm,STMEN=stmen,DMEN=dmen,
      RMEN=rmen,DRMEN=drmen,DCMEN=dcmen,GHRMEN=ghrmen,WGHRMEN=1,
      XMENQUANT=x0 x1,XMENQUALI=,
      SORT_MEN=ech_boot,SORT_RMEN=echr_boot);
```

C. SAS Program for two-stage sampling

In this section, we present the SAS macro developed to implement the proposed methodology for a sampling of households and a sub-sampling of individuals (two-stage sampling). The parametrization of the SAS program for computing bootstrap weights is presented in Section C.1. For clarity, a small example is presented in Section C.2.

C.1 Program for computing bootstrap weights

The SAS macro `%BOOTUP_2DEG` enables to compute bootstrap weights for a household survey with sub-sampling of individuals, and to account for correction of unit non-response via Response Homogeneity groups, and for the calibration of weights, both for households and individuals.

The parameters with equality sign are mandatory. All identifying variables must be of character type.

The parameters related to the database are:

- `BASE`: library containing the SAS tables `ECHMEN` and `ECHIND`. The default value is `BASE=WORK`.
- `BASESOR`: library containing the output. The default value is `BASESOR=WORK`.
- `ECHMEN=`: SAS table containing the list of sampled households in the population. The household non-respondents need also to be included in this table.
- `ECHIND=`: SAS table containing the list of sampled individuals inside all the responding households. The individual non-respondents need also to be included in this table.

The parameters related to the bootstrap are:

- `ITBOOT`: number of bootstrap iterations. The default value is `ITBOOT=1000`.

The parameters related to the variables needed in the household SAS table `ECHMEN` are:

- `IDMEN=`: list of variables identifying the household. This variable is required in both `ECHMEN` and `ECHIND`.
- `STMEN`: list of variables of stratification used for the sample selection.

- `DMEN`: sampling weight of the household.
- `RMEN`: response indicator of the household (1 for a respondent, 0 for a non-respondent).
- `DRMEN`: sampling weight of the household, corrected for non-response. The values are only needed for the respondents.
- `DCMEN`: calibrated weight. The values are only needed for the respondents.
- `GHRMEN`: list of variables identifying the response homogeneity groups for households.
- `WGHRMEN`: weighting used in the computation of the response probabilities inside RHGs:
    - With `WGHRMEN=0`, the response rates are not weighted. This is the default value.
    - With `WGHRMEN=1`, the response rates are weighted by the design weights `DMEN`.
- `XMENQUANT`: list of quantitative variables used in the calibration. The values are only needed for the respondents.
- `XMENQUALI`: list of qualitative variables used in the calibration. The values are only needed for the respondents.

The parameters related to the variables needed in the individual SAS table `ECHIND` are:

- `ID_IND=`: list of variables identifying the individual (character variable).
- `R_IND`: response indicator of the individual (1 for a respondent, 0 for a non-respondent).
- `DR_IND`: weight of the individual, corrected for both household and individual unit non-response. The values are only needed for the respondents.
- `DC_IND`: calibrated weight. The values are only needed for the respondents.
- `PIKSACI=`: conditional inclusion probability of the individual inside its household.
- `GHR_IND`: list of variables identifying the response homogeneity groups.
- `WGHR_IND`: weighting used in the computation of the response probabilities inside RHGs:
    - With `WGHR_IND=0`, the response rates are not weighted. This is the default value.
    - With `WGHR_IND=1`, the response rates are weighted by the design weights of individuals.
    - With `WGHR_IND=2`, the response rates are weighted by the weights of individuals, adjusted for household unit non-response.
- `XINDQUANT`: list of quantitative variables used in the calibration. The values are only needed for the respondents.
- `XINDQUALI`: list of qualitative variables used in the calibration. The values are only needed for the respondents.

The parameters related to the output are:

- `SORT_MEN`: SAS table containing all the sampled households, and the bootstrap sampling weights `WB_D1,...,WB_D&ITBOOT` for the whole sample.
- `SORT_RMEN`: SAS table containing all the responding households, and the bootstrap weights
    - `WB_N1,...,WB_N&ITBOOT` corrected for non-response,
    - `WB_C1,...,WB_C&ITBOOT` corrected for non-response and calibration.
- `SORT_RIND`: SAS table containing all the responding individuals inside the responding households, and the bootstrap weights

- `WB_N1,...,WB_N&ITBOOT` corrected for household non-response,
- `WB_NN1,...,WB_NN&ITBOOT` corrected for both household non-response and individual non-response,
- `WB_C1,...,WB_C&ITBOOT` corrected for non-response and calibration.

C.2 A small example

We consider the example treated in Section 2.2.4. The sample of households and the sample of individuals are as follows:

```
data echmen;
input idm$ stmen$ dmen rmen ghrmen$ drmen dcmen x0 x1;
cards;
A  1  4   1  aa  4.44   4.01   1  1
B  1  4   0  aa  .      .      .  .
C  1  4   0  bb  .      .      .  .
D  1  4   1  bb  5.54   4.87   1  0
E  1  16  1  bb  22.15  19.98  1  1
F  1  16  1  aa  17.78  15.63  1  0
G  1  16  0  bb  .      .      .  .
H  1  16  1  bb  22.15  19.98  1  1
I  1  16  1  bb  22.15  19.49  1  0
J  1  16  1  aa  17.78  16.03  1  1
;run;
```

```
data echind;
input idm$ idi$ piksaci dr1_ind rind ghrind$ phat_ind dr2_ind xi1 xi2 dc_ind;
cards;
A   i01   0.34   13.06   1   g1   0.75   17.41   1   3   19.61
D   i04   1.00   5.54    0   g2   0.33   .       .   .   .
E   i06   0.34   65.15   1   g1   0.75   86.86   1   2   53.93
F   i08   0.33   53.88   1   g1   0.75   71.84   1   3   78.43
H   i11   0.50   44.30   0   g1   0.75   .       .   .   .
I   i13   1.00   22.15   1   g2   0.33   67.12   1   1   48.04
J   i14   1.00   17.78   0   g2   0.33   .       .   .   .
;run;
```

We can obtain $B = 1{,}000$ bootstrap weights as follows. Since `WGHRMEN=1`, it is supposed that when unit non-response of households has been originally corrected by the method of RHGs, the response rates inside RHGs were weighted by the sampling weights. Since `WGHR_IND=0`, it is supposed that when unit non-response of individuals has been originally corrected by the method of RHGs, the response rates inside RHGs were unweighted.

```
%bootup_2deg(BASE=work,BASESOR=work,ECHMEN=echmen,ECHIND=echind,
     ITBOOT=1000,
     IDMEN=idm,STMEN=stmen,DMEN=dmen,RMEN=rmen,DRMEN=drmen,GHRMEN=ghrm
     en,WGHRMEN=0,
     DCMEN=dcmen,XMENQUANT=x0 x1,XMENQUALI=,
     ID_IND=idi,R_IND=rind,DR_IND=dr2_ind,PIKSACI=piksaci,GHR_IND=ghri
     nd,WGHR_IND=0,
     DC_IND=dc_ind,XINDQUANT=xi1 xi2,XINDQUALI=,
     SORT_MEN=sort_men,SORT_RMEN=sort_rmen,
     SORT_RIND=sort_rind);
```

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3), 445-458.

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80(1), 127-148.

Brick, J.M. (2013). Unit non-response and weighted adjustments: A critical review. *Journal of Official Statistics*, 29(3), 329-353.

Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. PhD thesis, University of Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chauvet, G., and Vallée, A.-A. (2018). Inference for two-stage sampling designs with application to a panel for urban policy. *arXiv preprint arXiv:1808.09758*.

Davison, A.C., and Hinkley, D.V. (1997). Bootstrap methods and their application, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*.

Davison, A.C., and Sardy, S. (2007). Resamping variance estimation in surveys with missing data. *Journal of Official Statistics*, 23(3), 371-386.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4882-eng.pdf.

Girard, C. (2009). The Rao-Wu rescaling bootstrap: from theory to practice. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, pages 2-4. Citeseer.

Haziza, D., and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1), 25-43.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Juillard, H., and Chauvet, G. (2018). Variance estimation under monotone non-response for a panel survey. *Survey Methodology*, 44, 2, 269-289. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54952-eng.pdf.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 35(4), 501-514.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101(473), 312-320.

Kott, P.S. (2012). Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups. *Survey Methodology*, 38, 1, 95-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11689-eng.pdf.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

McCarthy, P., and Snowden, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (95), 1-23.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 2, 209-217. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf.

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

Shao, J. (1994). *L*-Statistics in complex survey problems. *The Annals of Statistics*, 22(2), 946-967.

Shao, J., and Rao, J. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā: The Indian Journal of Statistics, Series B*, 393-414.

Shao, J., and Tu, D.S. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Tillé, Y. (2011). *Sampling Algorithms*. Springer.

Yeo, D., Mantel, H. and Liu, T.-P. (1999). Bootstrap variance estimation for the national population health survey. In *Proceedings of the Survey Research Methods Section*, American Statistical Association. Citeseer.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 349-359
Statistics Canada, Catalogue No. 12-001-X

An alternative jackknife variance estimator when calibrating weights to adjust for unit nonresponse in a complex survey

Phillip S. Kott and Dan Liao[1]

Abstract

Calibration weighting is a statistically efficient way for handling unit nonresponse. Assuming the response (or output) model justifying the calibration-weight adjustment is correct, it is often possible to measure the variance of estimates in an asymptotically unbiased manner. One approach to variance estimation is to create jackknife replicate weights. Sometimes, however, the conventional method for computing jackknife replicate weights for calibrated analysis weights fails. In that case, an alternative method for computing jackknife replicate weights is usually available. That method is described here and then applied to a simple example.

Key Words: Analysis weight; Linearization-based variance estimator; Delete-1 jackknife variance estimator; Replicate weight; Asymptotically unbiased; Bounded logistic response model.

1. Introduction

Calibration weighting is a method for adjusting the weights in probability-sampling theory by forcing the weighted sum of each variable in a set of survey variables to equal a specified target. When that happens, the analysis weights are said to satisfy the calibration equation. There are several reasons to calibrate analysis weights. The reason we focus on here is to remove potential selection bias resulting from unit nonresponse.

It is common in the survey-sampling literature to argue that a survey respondent's calibration-weight adjustment implicitly estimates the inverse of its probability of response (see, for example, Section 5.1 of Fuller, 2009). Kott and Liao (2012) show that using calibration weighting to adjust for unit nonresponse can provide *double protection* against nonresponse bias when estimating a population total. This means that if either a linear outcome model or an implied selection model holds, then the resulting estimator is asymptotically unbiased in some sense. They go on to describe a linearization-based variance estimator for an estimated total based on a stratified multistage (or single-stage) sample with calibration-adjusted analysis weights.

The brief treatment in Sections 2 and 3 of calibration weighting for nonresponse and of linearization-based variance estimation for a calibrated estimator of a population total are developed in more depth in Kott and Liao. Proofs of the various assertions made in these sections can be found there. Here they set up the theory behind variance estimation with a jackknife.

Given a stratified multistage probability sample, a traditional delete-1 jackknife variance estimator creates sets of replicate weights, one set corresponding to each selected primary sampling unit (PSU). One selected PSU is dropped at a time, and the replicate weights of its subsampled elements are set to zero. To

---

[1] Phillip S. Kott, Senior Research Statistician, RTI International, Rockville, Maryland 20852, U.S.A. E-mail: pkott@rti.org; Dan Liao, Research Statistician, RTI International, Rockville, Maryland 20852, U.S.A.

compensate, the replicate probability weights of the remaining elements in the same stratum as the dropped PSU are increased by the factor $n_h / (n_h - 1),$ where n_h is the original number of PSUs selected from the stratum. The replicate probability weights are calibrated in a manner analogous to the original analysis weights. Section 4 describes this jackknife and shows its near equality to the nearly-unbiased linearization-based variance estimator for a population total.

The advantage of a delete-1 jackknife over linearization for variance estimator is that once replicate weights are computed, estimating the variance of smooth function of estimated totals (such as a regression coefficient) is straightforward. Krewski and Rao (1981) provides a rigorous treatment of the delete-1 jackknife and its properties.

Sometimes no solution to a calibration equation exists when starting with a set of replicate probability weights. The main contribution of this paper is contained in the remainder of Section 4, where an alternative method of constructing jackknife replicate weights that can usually overcome this problem is described and justified. This method was introduced in Kott (2006) for another purpose.

Section 6 uses a weight-adjustment function described in Section 5 to illustrate how to implement this method. It then favorably compares the results of the method to those of two popular competitors. Section 7 discusses a variant of the alternative jackknife methodology.

