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

Survey Methodology

Survey Methodology
45-2

Release date: June 27, 2019

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

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

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

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

June 2019

•

Volume 45

•

Number 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 45, Number 2, June 2019

Contents

Waksberg Invited Paper Series

Regular Papers

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 185-186
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

This issue of *Survey Methodology* opens with the 17th paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Tommy Wright (Chair), Kirk Wolter, Danny Pfeffermann and Elizabeth Stuart for having selected Donald B. Rubin as the author of 2017 Waksberg Award paper.

2017 Waksberg Invited Paper

Author: Donald B. Rubin

Donald B. Rubin is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University. He was Chair of the department during 1985-1994 and 2000-2004. He is an Elected Fellow/Member of: the American Statistical Association (1977), the Institute of Mathematical Statistics (1977), the International Statistical Institute (1984), the American Association for the Advancement of Science (1984), and the American Academy of Arts and Sciences (1993). In 2008 Professor Rubin was elected an Honorary Member of the European Association of Methodology, in 2009 he was elected a Corresponding (foreign) Fellow of the British Academy, and in 2010 to the National Academy of Sciences. He has authored/coauthored over 400 publications (including 10 books), and has made important contributions to statistical theory and methodology, particularly in causal inference, the design and analysis of experiments and sample surveys, the treatment of missing data, and Bayesian data analysis. Among his many awards and honors, Professor Rubin has received the Samuel S. Wilks Medal (American Statistical Association, 1995), the Parzen Prize for Statistical Innovation (1996), the Fisher Lectureship (2004) and the George W. Snedecor Award of the Committee of Presidents of Statistical Societies (2007). He was named Statistician of the Year, American Statistical Association, Boston Chapter (1995), and Statistician of the Year, American Statistical Association, Chicago Chapter (2001). He was Associate editor for: Journal of Educational Statistics (1976-1979), Theory and Methods, The Journal of American Statistical Association (1975-1979), Coordinating Editor and Applications Editor, The Journal of American Statistical Association (1980-1982), Biometrika (1992-1995), Survey Methodology (1988-1993), Statistica Sinica (1993-2004). Professor Rubin has been, for many years, one of the most highly cited authors in mathematics in the world (according to ISI Science Watch); in 2002 he was ranked Seventh in the World for the Decade 1991-2000; he has 10 singly authored publications with over one thousand citations each. His biography is included in many places including, Who' Who in The World.

Waksberg Award honorees and their invited papers

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

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

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

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

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

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

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

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

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

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

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.

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

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

2014 Constance F. **Citro**, "From Multiple Modes for Surveys to Multiple Data Sources for Estimates". *Survey Methodology*, vol. 40, 2, 137-161.

2015 Robert M. **Groves**, "Towards a Quality Framework for Blends of Designed and Organic Data". Proceedings: Symposium 2016, Growth in Statistical Information: Challenges and Benefits.

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

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

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.

2019 Chris **Skinner**, Manuscript topic under consideration.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 187-198
Statistics Canada, Catalogue No. 12-001-X

Conditional calibration and the sage statistician

Donald B. Rubin[1]

Abstract

Being a calibrated statistician means using procedures that in long-run practice basically follow the guidelines of Neyman's approach to frequentist inference, which dominates current statistical thinking. Being a sage (i.e., wise) statistician when confronted with a particular data set means employing some Bayesian and Fiducial modes of thinking to moderate simple Neymanian calibration, even if not doing so formally. This article explicates this marriage of ideas using the concept of conditional calibration, which takes advantage of more recent simulation-based ideas arising in Approximate Bayesian Computation.

Key Words: Approximate Bayesian Computation (ABC); Bayesian inference; Fiducial inference; Fisher; Frequentist methods; Neyman.

1 Principled statisticians

There are many possible definitions for what makes a principled statistician, where by "principled" I do not necessarily imply "good" or "sage", but simply following clear principles of behavior. I think generally there are three major themes or philosophies of statistical inference. Neymanian frequentists, following ideas proposed originally by Neyman (1923, 1934), care about the operating characteristics of procedures (e.g., point estimates, interval estimates), under repeated sampling: point estimates should be approximately unbiased for their estimands (averaging over all possible samples), interval estimates should be conservative in the sense of having at least their nominal coverage of their estimands (again averaging over samples), and tests should be conservative in the sense of rejecting true null hypotheses at most at their nominal rates. These desiderata are widely viewed as being features of valid statistical inference (e.g., see Lehmann, 1959). Of course, all procedures that are valid are not equally desirable; valid point estimates with less variability are better, valid interval estimates that are shorter are better, and so forth.