2. Calibration weighting

Suppose we have a randomly drawn sample S from a finite population U. In the absence of nonresponse (as well as coverage error and measurement error), calibration weighting creates a set of analysis weights, $\{w_k \mid k \in S\},$ not dependent on the survey values of interest that

1. are close to the original inverse-probability weights, $d_k = 1/\pi_k$ where π_k is the selection probability of the k^{th} selected element; and
2. satisfy a set of linear calibration equations, one for each component of $\mathbf{z}_k,$ a vector of auxiliary variables with known population totals:

$$\sum_{k \in S} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k$$

"Close" means that as the sample grows arbitrarily large, the difference between w_k and d_k vanishes in probability. For a more formal treatment of the assumed asymptotic structure, see Isaki and Fuller (1982).

Most surveys experience unit nonresponse beyond a statistician's control. One is forced to assume, either explicitly or implicitly, some type of model to adjust for the nonresponse. An outcome model (also called a "prediction model") on a survey variable of interest usually assumes the response/nonresponse mechanism, like the sampling design, is ignorable. A response model assumes the response mechanism behaves like a phase of Poisson (i.e., independent) subsampling. Double protection means that if *either* the prediction or response model is specified correctly, the estimator will be nearly (i.e., asymptotically) unbiased in some sense. Here we will assume a correctly specified response model.

Let R be the subset of S containing respondents to the survey (for simplicity, we ignore the possibility of item nonresponse). The respondent sample can be calibrated to either the full population U:

$$\sum_{k \in R} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k \tag{2.1}$$

or to the original sample S:

$$\sum_{k \in R} w_k \mathbf{z}_k = \sum_{k \in S} d_k \mathbf{z}_k. \tag{2.2}$$

We assume a response model in which the probability of response for each $k \in U, p_k,$ is an independent function having the form $p_k = p(\boldsymbol{\gamma}^T \mathbf{x}_k),$ where $p(.)$ is a smooth monotonic function, and both the known vector $\mathbf{x}_k$ and unknown parameter vector $\boldsymbol{\gamma}$ have the same number of components as $\mathbf{z}_k$. In much of the literature $\mathbf{x}_k$ is equal to $\mathbf{z}_k,$ but most of the theory still follows when it does not.

If there is a vector $\mathbf{g}$ such that inserting $w_k = d_k / p(\mathbf{g}^T \mathbf{x}_k)$ solves either the calibration equation in (2.1) or (2.2), then $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$. Kott and Liao (2017) describe what to do when there are fewer components in $\mathbf{x}_k$ than in $\mathbf{z}_k$.

The function $f(\mathbf{g}^T \mathbf{x}_k) = 1/p(\mathbf{g}^T \mathbf{x}_k)$ is called the *weight-adjustment function*. The mean-value theorem tells us that under mild conditions $f(\mathbf{g}^T \mathbf{x}_k) - f(\boldsymbol{\gamma}^T \mathbf{x}_k) \approx f'(\mathbf{g}^T \mathbf{x}_k)(\mathbf{g} - \boldsymbol{\gamma})^T \mathbf{x}_k$. Consequently, as the respondent sample grows arbitrarily large $f(\mathbf{g}^T \mathbf{x}_k)$ converges to $f(\boldsymbol{\gamma}^T \mathbf{x}_k) = 1/p_k$ and $\mathbf{g}$ converges to $\boldsymbol{\gamma}$.

3. Linearization-based variance estimation

When calibrating the respondent sample to the full sample with (2.2), the calibration estimator for a population total, $t = \sum_R w_k y_k$ can be expressed as

$$\begin{aligned} t &= \sum_{k \in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k \in R} d_k f(\mathbf{g}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) \\ &\approx \sum_{k \in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k \in R} d_k f(\boldsymbol{\gamma}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) + \sum_{k \in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\left[(\mathbf{g} - \boldsymbol{\gamma})^T \mathbf{x}_k\right](y_k - \mathbf{z}_k^T \mathbf{b}) \\ &= \sum_{k \in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k \in R} d_k f(\boldsymbol{\gamma}^T \mathbf{x}_k)(y_k - \mathbf{z}_k^T \mathbf{b}) + (\mathbf{g} - \boldsymbol{\gamma})^T \sum_{k \in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\mathbf{x}_k (y_k - \mathbf{z}_k^T \mathbf{b}) \\ &= \sum_{k \in S} d_k \mathbf{z}_k^T \mathbf{b} + \sum_{k \in R} d_k p_k^{-1}(y_k - \mathbf{z}_k^T \mathbf{b}), \end{aligned} \tag{3.1}$$

where

$$\mathbf{b} = \left[\sum_{k \in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\mathbf{x}_k \mathbf{z}_k^T\right]^{-1} \sum_{k \in R} d_k f'(\mathbf{g}^T \mathbf{x}_k)\mathbf{x}_k y_k.$$

The key step here is that $\mathbf{b}$ has been defined so that $\sum_R d_k f_k'(\mathbf{g}^T\mathbf{x}_k)\,\mathbf{x}_k(y_k - \mathbf{z}_k^T\mathbf{b}) = 0.$ Observe that $f'(.)$ in $\mathbf{b}$ is the *derivative* of the weighting-adjustment function.

Let $\mathbf{b}^*$ be the probability limit of $\mathbf{b}$ as the respondent sample (of PSUs) grows arbitrarily large. The variance of t under the original design and the selection model is nearly equivalent to the variance of $\sum_S d_k q_k^*$, where

$$q_k^* = \mathbf{z}_k^T\mathbf{b}^* + p_k^{-1}(y_k - \mathbf{z}_k^T\mathbf{b}^*)\,I_k,$$

and $I_k = 1$ when k is a unit respondent and 0 otherwise.

For many designs, q_k^* can be approximated by replacing $\mathbf{b}^*$ with $\mathbf{b}$ and p_k^{-1} with $f(\mathbf{g}^T\mathbf{x}_k)$, and the variance of $\sum_S d_k q_k$ estimated under the original design as if the $q_k = \mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$ were constants. When calibrating the respondent sample to the population with equation (2.1), the $\sum_S d_k \mathbf{z}_k^T\mathbf{b}$ in equation (3.1) is replaced by $\sum_U \mathbf{z}_k^T\mathbf{b}$, which does not contribute to the variance, so $q_k = f(\mathbf{g}^T\mathbf{x}_k)(y_k - \mathbf{z}_k^T\mathbf{b})\,I_k$. Either way, replacing q_k^* with q_k tends to underestimate variances with *finite* samples (the replacement is asymptotically ignorable) because $e_k = (y_k - \mathbf{z}_k^T\mathbf{b})^2$ tends to be smaller than $e_k^* = (y_k - \mathbf{z}_k^T\mathbf{b}^*)^2$.

Given a stratified multistage probability sample with n_h sampled PSUs in each of H strata, let S_{hj} denote the subsample of elements within each PSU j in stratum h. A nearly unbiased linearization-based estimator for the variance of t is

$$v(t) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} d_k q_k\right)^2 - \frac{1}{n_h}\left(\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} d_\kappa q_\kappa\right)^2\right], \tag{3.2}$$

where $q_k = \alpha\mathbf{z}_k^T\mathbf{b} + f(\mathbf{g}^T\mathbf{x}_k)\,e_k I_k$, and $\alpha = 1$ when the respondent sample is calibrated to the original sample and 0 when the respondent sample is calibrated to the population. As is common in practice and continued here, equation (3.2) assumes that the little is lost by treating the PSU selection within strata as if it had been drawn with replacement, obviating the need for finite population correction.

4. Jackknife variance estimation

Let S_{h+} be all the sampled elements in stratum h. The conventional hj^{th} (delete-1) jackknife replicate for an estimated total $t = \sum_R w_k y_k$ is

$$t^{(hj)} = \sum_{k\in R} w_k^{(hj)} y_k = \sum_{k\in S} d_k^{(hj)} f(g^{(hj)T}\mathbf{x}_k)\,I_k y_k, \tag{4.1}$$

where $d_k^{(hj)} = 0$ when $k \in S_{hj}$

$d_k^{(hj)} = [n_h/(n_h - 1)]\,d_k$ when $k \in S_{h+}$ but $k \notin S_{hj}$

$$d_k^{(hj)} = d_k \qquad \text{otherwise,}$$

and the $\mathbf{g}^{(hj)}$ solve the replicate calibration equation:

$$\sum_{k \in R} d_k^{(hj)} f\,(\mathbf{g}^{(hj)T}\mathbf{x}_k)\,\mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k \text{ when the respondent sample is calibrated to } U, \text{ or}$$

$$\sum_{k \in R} d_k^{(hj)} f\,(\mathbf{g}^{(hj)T}\mathbf{x}_k)\,\mathbf{z}_k = \sum_{k \in S} d_k^{(hj)}\mathbf{z}_k \text{ when the respondent sample is calibrated to } S$$

for each hj. Observe that $t^{(hj)}$ is an estimate of the population total $\sum_U y_k$ with the PSU hj removed.

The delete-1 jackknife variance estimator for t is

$$\operatorname{var}_J(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (t^{(hj)} - t)^2. \tag{4.2}$$

Let $e_k^* = (y_k - \mathbf{z}_k^T \mathbf{b}^*)$. Consequently,

$$\begin{aligned} t^{(hj)} - t &= \sum_{k \in S} \left\{ (d_k^{(hj)} - d_k)\,\alpha \mathbf{z}_k^T \mathbf{b}^* + \left[d_k^{(hj)} f\,(\mathbf{g}^{(hj)T}\mathbf{x}_k) - d_k\, f\,(\mathbf{g}^T \mathbf{x}_k) \right] I_k e_k^* \right\} \\ &\approx \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b}^* + f\,(\mathbf{g}^T \mathbf{x}_k)\, I_k e_k^* \right]. \end{aligned}$$

From which we can conclude that $\operatorname{var}_J(t)$ is approximately equal to the delete-1 jackknife for $\sum_S d_k q_k^*$, where $q_k^* = \alpha \mathbf{z}_k^T \mathbf{b}^* + f\,(\mathbf{g}^T \mathbf{x}_k)\, I_k e_k^*$, based on a stratified multistage sample with H strata and n_h PSUs in stratum h. A little algebra will show that the delete-1 jackknife for $\sum_S d_k q_k^*$ is equal to $\operatorname{var}(t)$ in equation (3.2) with q_k^* replacing q_k because

$$\sum_{k \in S} (d_k^{(hj)} - d_k)\, q_k^* = \sum_{h=1}^{H} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k)\, q_k^*,$$

where

$$\begin{aligned} \sum_{k \in S_{h+}} (d_k^{(hj)} - d_k) q_k^* &= \sum_{k \in S_{h+}} \frac{1}{n_h - 1} d_k q^* - \frac{n_h}{n_h - 1} \sum_{k \in S_{hj}} d_k q^* \\ &= -\frac{n_h}{n_h - 1} \left(\sum_{k \in S_{hj}} d_k q^* - \frac{1}{n_h} \sum_{k \in S_{h+}} d_k q^* \right). \end{aligned}$$

Note that the contribution to the jackknife variance estimator for the hj^{th} replicate comes mostly from the hj^{th} PSU.

Observe that the small downward bias in finite samples caused by q_k replacing q_k^* in $\operatorname{var}(t)$ does not apply to $\operatorname{var}_J(t)$ in equation (4.2). The latter may have a slight tendency to be upwardly biased in finite samples because $\mathbf{g}^{(hj)}$ and $\mathbf{g}$, while both consistent estimators for $\boldsymbol{\gamma}$, need not be exactly equal.

There is sometimes a problem with computing the jackknife variance estimator $\operatorname{var}_J(t)$ in practice. That problem occurs when $f\,(.)$ is such that while there is a $\mathbf{g}$ satisfying the calibration equation in (2.1) or (2.2), no $\mathbf{g}^{(hj)}$ satisfies its analogue for at least one hj jackknife replicate. When that happens, one can follow a suggestion in Kott (2006) and compute the $w_k^{(hj)}$ in equation (4.1) with this alternative:

$$\tilde{w}_k^{(hj)} = d_k^{(hj)} f(g^T \mathbf{x}_k) + \left[\mathbf{c}^{(hj)} - \sum_{\kappa \in R} d_\kappa^{(hj)} f(g^T \mathbf{x}_\kappa) \mathbf{z}_\kappa^T \right] \left[\sum_{\kappa \in R} d_\kappa^{(hj)} f'(g^T \mathbf{x}_\kappa) \mathbf{x}_\kappa \mathbf{z}_\kappa^T \right]^{-1} d_k^{(hj)} f'(g^T \mathbf{x}_k) \mathbf{x}_k, \quad (4.3)$$

where $\mathbf{c}^{(hj)}$ is the calibration target for the hj^{th} replicate:

$$\mathbf{c}^{(hj)} = \sum_{\kappa \in U} \mathbf{z}_\kappa^T \text{ when the respondent sample is calibrated to } U,$$

and

$$\mathbf{c}^{(hj)} = \sum_{\kappa \in S} d_\kappa^{(hj)} \mathbf{z}_\kappa^T \text{ when the respondent sample is calibrated to } S.$$

By design $\sum_R \tilde{w}_k^{(hj)} \mathbf{z}_k = \mathbf{c}^{(hj)}$.

Letting $e_k = y_k - \mathbf{z}_k^T \mathbf{b}$, one can see that with $\tilde{t}^{(hj)} = \sum_R \tilde{w}_k^{(hj)} y_k$,

$$\begin{aligned} \tilde{t}^{(hj)} - t &= \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b} + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k \right] \\ &\approx \sum_{k \in S} (d_k^{(hj)} - d_k) \left[\alpha \mathbf{z}_k^T \mathbf{b}^* + f(\mathbf{g}^T \mathbf{x}_k) I_k e_k^* \right] \end{aligned}$$

so the alternative jackknife variance estimator $\text{var}_{AJ}(t)$ computed with $\tilde{t}^{(hj)}$ in place of $t^{(hj)}$ is nearly unbiased. Observe that the only possible restriction on the computation of $\tilde{w}_k^{(hj)}$ is that $\sum_R d_k^{(hj)} f'(g^T \mathbf{x}_k)\, \mathbf{x}_k \mathbf{z}_k^T$ be non-singular.

Observe that equation (4.3) can be rewritten as

$$\tilde{w}_k^{(hj)} = \tilde{d}_k^{(hj)} \tilde{f}(\tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k) \qquad (4.4)$$

where $\tilde{d}_k^{(hj)} = d_k^{(hj)} f(g^T \mathbf{x}_k)$,

$$\tilde{f}(\tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k) = 1 + \tilde{\mathbf{g}}^{(hj)T} \tilde{\mathbf{x}}_k,$$

$$\tilde{\mathbf{g}}^{(hj)T} = \left[\mathbf{c} - \sum_{\kappa \in R} \tilde{d}_\kappa^{(hj)} \mathbf{z}_\kappa^T \right] \left[\sum_{\kappa \in R} \tilde{d}_\kappa^{(hj)} \tilde{\mathbf{x}}_\kappa \mathbf{z}_\kappa^T \right]^{-1},$$

and

$$\tilde{\mathbf{x}}_k = \frac{f'(\mathbf{g}^T \mathbf{x}_k)}{f(\mathbf{g}^T \mathbf{x}_k)} \mathbf{x}_k.$$

This equation treats the hj^{th} replicate as the full sample. The weight-adjustment function $\tilde{f}(.)$ is linear, and the $\tilde{f}_k$ are not restricted to positive values even when the f_k are. In addition, observe that even when $\mathbf{x}_k = \mathbf{z}_k$, $\tilde{\mathbf{x}}_k$ will not equal $\mathbf{z}_k$ unless $f(.) = \exp(.)$.

5. The (bounded) logistic response model

Up until this point, we have not specified a response function, $p(.) = 1/f(.)$. Consider now a (bounded) logistic (or logit) response model having the form:

$$p(\boldsymbol{\gamma}^T \mathbf{x}_k) = \frac{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)/U}{L + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)}. \tag{5.1}$$

where $1 \leq L < U$. When $L = 1$ and U is infinite, this is a standard logistic response model, where the response probability can range from 0 to 1, not including the endpoints. For finite values of L and U, the bounded probability of response falls between $1/U$ and $1/L$. Consequently, the value of adjustment function ranges from L to U. In practice, L is usually set to 1, while U is frequently set as low as the sample allows for the calibration equation to hold.

The calibration procedures in the SUDAAN® language (Research Triangle Institute, 2012), WTADJUST for when $\mathbf{x}_k = \mathbf{z}_k$ and WTADJX otherwise, fit an equivalent weight-adjustment function:

$$f(\mathbf{g}^T \mathbf{x}_k) = \frac{L + B\exp(A\mathbf{g}^T \mathbf{x}_k)}{1 + B\exp(A\mathbf{g}^T \mathbf{x}_k)/U}, \tag{5.2}$$

where $A = \frac{U-L}{(C-L)(U-C)}$, $B = U\frac{C-L}{U-C}$, and $L < C < U$.

The choice of C helps determine what $\mathbf{g}$ satisfies the calibration equation but will not affect the value of the weight adjustment itself, $f_k = f(\mathbf{g}^T \mathbf{x}_k)$. Consequently, C can be any value between L and U. When $L = 1$, $C = 2$, and U is infinite, $\mathrm{A} = \mathrm{B} = 1$.

A little calculus reveals with the weight-adjustment function in equation (5.2):

$$f_k' = f'(g^T \mathbf{x}_k) = \frac{(U - f_k)(f_k - L)}{(U - C)(C - L)},$$

which is needed to compute equation (4.3) or (4.4). The general exponential model in the SUDAAN calibration procedures allow the L, C, and U to vary from element to element, a flexibility hard to interpret in response modeling and not considered here.

What will be useful here, although not for modeling, is the possibility that L in equation (5.2) is 0 and U is infinite. When iteratively solving a calibration equation for $\mathbf{g}$ with $f(\mathbf{g}^T \mathbf{x}_k) = \exp(\mathbf{g}^T \mathbf{x})$ using Newton's method, the SUDAAN calibration procedures first solve for $\mathbf{g}_1$ in the calibration equation with $f(\mathbf{g}_1^T \mathbf{x}_k) = 1 + \mathbf{g}_1^T \mathbf{x}_k$, which is a useful result when computing alternative jackknife weights. (The programs set the first iteration of the weight adjustment at $f_{k1} = \exp(\mathbf{g}_1^T \mathbf{x}_k)$ from which $1 + \mathbf{g}_1^T \mathbf{x}_k$ is easily derived.)

6. A simulation example

The MU281 population of municipalities in Särndal, Swensson and Wretman (1992; data from the slightly revised version is contained in http://lib.stat.cmu.edu/datasets/mu284; one of the municipalities was accidentally dropped in this analysis) has been augmented with an indicator (RESP) for whether an element (municipality) would respond if sampled. Probabilities of element response were generated using a logistic function of one of the data set's covariates (the log of the element's 1975 population in thousands). The average probability of response was roughly 70%.

A stratified simple random sample of 10 elements per each of 8 strata was simulated 1,716 times. In each simulated sample, the elements with RESP = 1 were treated as respondents, and the respondent sample was calibrated to the full sample using the weight-adjustment function in equation (5.2) with a lower bound of 1 and an upper bound of 5. In the calibration model, the two components of $\mathbf{x}_k$ were 1 and the log of the element's 1975 population in thousands; $\mathbf{z}_k$ was set equal to $\mathbf{x}_k$. 1,225 out of the 1,716 simulations had their respondent samples successfully calibrated on both components (i.e., satisfied the calibration equation in 2.2) and produced linearization-based standard-errors.

Estimated means (ratios of two estimated totals) and standard errors (square roots of estimated variances) were computed for four variables:

P85	1985 population (in thousands).
RMT85	Revenues from 1985 municipal taxation (in millions of kronor).
ME84	Number of municipal employees in 1984.
REV84	Real estate values according to 1984 assessment (in millions of kronor).

Although the SUDAAN procedure WTADJUST can compute standard errors when using a delete-1 jackknife, it will fail when one or more replicates fail to calibrate. Therefore, two versions of the conventional delete-1 jackknife standard errors were computed using a macro the authors created. In one, the set of the imperfect "calibrated" weights from the last iteration for the failed replicates were used. In the other, the replicates that failed to calibrate were dropped and this following modified jackknife variance estimator was computed:

$$\operatorname{var}_J^*(t) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h^*} \sum_{j=1}^{n_h^*} (t^{(hj)} - t)^2, \tag{6.1}$$

where n_h^* is the number of replicates in stratum h that successfully calibrated. This revised jackknife variance estimator is suggested by Rust (1985) when replicates are dropped at random, which is not what happens here. The SAS-callable (SAS Institute Inc., 2015) SUDAAN code used in the analysis for a single simulation is available from the authors upon request.

Among the 1,225 analyzable samples, 867 simulations had all the replicates using conventional delete-1 jackknife calibrate, while the remaining 358 simulations had at least one replicate that failed to calibrate after 50 iterations (the default is 10). Table 6.1 averages the results for both situations. When no conventional replicate failed, the alternative and conventional jackknife standard errors are close (on average) and slightly higher than those produced by linearization as theory predicts (note that the two versions of the conventional delete-1 jackknife are identical).

Table 6.1
Standard errors based on jackknife methods when calibrating for nonresponse with a bounded logistic model

	Variable	Estimated Mean	Linearization-based Standard Error	Alternative Jackknife Standard Error	Conventional Jackknife Standard Error Including Failed Replicates	Conventional Jackknife Standard Errors Dropping Failed Replicates
867 Simulations Where No Conventional Replicate Failed to Calibrate	P85	22.41	2.06	2.09	2.11	2.11
	RMT85	167.70	17.31	17.72	17.86	17.86
	ME84	1,215.87	124.67	127.27	128.15	128.15
	REV84	2,425.52	212.83	217.00	219.67	219.67
358 Simulations Where at Least One Conventional Replicate Failed to Calibrate	P85	22.78	2.24	2.31	2.95	2.04
	RMT85	170.48	18.93	19.47	24.39	16.60
	ME84	1,236.75	135.94	139.65	175.99	121.31
	REV84	2,451.95	239.27	239.18	296.77	208.45

When at least one replicate failed to calibrate for the conventional delete-1 jackknife, the alternative jackknife's standard errors are again close to linearization-based ones, even though it failed to calibrate in 114 out of these 358 simulations due to a (near) singularity in at least one of the replicates. However, including the failed replicates clearly overestimates standard error and dropping them clearly underestimates relative to linearization. It appears that the alternative jackknife variance estimator produces the more useful set of replicate weights in this situation.

Table 6.1 compares standard-errors from competing jackknifes to linearization-based standard errors rather than empirical standard errors because finite-population correction has been ignored. Moreover, the bounded logistic response model fit in the simulations was not the unbounded response model used to generate responses.

7. Discussion

There is a small chance (about 1.5% in our simulations) for equation (4.4) to return negative replicate weights. The canned procedures of many statistical packages (like SAS) cannot handle negative weights. Consequently, estimated totals computed from replicate weights may need to be calculated without the help of a canned procedure.

One does not need access to SUDAAN to compute alternative jackknife weights for calibration estimators. The *gencalib* routines in the 'Sampling' package in R (Tillé and Matei, 2016) can perform calibration not only under a bounded logistic response model but under linear calibration as well. Although there are SAS macros equivalent to WTADJUST, to our knowledge, there is currently no publicly-available SAS calibration-weighting macro that can be used when $\tilde{\mathbf{x}}_k$ in the weight-adjustment (equation 4.4) does not equal $\mathbf{z}_k$. Let us hope this is reversed soon.

Acknowledgements

The authors would like to thank the editors whose helpful suggestions improved the quality of this note.

References

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

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

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

Kott, P., and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration weighting routine. *Survey Research Methods,* 6, 105-111.

Kott, P., and Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, 5, 159-174.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Research Triangle Institute (2012). *SUDAAN Language Manual, Release 11.0.* RTI International, Research Triangle Park, NC.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

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

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. SAS Institute Inc., Cary, NC.

Tillé, Y., and Matei, A. (2016). Package 'sampling' is available online at http://cran.r-project.org/web/packages/sampling/sampling.pdf.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 361-370
Statistics Canada, Catalogue No. 12-001-X

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

Yong You[1]

Abstract

In this paper, we consider the Fay-Herriot model for small area estimation. In particular, we are interested in the impact of sampling variance smoothing and modeling on the model-based estimates. We present methods of smoothing and modeling for the sampling variances and apply the proposed models to a real data analysis. Our results indicate that sampling variance smoothing can improve the efficiency and accuracy of the model-based estimator. For sampling variance modeling, the HB models of You (2016) and Sugasawa, Tamae and Kubokawa (2017) perform equally well to improve the direct survey estimates.

Key Words: EBLUP; Hierarchical Bayes; Gibbs sampling; Log-linear model; Relative error; Sampling variance; Small area.