Bayesian statisticians (e.g., Savage, 1954; Lindley, 1971; de Finetti, 1972), in contrast to repeated-sampling operating characteristics, care about correct conditioning on observed data under a particular probabilistic specification. Fisherian statisticians (in the sense of Fiducialists, at least as I view the most central idea of this approach, Fisher (1956)) avoid conclusions that appear to be contradicted by observed data, which is at the heart of Fisher's randomization test in experiments; I have long resonated to the wisdom of this approach and its generalizations, as expressed in Rubin (1984). Nevertheless, I also think Bayesian thinking is critical to being a wise applied statistician in practice, for example by using posterior predictive p-values and checks, which assess whether a proposed model can (that is, is able to, not must) generate data that look like the observed data set we are facing – we return to this central idea later.

There is little doubt that frequentist thinking dominates current statistical thinking even though Bayesian procedures are becoming more common largely because of current computational advances, which allow

1. Donald B. Rubin, Professor, Yau Mathematical Sciences Center, Tsinghua University; Murray Shusterman Senior Research Fellow, Fox School of Business; Professor Emeritus, Harvard University. E-mail: dbrubin@me.com.

many complicated models to be fit routinely. Nevertheless, I believe that Bayesian and fiducial thinking are fundamental to being a sage (i.e., wise, not necessarily principled in the narrow sense of the following specific principles) statistician.

2 Should frequentists care about Bayesian procedures?

For example, why should frequentists ever use the sample mean to estimate the population mean? After all, the sample mean is essentially the center of the Bayesian posterior distribution of the population mean under a Gaussian model with relatively diffuse prior distributions on parameters, and therefore derived using an "unreliable (i.e., Bayesian) methodology". Of course this sentence is facetious, and not intended to be taken seriously, although there are serious points underlying it.

Serious Point #1: The original motivation for any statistical procedure, whether Bayesian or Fiducial or the result of some amazing dream, is irrelevant to the frequentist operating characteristics of that procedure. I used to hear this criticism directed at Multiple Imputation (MI), Rubin (1978). Because MI's initial justification was Bayesian, MI could never be trusted from a design-based (frequentist) perspective.

Serious Point #2: For creating procedures, especially in complex situations, such as those that easily arise with unintended missing data, Bayesian methods are far more generative of sensible answers than standard, frequentist arguments, such as those based on "principles" such as unbiasedness or minimizing mean squared error. Again, I think that the relative success of MI for missing data illustrates this point nicely (e.g., as argued in many places, including Rubin (1996)).

Serious Point #3: Nonetheless, frequentist evaluations (e.g., of bias of point estimates and coverage of interval estimates) are still highly relevant to the sage statistician because all idealizations, including Bayesian ones, are oversimplifications. As George Box said, "All models are wrong, but some are useful" (Box, 1976); also earlier, John von Neumann (1947) stated, "Truth is much too complicated to allow anything but approximations".

Two more Serious Points, an analogy, and some summarized points.

Serious Point #4: Frequentist criteria based on operating characteristics can be used to evaluate any procedure (really the same as Serious Point #1).

Serious Point #5: Therefore, we can, and moreover should, use Bayesian models to create procedures that appear to be appropriate under plausible assumptions, and use frequentist methods to evaluate these procedures in realistic situations, situations more general than those that were assumed when deriving the Bayesian answers.

Versions of these points have been made before, for example in Box (1980), Rubin (1984), and in Little (2008) and its discussions (e.g., Rubin, 2008), as well as earlier and later by various other authors. Many practicing statisticians would pretty much agree with all Serious Points, except perhaps Serious Points #2 and #5. Being a "calibrated" statistician generally means choosing procedures that have good operating characteristics over a broad range of circumstances. Being sage when confronted with a particular data set is more difficult to define, because it depends on the immediate context of the problem being confronted, and the consequences of resulting decisions, which formally can lead to decision theory (Wald, 1950). My own view is that although this framework is theoretically appealing, real decisions are made in contexts with many fuzzy and perpetually changing considerations, which disable the utility of the full formal structure of decision theory.