1. Introduction

Small area estimation is popular and important in survey data analysis. Model-based estimates have been widely used in practice to provide reliable estimates for small areas. In practice, area level models are usually used whenever direct survey estimates and area level auxiliary variables are available. Various area level models have been proposed to improve the precision of the direct survey estimates, see Rao and Molina (2015). Among the area level models, the Fay-Herriot model (Fay and Herriot, 1979) is a basic area level model widely used in small area estimation. The Fay-Herriot model has two components, namely, a sampling model for the direct survey estimates and a linking model for the small area parameter of interest. The sampling model assumes that a direct survey estimator y_i is design unbiased for the small area parameter θ_i such that

$$y_i = \theta_i + e_i, \quad i = 1, \ldots, m, \tag{1.1}$$

where e_i is the sampling error associated with the direct estimator y_i and m is the number of small areas. It is customary to assume that e_i's are independently normal random variables with mean $\mathrm{E}(e_i) = 0$ and sampling variance $\mathrm{Var}(e_i) = \sigma_i^2$. The linking model assumes that the small area parameter θ_i is related to auxiliary variables $x_i = (x_{i1}, \ldots, x_{ip})'$ through a linear regression model given as

$$\theta_i = x_i' \beta + v_i, \quad i = 1, \ldots, m, \tag{1.2}$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, and the v_i's are area-specific random effects assumed to be independent and identically distributed with $\mathrm{E}(v_i) = 0$ and $\mathrm{Var}(v_i) = \sigma_v^2$. The assumption of normality is generally included. Random effects v_i and sampling errors e_i are mutually

---

[1] Yong You, International Cooperation and Methodology Innovation Centre (ICMIC), Statistics Canada, Ottawa, Canada. E-mail: yong.you@statcan.gc.ca.

independent. The model variance σ_v^2 is unknown and needs to be estimated. Combining models (1.1) and (1.2) leads to a linear mixed model given as

$$y_i = x_i' \beta + v_i + e_i, \quad i = 1, \ldots, m. \tag{1.3}$$

Model (1.3) involves both design-based random errors e_i and model-based random effects v_i. For the Fay-Herriot model, the sampling variance σ_i^2 is usually assumed to be known. This is a very strong assumption. In practice, unbiased direct estimates of the sampling variances are generally available. To make use of the direct sampling variance estimates, two approaches are available in practice, namely, smoothing and modeling. For the smoothing approach, smoothed estimates of the sampling variances are used in the Fay-Herriot model and then treated as known. The smoothing approach requires external variables and external models such as use of the generalized variance function (GVF) and design effects. You and Hidiroglou (2012) particularly studied the GVF and design effects methods for sampling variance smoothing for proportions. In this paper, we will use a GVF model proposed in You and Hidiroglou (2012) for the sampling variance smoothing.

As an alternative to smoothing, sampling variance modeling is also commonly used in practice. Let s_i^2 denote the direct estimator for the sampling variance σ_i^2. We consider a custom model for s_i^2 as $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the i^{th} area. Rivest and Vandal (2002) and Wang and Fuller (2003) used empirical best linear unbiased prediction (EBLUP) method to obtain the model-based estimates. You and Chapman (2006) considered a hierarchical Bayes (HB) approach and combined the sampling variance model $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ with the small area model (1.3) to construct an integrated model. The integrated model borrows strength for small area estimates and sampling variance estimates simultaneously. The integrated HB modeling approach with $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ has thus been widely used in practice, for example, You (2008, 2016), Dass, Maiti, Ren and Sinha (2012), Sugasawa, Tamae and Kubokawa (2017), Ghosh, Myung and Moura (2018), and Hidiroglou, Beaumont and Yung (2019).

In this paper, we consider both the smoothing and modeling approaches for the sampling variances. In Section 2, we present the EBLUP method based on both the smoothed and direct estimates of the sampling variances. In Section 3, we present the Fay-Herriot HB model and three other HB models based on sampling variance modeling. We compare the effects of sampling variance smoothing and modeling in Section 4 through a real data analysis, and we offer some suggestions in Section 5.

2. Fay-Herriot model using EBLUP approach

Under the Fay-Herriot model (1.3), assuming σ_i^2 and σ_v^2 known in the model, we obtain the best linear unbiased prediction (BLUP) estimator of θ_i as $\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) x_i' \tilde{\beta}$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$ and $\tilde{\beta} = \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i x_i' \right)^{-1} \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i y_i \right)$. To estimate the variance component σ_v^2, we have to first assume σ_i^2 known. There are several methods available to estimate σ_v^2, and we use REML method to estimate σ_v^2. Then the EBLUP of the small area parameter θ_i is obtained as

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) x_i' \hat{\beta}, \tag{2.1}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \sigma_i^2)$ and $\hat{\sigma}_v^2$ is the REML estimator. The estimator for the mean squared error (MSE) of $\hat{\theta}_i$ is given by $\operatorname{mse}(\hat{\theta}_i) = g_{1i} + g_{2i} + 2g_{3i}$, where $g_{1i} = \hat{\gamma}_i \sigma_i^2$ is the leading term, g_{2i} accounts for the variability due to estimation of the regression parameter β, and g_{3i} is due to the estimation of the model variance σ_v^2; see Rao and Molina (2015) for details.

We may use the smoothed or direct estimate of σ_i^2 in (2.1). For sampling variance smoothing, we use a log-linear regression model on the direct sampling variance s_i^2 as suggested in You and Hidiroglou (2012), and the smoothing model is defined as:

$$\log(s_i^2) = \eta_0 + \eta_1 \log(n_i) + \varepsilon_i, \quad i = 1, \ldots, m, \tag{2.2}$$

where the model error term is $\varepsilon_i \sim N(0, \psi^2)$, and ψ^2 is unknown. Let $\hat{\eta}_0$ and $\hat{\eta}_1$ denote the ordinary least square estimates of the regression coefficients η_0 and η_1, and $\hat{\psi}^2$ be the estimated residual variance of the log-linear regression model (2.2). A smoothed estimator of the sampling variance σ_i^2 can be obtained as

$$\tilde{\sigma}_i^2 = \exp(\hat{\eta}_0 + \hat{\eta}_1 \log(n_i)) \exp(\hat{\psi}^2 / 2).$$

The smoothed sampling variances $\tilde{\sigma}_i^2$ can then be used in the EBLUP estimator (2.1) and its MSE computation. This procedure is a common practice, see Rao and Molina (2015).

If direct sampling variance estimate s_i^2 is used in the place of the true sampling variance σ_i^2 in (2.1), then an extra term accounting for the uncertainty of using s_i^2 is needed in the MSE estimator. This term, denoted as g_{4i}, is given as $g_{4i} = 4(n_i - 1)^{-1} \hat{\sigma}_v^4 s_i^4 (\hat{\sigma}_v^2 + s_i^2)^{-3}$; see Rivest and Vandal (2002) and Rao and Molina (2015), page 150. However, using s_i^2 directly in the EBLUP could lead to an over estimation of the model variance σ_v^2 (You, 2010; Rubin-Bleuer and You, 2016), as well as less accurate estimates. We will compare the EBLUP estimates with the HB estimates based on the smoothed and direct sampling variances in Section 4.

3. Fay-Herriot model using HB approach with sampling variance modeling

In this section we first present the Fay-Herriot model in a HB framework. Then we consider three models for the sampling variance modeling. The first model is the one considered in You and Chapman (2006) in which an inverse gamma model is used for the sampling variance σ_i^2 with known vague parameter values. The second model is introduced in You (2016) whereby a log-linear model with random error is used for σ_i^2. The third model is one proposed by Sugasawa et al. (2017) where an inverse gamma model is used for σ_i^2 but with different parameter settings.

HB Model 1: Fay-Herriot model in HB, denoted as FH-HB:

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1,\ \pi(\sigma_v^2) \propto 1.$

Note that in the FH-HB model, the sampling variance σ_i^2 is assumed to be known. Either a smoothed sampling variance $\tilde{\sigma}_i^2$ or a direct sampling variance estimate s_i^2 will be used in place of σ_i^2.

HB Model 2: You-Chapman Model (You and Chapman, 2006), denoted as YCM:

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi_{d_i}^2,\ d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$, where $a_i = 0.0001,\ b_i = 0.0001,\ i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1,\ \pi(\sigma_v^2) \propto 1.$

The full conditional distributions for the Gibbs sampling procedure under both FH-HB and YCM can be found in You and Chapman (2006).

HB Model 3: You (2016) Log-linear model on sampling variances, denoted as YLLM:

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi_{d_i}^2, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\log(\sigma_i^2) \sim N\left(\delta_1 + \delta_2 \log(n_i), \tau^2\right), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1,\ \pi(\delta_1, \delta_2) \propto 1,\ \pi(\sigma_v^2) \propto 1,\ \pi(\tau^2) \propto 1.$

Note that model YLLM uses a log-linear model for the sampling variance σ_i^2, and extends the model proposed by Souza, Moura and Migon (2009) for sampling variances by using $\log(n_i)$ and adding a random effect to the regression part in the model. The full conditional distributions for the Gibbs sampling procedure are given in the Appendix.

HB Model 4: Sugasawa, Tamae and Kubokawa (2017) model shrinking both means and variances, denoted as STKM:

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi_{d_i}^2, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2), \quad i = 1, \ldots, m;$

- $\pi(\sigma_i^2) \sim \mathrm{IG}(a_i, b_i\gamma)$, where a_i and b_i are known constants, $a_i = O(1)$, $b_i = O(n_i^{-1})$;
- Flat priors for unknown parameters: $\pi(\beta) \propto 1$, $\pi(\sigma_v^2) \propto 1$, $\pi(\gamma) \propto 1$.

Note that in STKM, for the inverse gamma model of σ_i^2, we choose $a_i = 2$ and $b_i = n_i^{-1}$ as suggested by Sugasawa et al. (2017). Ghosh et al. (2018) also used the same setting in their study of comparing HB estimators. The full conditional distributions for STKM can be found in Sugasawa et al. (2017).

Note that the Chi-squared sampling variance modeling $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ in the above HB Models 2-4 is based on normality and simple random sampling (Rivest and Vandal, 2002). For complex survey designs, the degrees of freedom d_i may need to be determined more carefully. There is no sound theoretical result for determining the degrees of freedom (Dass et al., 2012). The approximation formula based on non-normal unit level errors provided by Wang and Fuller (2003) and the simulation based guideline of Maples, Bell and Huang (2009) could be useful but require unit level data and an extensive simulation study. A careful determination of the degrees of freedom may provide a reasonably useful approximation. Moreover, Bayesian model fit analysis can also be helpful for model determination.

4. Application

In this section, we apply the models in Sections 2 and 3 to the Canadian Labour Force Survey (LFS) data and compare the EBLUP and HB estimates. The LFS releases monthly unemployment rate estimates for large areas such as the nation and provinces as well as local areas such as Census Metropolitan Areas (CMAs) and Census Agglomerations (CAs) across Canada. The direct LFS estimates for some local areas are not reliable exhibiting very large coefficient of variations (CVs) due to small sample sizes. Model-based estimators are considered to improve the direct LFS estimates. As an illustration, we apply the Fay-Herriot model to the May 2016 unemployment rate estimates at the CMA/CA level, and compare the model-based estimates and the direct estimates with the census estimates to compare the effects of sampling variance smoothing and modeling. Hidiroglou et al. (2019) also compared the model-based LFS estimates with the census estimates. For the unemployment rate estimation, the local area employment insurance monthly beneficiary rate is used as an auxiliary variable in the model. For comparison of point estimates, we compute the absolute relative error (ARE) of the direct and model estimates with respect to the census estimates for each CMA/CA as follows:

$$\mathrm{ARE}_i = \left| \frac{\theta_i^{\mathrm{Census}} - \theta_i^{\mathrm{Est}}}{\theta_i^{\mathrm{Census}}} \right|,$$

where θ_i^{Est} is the direct or the EBLUP/HB estimate and $\theta_i^{\mathrm{Census}}$ is the corresponding census value of the unemployment rate. Then we take the average of AREs over CMA/CAs. For CV, we compute the average CVs of the direct and model-based estimates. We prefer a model with smaller ARE and smaller CV.

We first apply the models to all the 117 CMA/CAs with sample size $\geq 2,$ and then apply them to 92 CMA/CAs with sample size $\geq 5,$ and finally 79 CMA/CAs with sample size $\geq 7.$ Table 4.1 presents the average ARE and the corresponding average CV (in brackets). In Table 4.1, the model with Smoothed sv indicates that a smoothed sampling variance is used, Direct sv indicates that a direct sampling variance estimate is used.

With Smoothed sv, both FH-EBLUP and FH-HB substantially improve the direct survey estimates with much smaller ARE and CV. In particular, FH-HB has the smallest ARE, and FH-EBLUP has the smallest CV. For example, over the 117 areas, the direct LFS estimator has ARE 0.263 with average CV 0.329, FH-EBLUP Smoothed sv has ARE 0.124 with average CV 0.087, FH-HB Smoothed sv has ARE 0.118 with average CV 0.116. The good performance of FH-EBLUP and FH-HB with Smoothed sv indicates that the smoothing GVF (2.2) is very useful and effective in improving the model-based estimates.

With Direct sv, both FH-EBLUP and FH-HB perform the worst among all the models, with almost identical results under this scenario. The other three HB models perform better than the FH-EBLUP and FH-HB using direct sv. YLLM and STKM perform better than YCM with smaller ARE and smaller CV. YLLM and STKM perform very similarly for all the CMA/CA groups, and YLLM consistently has slightly smaller ARE than STKM, but YLLM has slightly larger CV than STKM. For example, over the 117 areas, YLLM has ARE 0.135, STKM has ARE 0.137, and YLLM has average CV 0.123, and STKM has average CV 0.122. YCM has ARE 0.148 with CV 0.136, FH-HB has ARE 0.171 with CV 0.221.

Table 4.1
Comparison of average absolute relative error (ARE) and average CV in parenthesis

CMA/CAs	Direct LFS	FH-EBLUP Smoothed sv	FH-HB Smoothed sv	FH-EBLUP Direct sv	FH-HB Direct sv	YCM Direct sv	YLLM Direct sv	STKM Direct sv
Average over 117 CMA/CAs	0.263	0.124	0.118	0.170	0.171	0.148	0.135	0.137
(sample size ≥ 2)	(0.329)	(0.087)	(0.116)	(0.238)	(0.221)	(0.136)	(0.123)	(0.122)
Average over 92 CMA/CAs	0.216	0.124	0.116	0.133	0.132	0.132	0.125	0.127
(sample size ≥ 5)	(0.262)	(0.076)	(0.103)	(0.123)	(0.123)	(0.121)	(0.117)	(0.116)
Average over 79 CMA/CAs	0.181	0.122	0.113	0.126	0.122	0.122	0.118	0.120
(sample size ≥ 7)	(0.232)	(0.057)	(0.094)	(0.115)	(0.115)	(0.115)	(0.114)	(0.113)

Now we present a Bayesian model comparison using conditional predictive ordinate (CPO) for the four HB models with Direct sv. CPOs are the observed likelihoods based on the cross-validation predictive distribution $f\left(y_i \mid y_{\text{obs}(i)}\right).$ We compute the CPO values for each observed data point $y_{i,\text{obs}}$ and larger CPO indicates that $y_{i,\text{obs}}$ supports the model and a better model fit. For model choice, we can compute the CPO ratio of model A against model B. If this ratio is greater than 1, then $y_{i,\text{obs}}$ supports model A. We compute the CPO ratio for YCM/FH-HB, YLLM/FH-HB and STKM/FH-HB, and count the number of the CPO ratios are larger than 1. We can also plot the CPO values or summarize the CPO values by taking the average of the estimated CPOs. For more detail on CPO, see for example, Gilks, Richardson and Spiegelhalter (1996), page 153, You and Rao (2000), and Molina, Nandram and Rao (2014). Table 4.2 presents the CPO mean and median values over the 117 CMA/CAs and the number of CPO ratios larger than 1.

Table 4.2
Summary of CPO values and CPO ratios over 117 CMA/CAs

	FH-HB Direct sv	**YCM Direct sv**	**YLLM Direct sv**	**STKM Direct sv**
CPO Mean	0.1053	0.1222	0.1242	0.1238
CPO Median	0.0976	0.1004	0.1045	0.1051
# of CPO ratio >1	-	72	78	76

It is clear from Table 4.2 that YCM, YLLM and STKM have larger CPO values than FH-HB, which indicate that the HB model with sampling variance modeling is preferred when the direct sampling variance estimates are used, and YLLM and STKM are better than YCM. For CPO ratios, among the 117 areas, 72 areas/observations support YCM, 78 areas support YLLM and 76 areas support STKM. Therefore more observations support YCM, YLLM and STKM over FH-HB, and YLLM has the most number of CPO ratios that are larger than 1. The CPO comparison is consistent with the results reported in Table 4.1. For other model checking and evaluation methods, see Hidiroglou et al. (2019).

5. Conclusion

In this paper, we compare the model-based estimates under the Fay-Herriot model when sampling variances are smoothed and modeled. As in Hidiroglou et al. (2019), our results indicate that the Fay-Herriot model can provide great improvement for the direct survey estimates for LFS rate estimation, even though more complex models such as unmatched models or time series models could be used (e.g., You, 2008). Among all the estimators, FH-EBLUP and FH-HB using smoothed sampling variances perform the best in terms of ARE and CV reduction. Both FH-EBLUP and FH-HB using direct sampling variance estimates perform the worst. For HB modeling approach, both YLLM and STKM perform very well and are better than YCM, and YLLM is slightly better than STKM in our study. Thus if direct sampling variance estimates are used, YLLM or STKM model is suggested. Alternatively, smoothed sampling variances should be used in the Fay-Herriot model to overcome the sampling variance modeling difficulty as discussed in Section 3. The smoothed sampling variances based on the GVF model given by (2.2) in Section 2 can perform very well as shown in our study.

Appendix

Full conditional distributions and sampling procedure for YLLM

- $\left[\theta_i \,\middle|\, y, \beta, \sigma_i^2, \sigma_v^2\right] \sim N\left(\gamma_i y_i + (1-\gamma_i)\, x_i'\beta, \gamma_i \sigma_i^2\right)$, where $\gamma_i = \sigma_v^2 / \sigma_v^2 + \sigma_i^2$, $i = 1, \ldots, m$;
- $\left[\, \beta \,\middle|\, y, \theta, \sigma_i^2, \sigma_v^2 \,\right] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1} \left(\sum_{i=1}^m x_i \theta_i\right), \sigma_v^2 \left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right)$;
- $\left[\, \sigma_v^2 \,\middle|\, y, \theta, \beta, \sigma_i^2 \,\right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2} \sum_{i=1}^m (\theta_i - x_i'\beta)^2\right)$;
- $\left[\sigma_i^2 \,\middle|\, y, \theta, \beta, \sigma_v^2, \delta, \tau^2\right] \propto f(\sigma_i^2) \cdot h(\sigma_i^2)$, where $f(\sigma_i^2)$ and $h(\sigma_i^2)$ are $f(\sigma_i^2) \sim \mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, and $h(\sigma_i^2) = \exp\left(-\frac{(\log(\sigma_i^2) - z_i'\delta)^2}{2\tau^2}\right)$;
- $\left[\, \delta \,\middle|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \tau^2 \,\right] \sim N_2\left(\left(\sum_{i=1}^m z_i z_i'\right)^{-1} \left(\sum_{i=1}^m z_i \log(\sigma_i^2)\right), \tau^2 \left(\sum_{i=1}^m z_i z_i'\right)^{-1}\right)$;
- $\left[\, \tau^2 \,\middle|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \delta \,\right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2} \sum_{i=1}^m \left(\log(\sigma_i^2) - z_i'\delta\right)^2\right)$.

We use Metropolis-Hastings rejection step to update σ_i^2:

(1) Draw σ_i^{2*} from $\mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$;

(2) Compute the acceptance probability $\alpha(\sigma_i^{2*}, \sigma_i^{2(k)}) = \min\left\{ h(\sigma_i^{2*}) / h(\sigma_i^{2(k)}), 1 \right\}$;

(3) Generate u from Uniform (0, 1), if $u < \alpha(\sigma_i^{2*}, \sigma_i^{2(k)})$, the candidate σ_i^{2*} is accepted, $\sigma_i^{2(k+1)} = \sigma_i^{2*}$; otherwise σ_i^{2*} is rejected, and set $\sigma_i^{2(k+1)} = \sigma_i^{2(k)}$.

Acknowledgements

I would like to thank the Editor, the Associate Editor and one referee for their constructive comments and suggestions to improve the paper.

References

Dass, S.C., Maiti, T., Ren, H. and Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, 38, 2, 173-187. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012002/article/11756-eng.pdf.

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

Ghosh, M., Myung, J. and Moura, F.A.S. (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44, 1, 101-115. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54959-eng.pdf.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

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

Maples, J., Bell, W. and Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.

Molina, I, Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics*, 8, 852-885.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, 2nd Edition. New York: John Wiley & Sons, Inc.

Rivest, L.P., and Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.

Rubin-Bleuer, S., and You, Y. (2016). Comparison of some positive variance estimators for the Fay-Herriot small area model. *Survey Methodology*, 42, 1, 63-85. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016001/article/14542-eng.pdf.

Souza, D.F., Moura, F.A.S. and Migon, H.S. (2009). Small area population prediction via hierarchical models. *Survey Methodology*, 35, 2, 203-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11042-eng.pdf.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10614-eng.pdf.

You, Y. (2010). Small area estimation under the Fay-Herriot model using different model variance estimation methods and different input sampling variances. Methodology branch working paper, SRID-2010-003E, Statistics Canada, Ottawa, Canada.

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Methodology branch working paper, ICCSMD-2016-03-E, Statistics Canada, Ottawa, 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 Hidiroglou, M. (2012). Sampling variance smoothing methods for small area proportion estimators. Methodology Branch Working Paper, SRID-2012-08E, Statistics Canada, Ottawa, Canada; Presented as an invited paper at the Fields Institute Symposium on the Analysis of Survey Data and Small Area Estimation, Carleton University, Ottawa, 2012.

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, December 2021
Vol. 47, No. 2, pp. 371-382
Statistics Canada, Catalogue No. 12-001-X

Assessing the coverage of confidence intervals under nonresponse. A case study on income mean and quantiles in some municipalities from the 2015 Mexican Intercensal Survey

Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz and Guillermina Eslava-Gómez[1]

Abstract

This note presents a comparative study of three methods for constructing confidence intervals for the mean and quantiles based on survey data with nonresponse. These methods, empirical likelihood, linearization, and that of Woodruff's (1952), were applied to data on income obtained from the 2015 Mexican Intercensal Survey, and to simulated data. A response propensity model was used for adjusting the sampling weights, and the empirical performance of the methods was assessed in terms of the coverage of the confidence intervals through simulation studies. The empirical likelihood and linearization methods had a good performance for the mean, except when the variable of interest had some extreme values. For quantiles, the linearization method had a poor performance, while the empirical likelihood and Woodruff methods had a better one, though without reaching the nominal coverage when the variable of interest had values with high frequency near the quantile of interest.

Key Words: Confidence interval estimation; Empirical likelihood; Linearization; Missing at random; Nonresponse; Two-phase sampling.

1. Introduction

The 2015 Mexican Intercensal survey (MIC2015) conducted by the National Institute of Statistics and Geography (INEGI, 2015) collected information nationwide, using a probability sampling design in 1,643 municipalities and through a census in 814 municipalities. In this study we use the census data corresponding to 441 municipalities from the state of Oaxaca.

We focus on *income* as the variable of interest which exhibits a nonresponse rate of about 22.5%. Considering the respondents, the distribution of *income* has a high skewness mainly due to the presence of extreme values, and shows some values with high frequency.

The objective of this study is to assess the empirical coverage rate of confidence intervals (CI) computed by three methods for the population mean and population quantiles, 0.1, 0.5 and 0.9, in survey data with nonresponse. Two-phase sampling is used with a random sample selected in the first phase, while in the second the sample is split into respondents and nonrespondents considering the nonresponse pattern of *income* in the census data. A response propensity model is used to adjust the weights for nonresponse.

---

[1] Omar De La Riva Torres, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: odelariva@ciencias.unam.mx; Gonzalo Pérez-de-la-Cruz, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: gonzalo.perez@ciencias.unam.mx; Guillermina Eslava-Gómez, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: eslava@ciencias.unam.mx.

For the population mean, we consider the Hájek estimator and two methods for computing CIs: empirical likelihood (Berger, 2020) and linearization (Särndal, Swenson and Wretman, 1992, Sections 5.2 and 5.7). Concerning the population quantiles, we consider the point estimator obtained by interpolation of the distribution function as in Woodruff (1952) and Graf and Tillé (2014), and three methods for computing CIs: empirical likelihood (Berger, 2020), Woodruff (Woodruff, 1952) and linearization (Deville, 1999). These methods are described in Section 2, the numerical results are presented in Section 3 for the MIC2015 data and in Section 4 for some simulated populations. Some final comments are given in Section 5.