3 On the elusive goal of being calibrated and sage

Bayesians condition on what is observed, and so in principle, try to be appropriate to the data at hand. True Bayesian calibration, however, in the sense of creating interval estimates that have accurate Bayesian coverage of the true posterior distribution no matter what "Truth" generated the observed data, is essentially impossible in practice. This was illustrated to me in fairly trivial examples, first in Rubin (1983) when I was attempting to demonstrate the superiority of the Bayesian approach in the context of survey inferences, then in Rosenbaum and Rubin (1984), which documented the relevance of stopping rules on the Bayesian validity of Bayesian inferences, unless all model and prior distributions were correct, and more recently in Ferriss (Harvard PhD. Thesis, 2018), which considered the implications of re-randomization in experiments on the Bayesian validity of Bayesian inference. But despite this inability to approach the Bayesian ideal when there is the absence of knowledge of correct models, a statistician can still seek to be calibrated, in some important sense, and sage in the fiducial sense of avoiding conclusions that are contradicted by the data set actually being analyzed. I refer to this as being "conditionally calibrated" and explicate this surprisingly elusive idea here.

A personal aside relevant to this idea of being conditionally calibrated: When I was visiting the University of California, Berkeley in the 1970's and had a visitor's office next to, the then retired, but still intellectually vibrant and feisty, Jerzy Neyman, he clearly expressed to me his view that such conditioning for statistical inference was essentially impossible to define correctly, at least in the context of our 1970's discussion of Fisher's desiderata to condition on ancillary statistics when drawing inferences.

Another relevant aside: my reading is that fundamentally, both Neyman and Fisher wanted, at least in their youths, to be effectively Bayesian in that they both sought a distribution for the estimand conditional on the observed data, but took very different mathematical approaches to finding that distribution, as discussed in (Rubin, 2016). Fisher (1956) was totally forthright about this fiducial objective: "The Fiducial argument uses the observations to change the logical status of the parameter [the unknown estimand] from one in which nothing is known of it, and no probability statement can be made, to the status of a random

variable having a well-defined distribution". Values of the estimand with little support in this fiducial distribution, were those values that were stochastically contradicted by the observed data, that is, if true, they were unlikely to generate the observed data – a stochastic proof by contradiction. Despite the intuitive appeal of this approach, mathematical foundations for it have not enjoyed universal acceptance (e.g., Dempster, 1967; Martin and Liu, 2016).

Neyman was not direct as Fisher when seeking a distribution for the estimand, but consider his original definition of "confidence intervals" (Neyman, 1934), which was openly based on some Bayesian logic:

> Suppose we are taking samples, Σ, from some population π. We are interested in a certain collective character of this population, say θ. Denote by x a collective character of the sample Σ and suppose that we have been able to deduce its frequency distribution, say $p(x|\theta)$, in repeated samples and that this is dependent on the unknown collective character, θ, of the population π....
>
> Denote now by $\varphi(\theta)$ the unknown probability distribution *a priori* of θ...
>
> ...[T]he probability of our being wrong is less than or at most equal to $1-\varepsilon$, and this whatever the probability law a priori, $\varphi(\theta)$.
>
> The value of ε, chosen in a quite arbitrary manner, I propose to call the "confidence coefficient." If we choose, for instance, $\varepsilon = \cdot 99$ and find for every possible x the intervals $[\theta_1(x), \theta_2(x)]$ having the properties defined, we could roughly describe the position by saying that we have 99 per cent. confidence in the fact that θ is contained between $\theta_1(x)$ and $\theta_2(x)$....
>
> ...[I]call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals, corresponding to the confidence coefficient ε.

Figure 3.1 Neyman's (1934, pages 589-590) definition of confidence intervals.

4 Calibration – A simulation perspective

Consider how to evaluate a proposed procedure, generically called P, which is to be applied to a data set, generically called Y, yet to be collected from a population; P is a specified function of the data Y and will be used to estimate the estimand, here a scalar quantity, Q, which describes some aspect of the population from which Y is drawn. For descriptive simplicity, suppose P is a purported 95% interval for Q; for P to be exactly calibrated means that P includes Q in exactly 95% of repeated samples. Further, suppose Y is drawn from its population using design D, which is known and fixed throughout this discussion; for concreteness, D is simple random sampling. Although D is known, at the design stage, the data set Y is not yet known. Also suppose, again for simplicity, that all "experts" interested in this problem agree on a set of K possible "Truths" for describing the unknown population to which D will be applied to obtain data set Y; call these possible Truths $T_1, T_2, \ldots, T_K$, and the values of their associated local (local to each Truth) estimands $Q_1, Q_2, \ldots, Q_K$, where $Q_k = \tilde{Q}(T_k)$, for $k = 1, \ldots, K$, for the function $\tilde{Q}$, common to all possible truths. The estimand Q is the value of the function $\tilde{Q}$ evaluated at the actual Truth.