2. Three methods for estimating confidence intervals

2.1 Estimation using two-phase sampling

We consider a finite population $U = \{1, 2, \ldots, N\}$ and a probability sample $s \subset U$ of fixed size n, with first and second-order inclusion probabilities π_k and π_{kl}, $k, l \in U$, $k \neq l$. Let y_k be the k^{th} value of the variable of interest y, and let θ_0 be a population parameter and $\hat{\theta}_0$ an estimator of θ_0.

We assume that the value y_k is available for a subset $r \subset s$ only. Let ϕ_k denote the response probability for unit k. Let I_k be a response indicator variable such that $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s \setminus r$. We also assume that there is a vector of auxiliary variables $\mathbf{x}$ observed for all $k \in s$. We make the missing at random (MAR) assumption:

$$P(I_k = 1 \mid y_k, \mathbf{x}_k) = P(I_k = 1 \mid \mathbf{x}_k) = \phi_k \ \forall\, k \in U.$$

The response probabilities ϕ_k are used for adjusting the design weights. We assume that the sampling design and the response mechanism are independent as in Berger (2020). Borrowing from two-phase sampling theory (Särndal et al., 1992, Section 9.3) the weights adjusted for nonresponse are defined as $\pi_k^{*-1} = 1/(\pi_k \phi_k)$ and the second-order inclusion probabilities as $\pi_{kl}^{*} = \pi_{kl}\, \phi_k\, \phi_l, k, l \in U,\ k \neq l.$

The interest lies in estimating the population mean $\bar{Y} = \sum_{k \in U} y_k / N$ and the population quantile given by

$$Y_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right)\left[qN - N_{(d-1)}\right]}{N_{(d)} - N_{(d-1)}}, \tag{2.1}$$

where $y_{(i)}$ is the value for the i^{th} unit arranged in increasing order, $d = \min\{l : qN < N_{(l)}, l = 1, \ldots, N\}$ and $N_{(l)} = \sum_{j \in U} I\left(y_j \leq y_{(l)}\right), l = 1, \ldots, N.$ Formula (2.1) is obtained by considering a piecewise linear interpolation of the step distribution function $F(y) = \sum_{k \in U} I(y_k \leq y) / N,$ where $I\left(y_k \leq y\right) = 1$ when $y_k \leq y.$

These population parameters are respectively estimated by

$$\hat{\hat{Y}} = \frac{\sum_{k \in r} y_k / \pi_k^*}{\sum_{k \in r} 1/\pi_k^*}, \tag{2.2}$$

and

$$\hat{Y}_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right)\left(q\hat{N} - \hat{N}_{(d-1)}\right)}{\hat{N}_{(d)} - \hat{N}_{(d-1)}}, \tag{2.3}$$

where $d = \min\left\{l : q\hat{N} < \hat{N}_{(l)}, l = 1, \ldots, n_r\right\}$, $\hat{N} = \sum_{k \in r} 1/\pi_k^*$, $\hat{N}_{(l)} = \sum_{k \in r} I(y_k \le y_{(l)}) / \pi_k^*$ and $n_r = \sum_{k \in s} I_k$.

These estimators and the CIs described in the following subsection are based on the assumption that the response probabilities ϕ_k are known, unlike Berger (2020) and Kim and Kim (2007). However, we use $\hat{\phi}_k$ instead of ϕ_k in the simulation studies, where

$$\hat{\phi}_k = \frac{\exp\left(\mathbf{x}_k^{\top}\hat{\beta}\right)}{1 + \exp\left(\mathbf{x}_k^{\top}\hat{\beta}\right)} \ \forall \ k \in s,$$

with $\hat{\beta}$ obtained by fitting a logistic regression using s. This leads to the estimators known as the empirical double expansion estimators (Haziza and Beaumont, 2017).

2.2 Methods for estimating confidence intervals

2.2.1 Linearization

The linearization method relies on the assumption that the distribution of $\hat{\theta}_0$ is approximately normal. A CI for θ_0 is

$$\left[\hat{\theta}_0 - z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2},\ \hat{\theta}_0 + z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2}\right], \tag{2.4}$$

where $1 - \alpha$ is the confidence level, also known as nominal coverage; see Särndal et al. (1992, expression 5.2.3). In practice $V(\hat{\theta}_0)$ is estimated. For the estimators given by (2.2) and (2.3), a variance estimator is given by

$$\hat{V}(\hat{\theta}_0) = \sum_{k \in r}\sum_{l \in r} \frac{(\pi_{kl}^* - \pi_k^*\pi_l^*)}{\pi_{kl}^*} \frac{\hat{z}_k}{\pi_k^*} \frac{\hat{z}_l}{\pi_l^*}, \tag{2.5}$$

where $\hat{z}_k = \left(y_k - \hat{\hat{Y}}\right) / \hat{N}$ for $\hat{\hat{Y}}$ (Särndal et al., 1992, Result 5.7.1) and $\hat{z}_k = -\left(I(y_k \le \hat{Y}_q) - q\right) / \left(f(\hat{Y}_q)\hat{N}\right)$ for $\hat{Y}_q$ (Deville, 1999). The *density function* f was obtained in two ways: a) using a Gaussian kernel as in Osier (2009) and b) using the nearest neighbour technique as in Graf and Tillé (2014). We present the results pertaining to a), since the technique in b) led to similar results.

We note that using (2.5) with $\hat{\phi}_k$ instead of ϕ_k might lead to overestimation of the variance of the empirical double expansion estimator and to wider CIs; see the expression (17) in Kim and Kim (2007) associated with the estimators of $\bar{Y}$.

2.2.2 Empirical likelihood method

The empirical likelihood approach assumes that θ_0 is the unique solution of the estimating equation $G(\theta) = \sum_{k\in U} g_k(\theta) = 0$ for a given function g_k. In particular, we use:

i) $g_k(\theta) = y_k - \theta$ for $\theta_0 = \bar{Y}$.

ii) $g_k(\theta) = \rho(y_k, \theta) - q$ for $\theta_0 = Y_q$, where $\rho(y_k, \theta) = I(y_k < y_{(l)}) + I(y_k = y_{(l)})(\theta - y_{(l-1)}) / (y_{(l)} - y_{(l-1)})$ and $l = \min\{j : y_{(j)} > \theta\}$.

The empirical log-likelihood function in Berger and De La Riva Torres (2016) for a one-stage sampling design without stratification or auxiliary information is

$$\ell_{\max}(\theta) = \max_{m_k : k\in s} \left\{ \sum_{k\in s} \log(m_k) : m_k > 0,\ \sum_{k\in s} m_k g_k(\theta) = 0,\ \sum_{k\in s} m_k \pi_k = n \right\}, \tag{2.6}$$

where $\{m_k : k \in s\}$ satisfies the design and the parameter constraints $\sum_{k\in s} m_k \pi_k = n$ and $\sum_{k\in s} m_k g_k(\theta) = 0.$

In the presence of nonresponse, we use (2.6) replacing $\sum_{k\in s} m_k g_k(\theta) = 0$ with $\sum_{k\in s} m_k I_k g_k(\theta) / \phi_k = 0.$ A CI for θ_0 is given by

$$\left[\min\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\},\ \max\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\}\right], \tag{2.7}$$

where $\hat{R}(\theta) = 2\{\ell_{\max}(\hat{\theta}_0) - \ell_{\max}(\theta)\}$ and $\chi_1^2(\alpha)$ is the $(1-\alpha)$-quantile of the χ_1^2 distribution. The estimator $\hat{\theta}_0 := \text{argmax}_{\{\theta\}}\, \ell_{\max}(\theta)$ corresponds to (2.2) and (2.3), respectively.

We computed (2.7) using a root search method, calculating $\hat{R}(\theta)$ for several values of $\theta,$ where $\ell_{\max}(\theta)$ for a given value θ was obtained by a modified Newton-Raphson algorithm as in Wu (2004).

2.2.3 Woodruff method for quantiles

The method of Woodruff (1952) is based on the estimated distribution function $\hat{F}(y).$ For a quantile $Y_q,$ the variance of $\hat{F}(Y_q)$ can be approximated using the Taylor linearization method with linearized variable $z_k = (I(y_k \le Y_q) - q) / N,$ while the variance is estimated using (2.5) with $\hat{z}_k = (I(y_k \le \hat{Y}_q) - q) / \hat{N}.$ Assuming normality of $\hat{F}(Y_q)$ and using (2.4), it is possible to find a CI $[c_1, c_2]$ for $F(Y_q),$ which leads to $[\hat{F}^{-1}(c_1),\ \hat{F}^{-1}(c_2)]$ for $Y_q.$

3. Empirical study based on data from the populations MIC2015_Oax and MIC2015_Oax$_{\text{trunc}}$

The population census considered in this work consisted of 208,101 inhabitants with complete responses in a vector $\mathbf{x}$ of six auxiliary variables: *age*, *educational level*, *employment status*, *gender*, *indigenous language*, and *marital status*. The variable of interest y corresponds to the monthly *income*.

A logistic regression with some two-way interactions was fitted to the 208,101 observations, with response variable $I_k = 1$ if an *income* value was given by individual k, and $I_k = 0$ if it was not, and the vector $\mathbf{x}$ of six explanatory variables. This model was then applied to the population of 161,296 individuals, which corresponds to those with $I_k = 1.$ This led to a set of response propensity values $\phi = (\phi_1, \ldots, \phi_{161,296}).$

The set of 161,296 respondents, with response propensities $\phi_1, \ldots, \phi_{161,296},$ is referred to as MIC2015_Oax population. The distribution of *income* in this population is highly asymmetric partly due to the presence of some very large values. When removing the 80 observations with *income* larger than or equal to 50,000, we obtain a truncated population referred to as MIC2015_Oax$_{\text{trunc}}$, which is also used in our experiments.

The step distribution function of *income* has only 913 and 887 jumps in each population respectively, with some large jumps at *income* values that are near the quantiles of interest. In particular, $y =$ 2,571 accounts for 7.3% of the distribution and is very close to the quantile $Y_{0.5}$; $y = 643$ (1.1%) and $y = 857$ (4.2%) are close to $Y_{0.1}$; whereas $y =$ 6,429 (2.8%) and $y =$ 7,000 (1.1%) are close to $Y_{0.9}$.

3.1 Numerical results

For each population, the coverage rate of the CI for each method was estimated as follows:

1. A simple random sample s of size $n \in \{1{,}000; 5{,}000\}$ was selected.
2. For each unit $k \in s$ with response propensity ϕ_k, we generated I_k from a $\text{Bernoulli}(\phi_k)$.
3. Two cases were considered: a) full response and b) average nonresponse rate of 22.5%. For the latter, a logistic regression with two-way interactions, with I as the response and the six explanatory variables, was selected by forward selection using the BIC criterion. The estimated response probabilities $\hat{\phi}_k$ were obtained with the selected model.
4. 90% CIs were computed using linearization (Lin), empirical likelihood (EL) and the Woodruff (W) method for quantiles.
5. Steps 1 to 4 were repeated $M =$ 5,000 times and the coverage rate for each method and for each parameter was calculated as the proportion of CIs that covered the corresponding parameter value.

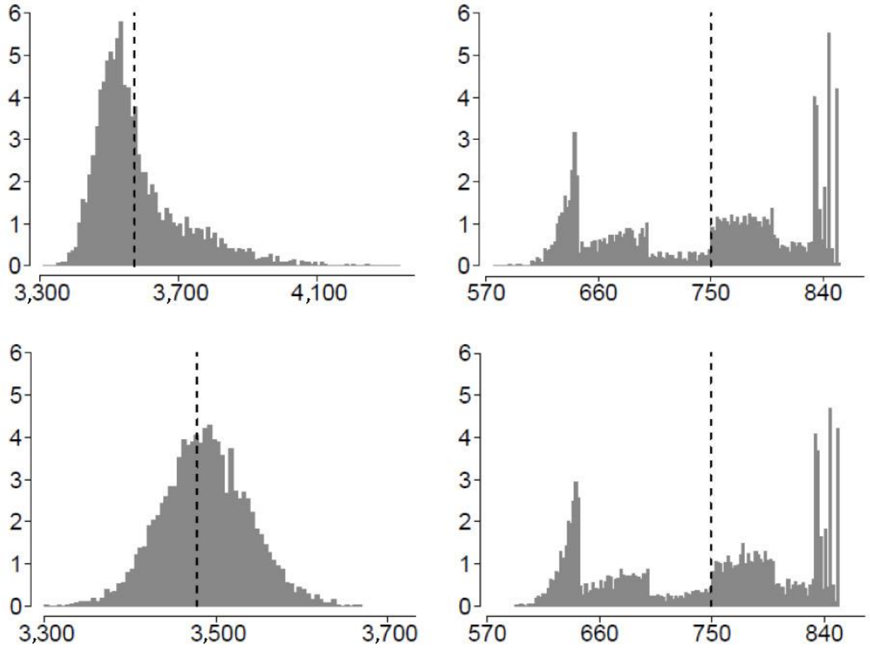
Table 3.1 shows the results for $n = 5{,}000$. Table 3.2 shows the absolute value of the percent relative bias, $\mathrm{RB} = 100\left(\bar{\hat{\theta}} - \theta_0\right) \Big/ \theta_0$ with $\bar{\hat{\theta}}_0 = \sum \hat{\theta}_{0i} \Big/ M$, and of the percent relative root mean square error, $\mathrm{RRMSE} = 100\left\{ \sum (\hat{\theta}_{0i} - \theta_0)^2 \Big/ M \right\}^{1/2} \Big/ \theta_0$. Figure 3.1 presents the distribution of the $M = 5{,}000$ estimates for the nonresponse scenario for each parameter; the corresponding distributions for the full response scenario are qualitatively similar. The results for $n = 1{,}000$ are omitted since they were similar to those obtained with $n = 5{,}000$. From Tables 3.1 and 3.2 and Figure 3.1, we make the following remarks:

a) For $\bar{Y}$, Lin and EL methods perform similarly: they have a poor performance (coverage as low as 72.9%) for MIC2015_Oax, and a good one for MIC2015_Oax$_{\text{trunc}}$, reaching the nominal level with similar tail error rates and CI average length. Figure 3.1 a) shows that the distribution of $\hat{\bar{Y}}$ is symmetric for MIC2015_Oax$_{\text{trunc}}$ and highly asymmetric for MIC2015_Oax; this asymmetry seems to be related to the 80 extreme *income* values not present in MIC2015_Oax$_{\text{trunc}}$.

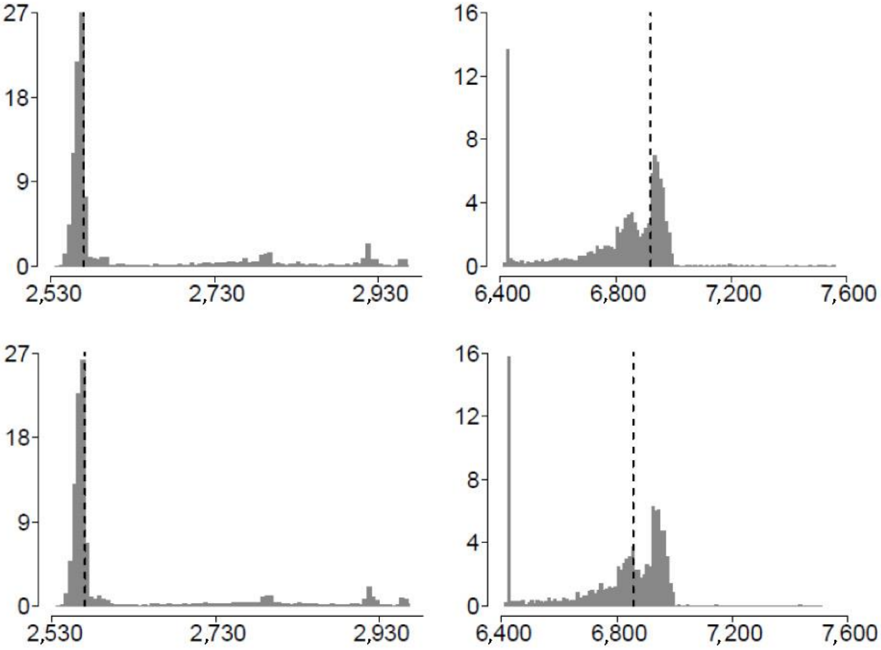
b) For quantiles, Lin method has a poor performance with the shortest CI average length in both scenarios, in spite of the expected overestimation of the variance in the nonresponse scenario. This method relies on the normality of $\hat{Y}_q$, but Figures 3.1 b), c) and d) show that the distribution of $\hat{Y}_q$ is far from being symmetric and unimodal, with modes around *income* values with high frequency. Especially for $Y_{0.1}$, where the coverage rate is as low as 31.4%, the distribution of $\hat{Y}_{0.1}$ is multimodal with a high proportion of values that are farther from $Y_{0.1}$ than half the CI average length. EL and W methods generally perform well, except for $Y_{0.5}$ in the full response scenario and for $Y_{0.9}$ in MIC2015_Oax. The low coverages seem to be related to the observed high frequency of the two *income* values 2,571 (7.3%) and 6,429 (2.8%). The first one is very close to $Y_{0.5} = 2{,}570$ and some of the CIs for $Y_{0.5}$ are too narrow when $\hat{Y}_{0.5} < 2{,}571$. The second one is farther from $Y_{0.9}$ in MIC2015_Oax than in MIC2015_Oax$_{\text{trunc}}$, reducing the proportion of CIs that cover $Y_{0.9}$ when $\hat{Y}_{0.9} \approx 6{,}429$, see Figure 3.1 d), where $Y_{0.9} = 6{,}921$ in MIC2015_Oax and $Y_{0.9} = 6{,}856$ in MIC2015_Oax$_{\text{trunc}}$.

c) Table 3.2 shows that the RB is small, less than 3.3%, for all parameters. When only a simple adjustment with the percentage of nonresponse is applied (not shown in this note), the RB is larger and all the methods have a very poor performance. These results suggest that the use of a propensity model helps to obtain a RB comparable with that of the full response case. For $Y_{0.1}$, the empirical double expansion estimator is even less biased than the one associated with the full response scenario; however their RRMSE are comparable and the largest among those for the parameters of interest, since the distribution of the estimators is multimodal in both scenarios, see Figure 3.1 b).

Figure 3.1 **Distribution of the $M = 5{,}000$ estimates of $\bar{Y}$ in a), and of $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$ in b) to d) for the case with an average nonresponse of 22.5% and $n = 5{,}000$. The upper panel corresponds to MIC2015_Oax and the lower to MIC2015_Oax$_{\text{trunc}}$. The dotted lines indicate the population values $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$.**



a) $\widehat{\bar{Y}}$ b) $\widehat{Y}_{0.1}$



c) $\widehat{Y}_{0.5}$ d) $\widehat{Y}_{0.9}$

Table 3.1
Coverages of 90% CIs for the parameters $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, for y = *income*. Average nonresponse of 22.5% (*NR*) and Full response (*Full*)

Parameter θ_0	Method	Coverage %		Lower tail err. rates %		Upper tail err. rates %		CI average length	
		NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
MIC2015_Oax									
$\bar{Y}$	EL	79.1*	72.9*	4.5	3.8*	16.4*	23.3*	370.8	335.4
	Lin	80.6*	73.8*	0.3*	0.1*	19.1*	26.1*	345.3	309.9
$Y_{0.1}$	EL	91.1*	90.5	6.0*	4.1*	2.9*	5.4	192.8	179.5
	W	90.6	89.8	6.2*	4.2*	3.2*	6.0*	191.2	177.1
	Lin	37.9*	35.1*	28.9*	19.2*	33.3*	45.7*	114.4	89.4
$Y_{0.5}$	EL	88.2*	82.3*	1.6*	0.7*	10.2*	17.1*	274.6	225.8
	W	88.0*	81.0*	1.7*	0.7*	10.3*	18.3*	274.4	224.1
	Lin	79.0*	88.0*	21.0*	12.0*	0.0*	0.0*	152.2	127.3
$Y_{0.9}$	EL	84.0*	83.0*	2.6*	2.7*	13.3*	14.3*	527.2	470.2
	W	86.1*	84.3*	2.5*	2.7*	11.4*	13.0*	533.8	474.2
	Lin	72.3*	73.5*	0.4*	0.1*	27.3*	26.4*	392.8	346.6
MIC2015_Oax$_{\text{trunc}}$									
$\bar{Y}$	EL	90.5	90.6	6.4*	4.4	3.0*	5.0	173.7	147.5
	Lin	90.8*	90.1	5.7*	4.2*	3.5*	5.6*	171.2	145.0
$Y_{0.1}$	EL	89.8	89.9	6.8*	4.0*	3.4*	6.1*	191.7	178.7
	W	89.1*	89.1*	7.0*	4.1*	4.0*	6.8*	190.0	176.2
	Lin	35.5*	31.4*	29.4*	21.0*	35.1*	47.6*	104.9	81.4
$Y_{0.5}$	EL	87.2*	80.4*	1.6*	0.9*	11.2*	18.7*	267.8	218.5
	W	87.1*	79.3*	1.6*	0.9*	11.3*	19.8*	267.7	216.7
	Lin	80.0*	87.4*	20.0*	12.6*	0.0*	0.0*	144.9	121.0
$Y_{0.9}$	EL	90.3	90.1	4.4	4.4*	5.3	5.5	521.3	470.8
	W	92.3*	91.9*	4.3*	4.3*	3.5*	3.8*	528.2	475.3
	Lin	75.6*	77.0*	0.1*	0.1*	24.3*	23.0*	411.7	365.5

$*$ Coverages and tail error rates significantly different from 90% and 5% respectively (Feller, 1968, page 182). p-value < 5%

MIC2015_Oax: $N = 161{,}296$, $\rho = 0.08$, $\gamma = 89.9$; MIC2015_Oax$_{\text{trunc}}$: $N = 161{,}216$, $\rho = 0.21$, $\gamma = 3.48$, where $\rho = \operatorname{corr}(y, \phi)$ and $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Table 3.2
Percent relative bias (RB) and percent relative root mean squared error (RRMSE) of estimators of $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, based on 5,000 samples. Average nonresponse of 22.5% (*NR*) and Full response (*Full*)

Population	$\bar{Y}$				$Y_{0.1}$				$Y_{0.5}$				$Y_{0.9}$			
	$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$	
	NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
MIC2015_Oax	0.30	0.01	3.8	3.2	0.58	3.13	10.4	9.9	1.96	0.93	4.8	3.4	1.79	1.68	3.3	3.0
MIC2015_Oax$_{\text{trunc}}$	0.27	0.02	1.5	1.3	0.71	3.23	10.5	10.0	1.87	0.96	4.8	3.5	1.05	0.95	3.0	2.8

4. Simulated populations

In order to control the asymmetry of the distribution of y, the correlation $\operatorname{corr}(y, \phi)$ and the percentage of nonresponse, we simulated two symmetric and two asymmetric populations of size N = 50,000 with a variable of interest y and six auxiliary variables $x_1, \ldots, x_6$, as follows.

1. 50,000 simple random samples for each of the variables $x_1, \ldots, x_6$ were generated independently from a $N(0, 1)$ distribution.
2. The response probabilities ϕ_k were obtained using a logistic regression with $\beta_1 = \cdots = \beta_6 = -0.3$ and β_0 chosen so that the average nonresponse was equal to 24.8%.
3. Two settings were considered for the distribution of y:
   - i) Symmetric. y_k was generated from a $N(\mu, \sigma^2)$, with $\mu = 1 + 2.16\ \mathrm{corr}(y, x) \sum_{j=1}^{6} x_{kj}$ and $\sigma^2 = 4.67 * (1 - 6\,\mathrm{corr}^2(y, x))$, where $\mathrm{corr}(y, x) = \mathrm{corr}(y, x_j)$, $j \in \{1, \ldots, 6\}$.
   - ii) Asymmetric. $y_k = \exp(z_k)$, where z_k was generated from a $N(\mu, \sigma^2)$, with $\mu = \mathrm{corr}(z, x) \sum_{j=1}^{6} x_{kj}$ and $\sigma^2 = 1 - 6\,\mathrm{corr}^2(z, x)$, where $\mathrm{corr}(z, x) = \mathrm{corr}(z, x_j)$, $j \in \{1, \ldots, 6\}$.

   Both $\mathrm{corr}(y, x)$ and $\mathrm{corr}(z, x)$ were chosen so that $\mathrm{corr}(y, \phi)$ was approximately equal to -0.2 or -0.8.

4.1 Numerical results

The coverage rate of the CIs was computed as in Section 3.1, with $n = 500$ and using a logistic regression without interactions to obtain $\hat{\phi}_k$, $k \in s$.

Table 4.1 reports the results for the populations with $\mathrm{corr}(y, \phi) = -0.8$. Table 4.2 shows the $|\mathrm{RB}|$ and $|\mathrm{RRMSE}|$ of the estimators. Numerical results for populations with $\mathrm{corr}(y, \phi) = -0.2$ are omitted since they were similar to those obtained for populations with $\mathrm{corr}(y, \phi) = -0.8$. We make the following observations:

a) EL and Lin methods have a similar reasonable performance for $\bar{Y}$, though the upper tail error rates are larger than 5% in the asymmetric population.