The Q_k are here called local estimands, that is, local to the truths. As far as I can tell, Neyman never fomally considered such local estimands, but I see them as important bridges to the Bayesian perspective as well as to being a sage statistician. Only one of the possible truths is the actual truth. The inferential objective is the value of $\tilde{Q}$ for the Truth that generated the yet-to-be observed data Y.

The collection of K possible Truths can often be compactly described mathematically, so that K can be essentially infinite. One example of such Truths, and their associated local estimands, could be K Gaussian univariate populations, with unknown local means, μ_k, and with the scalar estimand Q equal to the mean of the one true population. Or the Truths could be all possible N-dimensional vectors of real numbers; this is the standard finite population set-up for survey sampling with N units and one scalar variable, as in Cochran (1963) and Kish (1965), where the estimand Q is typically the mean of the N values for the true population.

We continue by defining simple calibration using a simulation to fix ideas; this simulation will be used to define concepts throughout this manuscript, including the key concept of conditional calibration. Suppose that for each possible truth, T_k, $k = 1, \ldots, K$, with local estimand Q_k, we have drawn J data sets, labeled Y_{jk}, $j = 1, \ldots, J$, each drawn using design D. To each of these data sets, we apply procedure P to the data to create an interval estimate for Q_k, where for each k, Q_k is the same for all Y_{jk} $(j = 1, \ldots, J)$ because all such Y_{jk} arose from the same truth T_k. We then assess whether when P is applied to Y_{jk}, the resulting interval includes the local estimand Q_k. The proportion of data sets, $\{Y_{jk},\ j = 1, \ldots, J\}$, for which the interval P includes Q_k is called here the local calibration (or local coverage) of the procedure P for the k^{th} Truth, notationally written C_k for $k = 1, \ldots, K$. For evaluating point estimators, rather than interval estimators, the calibration of P for Q_k could be replaced by the bias or mean squared error of the point estimate of Q_k.

This simulation is depicted in Table 4.1, where each column represents a possible truth, and the J rows represent the J data sets generated under each truth.

Table 4.1
Display of simulation (Each column represents a possible truth)

Local estimands:	$Q_1 = \tilde{Q}(T_1)$	…	$Q_k = \tilde{Q}(T_k)$	…	$Q_K = \tilde{Q}(T_K)$
	Y_{11}	…	Y_{1k}	…	Y_{1K}
	⋮		⋮		⋮
	Y_{j1}	…	Y_{jk}	…	Y_{jK}
	⋮		⋮		⋮
	Y_{J1}	…	Y_{Jk}	…	Y_{JK}
Calibration of P for Q_k:	C_1	…	C_k	…	C_K

Now we define local calibration using 95% to represent any level of coverage. A 95% interval estimate of Q, P, is called "locally (for truth T_k) conservatively calibrated" if $C_k >= 95\%$; we could say that P is "approximately locally calibrated" (for Truth T_k) if C_k is close to 95%, but this idea was never formally defined by Neyman, although in Fisher's (1934) discussion of Neyman (1934), we can see Fisher had something like this in mind with his criticism of Neyman's formulation.

Next, following Neyman, the interval estimate P is called "confidence calibrated" across the ensemble of possible truths, $\{T_k, k = 1, \ldots, K\}$, if all $C_k >= 95\%$, or returning to Neyman's original definition, P is then simply called a 95% confidence interval for Q. The critical point here for calibration is that all that matters to a die-hard Neymanian frequentist, when evaluating a procedure, P, for its validity, is whether the collection of C_k values for procedure P are all greater than the nominal level for P. The word "confidence" arises because when confronted with the results of Table 4.1 for procedure P and with a critic who selected one Truth from the collection of possible truths, you should be "confident" that the result of applying P to Y^* will be an interval that includes Q.

These assessments of 95% confidence calibration are well-defined no matter what the etiology of the procedure P. BUT, are they statistically apposite for evaluating P as a 95% interval estimate of the unknown Q after seeing a specific data set, call it Y^*? That is, after seeing a specific instance of Y, *now known to be* Y^*, does the 95%-attached to P necessarily reflect the judgment of a sage statistician? Maybe we should seek only procedures that are approximately calibrated for truths that plausibly could have generated the observed Y^*?