b) For quantiles, Lin method has the lowest and the highest coverage of 85.7 and 97.9% respectively. Unlike the distribution of $\hat{Y}_q$ shown in Figure 3.1, the distribution of $\hat{Y}_q$ for the simulated populations is symmetric and unimodal in all cases. EL and W methods perform well, reaching the nominal level in all cases with comparable tail error rates and CI average length.

c) Weighting adjustment in the nonresponse scenario helps to get a RB similar to that of the full response.

d) The coverage rate of the CI for each method is larger in the nonresponse scenario than in the full response, and in some cases it is also larger than the nominal level. This might be related to the impact of having treated the $\hat{\phi}_k$ as fixed in the nonresponse case.

Table 4.1
Simulated populations. Coverages of 90% CIs for $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, for y with $\mathrm{corr}(y, \phi) = -0.8$. Average nonresponse of 24.8% (*NR*) and Full response (*Full*)

Parameter θ_0	Method	Coverage %		Lower tail err. rates %		Upper tail err. rates %		CI average length	
		NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
Asymmetric									
$\bar{Y}$	EL	90.9*	88.6*	2.2*	5.1	6.9*	6.3	0.50	0.32
	Lin	90.1	88.6*	0.5*	3.2*	9.4*	8.2*	0.48	0.31
$Y_{0.1}$	EL	90.6	89.5	4.0*	4.6	5.4	5.9	0.07	0.07
	W	90.3	89.5	3.6*	4.1*	6.2*	6.4*	0.07	0.07
	Lin	97.9*	97.4*	1.2*	1.5*	1.0*	1.2*	0.10	0.10
$Y_{0.5}$	EL	93.6*	90.5	3.0*	4.4*	3.4*	5.1	0.22	0.19
	W	93.5*	90.4	2.9*	4.4	3.6*	5.1	0.22	0.19
	Lin	92.5*	88.9*	2.9*	5.0	4.6	6.2*	0.21	0.18
$Y_{0.9}$	EL	92.7*	90.3	2.6*	4.1*	4.7	5.7*	1.35	0.91
	W	93.0*	90.4	2.7*	4.6	4.3*	5.0	1.36	0.92
	Lin	87.4*	85.7*	2.6*	4.1*	10.1*	10.2*	1.23	0.87
Symmetric									
$\bar{Y}$	EL	93.6*	90.2	3.3*	5.0	3.1*	4.8	0.40	0.32
	Lin	93.5*	89.9	3.1*	5.2	3.4*	4.9	0.39	0.32
$Y_{0.1}$	EL	91.2*	90.9*	3.6*	3.9*	5.2	5.2	0.59	0.55
	W	91.0*	90.6	3.2*	3.6*	5.8*	5.8*	0.59	0.56
	Lin	88.6*	88.2*	5.8*	5.9*	5.7*	6.0*	0.57	0.53
$Y_{0.5}$	EL	92.8*	90.1	3.3*	4.6	3.9*	5.3	0.46	0.39
	W	92.9*	90.1	3.1*	4.6	4.0*	5.3	0.46	0.39
	Lin	92.4*	90.2	3.4*	4.7	4.2*	5.2	0.46	0.39
$Y_{0.9}$	EL	92.9*	90.4	2.2*	3.6*	4.9	6.0*	0.76	0.54
	W	93.3*	90.3	2.2*	4.2*	4.5	5.4	0.77	0.54
	Lin	90.3	89.1*	2.1*	3.2*	7.6*	7.7*	0.74	0.53

* Coverages and tail error rates significantly different from 90% and 5% respectively (Feller, 1968, page 182). p-value < 5%
Symmetric: $\gamma = 0.02$; Asymmetric: $\gamma = 6.2$; where $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Table 4.2
Percent relative bias (RB) and percent relative root mean squared error (RRMSE) of estimators of $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, based on 5,000 samples. Average nonresponse of 24.8% (*NR*) and Full response (*Full*)

Population	$\bar{Y}$				$Y_{0.1}$				$Y_{0.5}$				$Y_{0.9}$			
	$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$		$\lvert$RB$\rvert$		$\lvert$RRMSE$\rvert$	
	NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
Asymmetric	0.11	0.13	9.0	5.9	0.56	0.53	7.8	7.5	0.00	0.07	6.0	5.7	0.53	0.43	9.7	7.6
Symmetric	0.12	0.09	10.3	9.5	0.98	0.91	10.2	9.7	0.33	0.35	12.7	11.8	0.43	0.34	5.5	4.3

5. Conclusions

Considering the distribution of y (*income*), it was observed that a poor performance of a method in the full response scenario generally corresponded to a poor one in the nonresponse scenario, although in several cases the coverage rate was larger in the latter. This suggests that having treated the $\hat{\phi}_k$ as fixed had little effect on the performance of the methods as compared to the impact of the characteristics of the distribution of y. Extreme values were related to a low coverage of the CIs for the mean for both empirical likelihood and linearization methods. The presence of values with high frequency near a quantile of interest also had an impact on the coverage of its CIs; this might be related to the behavior of the step distribution function, where the jumps in $F(y)$ and $\hat{F}(y)$ are usually required to be small in order to obtain a good performance of the Woodruff method (Lohr, 2010, page 390). In general, the linearization method had a poor performance for quantiles, while the performance of empirical likelihood and of Woodruff were similar and better; this behavior has also been observed in Berger and De La Riva Torres (2016). While Woodruff method is simple and easy to implement, an advantage of the empirical likelihood method is that it can be used for parameters other than quantiles.

Acknowledgements

Omar De La Riva Torres was supported by the Post Doctoral Scholarship Program of the National Autonomous University of Mexico.

References

Berger, Y.G. (2020). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, 72, 91-121.

Berger, Y.G., and De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 2, 319-341.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-eng.pdf.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications: Volume I*. New York: John Wiley & Sons, Inc.

Graf, E., and Tillé, Y. (2014). Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology*, 40, 1, 61-79. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14000-eng.pdf.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 2, 206-226.

INEGI (2015). *Encuesta Intercensal 2015*. Instituto Nacional de Estadística, Geografía e Informática. https://www.inegi.org.mx/programas/intercensal/2015/.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 4, 501-514.

Lohr, S.L. (2010). *Sampling: Design and Analysis.* Brooks/Cole.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 3, 167-195.

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

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2021.

- P.G. Andersson, *Stockholm University*
- S. Arima, *Uniersità del Salento*
- B. Baffour, *Australian National University*
- G. Barcarolli, *Independent consultant*
- E. Ben-David, *U.S. Census Bureau*
- J. Billiet, *Ku Leuven*
- C. Bocci, *Statistics Canada*
- D. Bonnery, *University of Cambridge*
- H.J. Boonstra, *Statistics Netherlands*
- C. Boulet, *Statistics Canada*
- J.M. Brick, *Westat Inc.*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- I. Carrillo, *Statistics Canada*
- L. Castro-Martin, *Roma Tre University*
- S. Chatterjee, *University of Minnesota*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- Q. Chen, *Columbia University*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- M. Cohen, *American Institutes for Research*
- D. da Silva, *Federal University of Rio Grande do Norte*
- S. Das, *Maastricht University*
- G. Datta, *University of Georgia*
- T. De Waal, *Centraal Bureau voor de Statistiek*
- J. Dever, *RTI International*
- P. Duchesne, *Université de Montréal*
- L. Dumitrescu, *Victoria University of Wellington*
- S. Eckman, *Research Triangle Institute*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- A. Erciulescu, *Westat Inc.*
- V. Estevao, *Statistics Canada*
- B. Fay, *Westat Inc.*
- W.A. Fuller, *Iowa State University*
- C. Girard, *Statistics Canada*
- D. Haziza, *University of Ottawa*
- S. Heeringa, *University of Michigan*
- M.A. Hidiroglou, *Statistics Canada*
- M. Hirose, *Kyushu University*
- N. Hoshino, *Kanazawa University - Kakuma Campus*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
- K. Irimata, *U.S. Census Bureau*
- R. Jauslin, *Université de Neuchâtel*
- D. Judkins, *ABT Associates Inc Bethesda*
- A. Junior, *Senac, prospecting and evaluation management*
- A. Kaplan, *Colorado State University*
- A. Keller, *U.S. Census Bureau*
- C. Kern, *University of Mannheim*
- J.K. Kim, *Iowa State University*
- H. Kim, *University of Cincinnati*
- S. Kolenikov, *Abt Associates*
- P.S. Kott, *RTI International*
- P. Lahiri, *University of Maryland*
- T. Laitila, *Örebro University*
- D. Lee, *University of Alabama*
- S. Lee, *University of Michigan*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- D. Liao, *Research Triangle Institute*
- B. Liseo, *University of Rome La Sapienza, MEMOTEF*
- Y. Liu, *East China Normal University*
- G. Loosveldt, *KU Leuven*
- J. Maples, *U.S. Census Bureau*
- D. Marella, *Roma Tre University*
- K. McConville, *Reed College*
- T. Merkouris, *Athens University of Economics and Business Statistics*
- I. Molina, *Universidad Carlos III de Madrid*
- B. Nandram, *Worcester Polytechnic Institute*
- D. Nedyalkova, *Federal Statistical Office*
- A. Norberg, *Statistics Sweden*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- M. Ranalli, *Università degli Studi di Perugia*
- J.N.K. Rao, *Carleton University*
- J. Reiter, *Duke University*
- L.-P. Rivest, *Université Laval*
- M. Scanu, *Italian National Institute of Statistics*
- F.J. Scheuren, *National Opinion Research Center*
- M. Schonlau, *University of Waterloo*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- L. Stokes, *Southern Methodist University*
- S. Sugasawa, *University of Tokyo*
- M. Templ, *Zurich University of Applied Sciences*
- N. Thuan, *Oregon Health & Science University*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- A.-A. Vallée, *Université Laval*
- S. van Buuren, *Universiteit Utrecht*
- A. Van Delden, *Statistics Netherlands*
- J. van den Brakel, *Statistics Netherlands*
- A. Veiga, *Instituto Brasileiro de Geografia e Estatistica*
- F. Verret, *Statistics Canada*
- J. Wagner, *University of Michigan*
- Z. Wang, *Wilfrid Laurier University*
- N. Wlömert, *Vienna University of Economics and Business*
- C. Wu, *University of Waterloo*
- C. Xu, *The University of Oklahoma Health Sciences Center, Biostatistics and Epidemiology*
- D. Yang, *U.S. Bureau of Labor Statistics*
- Y. You, *Statistics Canada*
- C. Yu, *Iowa State University*
- W. Yung, *Statistics Canada*
- L.-C. Zhang, *University of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2021 issues: Céline Ethier of International Cooperation and Methodology Innovation Centre; the team from Dissemination Division, in particular: Chantal Chalifoux, Sophie Pilon, Eric Chenier, Kathy Charbonneau, Beth Chepita, Travis Robinson, Brianna Bucci and Karo-Lynn Audy as well as our partners in the Communications Division

ANNOUNCEMENTS

Nominations Sought for the 2023 Waksberg Award

The journal *Survey Methodology* has established 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 to write a paper 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.

The recipient of the Waksberg Award will receive an honorarium and give the 2023 Waksberg Invited Address at the Statistics Canada Symposium, expected to be held in the autumn of 2023. The paper will be published in an upcoming issue of *Survey Methodology* (Targeted for December 2023).

The author of the 2023 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*. **Nomination of individuals to be considered should be sent by email before February 28, 2022 to the chair of the committee, Jack Gambino (jack.gambino@gmail.com). Nominations should include a CV and a nomination letter.** Nominations will remain active for 5 years.

Members of the Waksberg Paper Selection Committee (2021-2022)

Jack Gambino, *Statistics Canada* (Chair)
Giovanna Ranalli, *University of Perugia*
Denise Silva, *Brazilian Institute of Geography and Statistics*
Kristen Olson, *University of Nebraska-Lincoln*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)
Louis-Paul Rivest (2014 - 2015)
Tommy Wright (2015 - 2016)
Kirk Wolter (2016 - 2017)
Danny Pfeffermann (2017 - 2018)
Michael A. Hidiroglou (2018 - 2019)
Robert E. Fay (2019 - 2020)
Jean Opsomer (2020 - 2021)

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. 2, June 2021

Special Issue on New Techniques and Technologies for Statistics

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. 3, September 2021

Special Issue on Population Statistics for the 21st Century

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 49, No. 1, March/mars 2021

Special Issue: Special Issue on Neuroimaging Data Analysis

Issue Information

Editorial

Research Article

Discussions

Rejoinder

Research Articles

Acknowledgement

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 49, No. 2, June/juin 2021

Issue Information

Research 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 "$\exp(\cdot)$" and "$\log(\cdot)$", 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.