We now consider the formal Bayesian perspective because it sheds light on this concept of being sage after seeing $Y = Y^*$.

5 The Bayesian posterior distribution of *Q*

The Bayesian approach differs from the Neymanian approach (and from Fisher's fiducial approach) by formulating the problem so that a real conditional probability distribution for the estimand Q can be calculated, using the laws of probability theory to condition on the fact that the observed data equals Y^*- this distribution is called the posterior distribution of Q, that is, posterior after seeing $Y = Y^*$. To conduct this activity formally, Q must be a random variable, and thus Q needs to have a "starting" probability distribution, called its prior distribution, meaning prior to seeing any data; in the context of our setup, this prior distribution is a distribution over the possible local estimands, that is, a set of K probabilities (summing to one), one probability for each possible Truth. This prior distribution is essentially a set of K weights $\{W_k, k = 1, \ldots, K\}$ reflecting the prior beliefs of experts that each of the K possible local estimands is the correct one. The Neymanian frequentist has no use for such weights over the set of possible Truths, because the 95% is supposed to hold for any set of weights, and thus for each possible Truth (i.e., for all K point mass prior distributions).

Now comes the part of the argument that hints at a departure from Neyman's 1970's claim to me that conditional inference is too difficult. In the context of the simulation just described, and admitting some Bayesian or fiducial logic, when confronted with actual observed data set Y^*, attention should be focused on the parts of the simulation where the generated Y_{jk} equals Y^*; the other Y_{jk} can be ignored (at least in the context of the idealized description here, where J is essentially infinite) because, to be fully Bayesian, we want to condition on Y equaling Y^*.

In fact, let us use the simulation itself to describe the Bayesian posterior distribution of Q, i.e., the distribution of Q conditioning on the fact that $Y = Y^*$. Let M_k^* be the proportion of the J values of Y_{jk} that match Y^*, for $k = 1, \ldots, K$; that is, for truth T_k, M_k^* is the proportion of the generated data sets from truth T_k that match the actual data set Y^*. For example, if M_k^* is zero, then the a priori possible truth T_k could not be the actual truth because it could not have generated observed data Y^*. The posterior probability that the estimand Q equals Q_k, the local value of $\tilde{Q}$ for Truth T_k, is the weighted average of the proportions, M_k^*, weighted by W_k, the prior probability that T_k is the correct truth. Here, this weighted average of proportions is generally labeled π_k, where π_k for the observed data Y^* is labelled π_k^* and equals $M_k^* W_k \Big/ \sum_{k'=1}^{K} [M_{k'}^* W_{k'}]$; we could call π_k^* the estimated ability of Truth T_k to match observed data Y^*. This description of the posterior distribution of Q using simulation is from Rubin (1984); see Figure 5.1.

> Suppose we first draw equally likely values of θ from $p(\theta)$, and label these $\theta_1, \ldots, \theta_s$. The $\theta_j, j = 1, \ldots, s$ can be thought of as representing the possible populations that might have generated the observed X. For each θ_j, we now draw an X from $f(X|\theta=\theta_j)$; label these $X_1, \ldots, X_s$. The X_j represent possible values of X_j that might have been observed under the full model $f(X|\theta)\, p(\theta)$. Now some of the X will look just like the observed X and many will not; of course, subject to the degree of rounding and the number of possible values of X, s might have to be very large in order to find generated X_j that agree with observed X, but this creates no problem for our conceptual experiment. Suppose we collect together all X_j that match the observed X, and then all θ_j that correspond to these X_j. This collection of θ_j represents the values of θ that could have generated the observed X; formally, this collection of θ values represents the posterior distribution of θ. An interval that includes 95% of these values of θ is a 95% probability interval for θ and has the frequency interpretation that under the model, 95% of populations that could have generated the data are included within the 95% interval.

Figure 5.1 Description of posterior distribution from Rubin (1984).

There are objections to this approach. First, where do the prior weights W_k come from and who are the experts providing these weights? Perhaps we should find some way to avoid using these potentially overly subjective prior weights? Second, perhaps the requirement for exact equality between a generated data set Y_{jk} and the observed data set Y^* should be relaxed in some way so that a generated Y_{jk} does not have to equal Y^* exactly but only "look like" it came from the same distribution as did Y^*, and so match Y^* in some way?

More on this second point first, which is clearly important when trying to conduct an actual simulation like this idealized one with a finite budget. The approximate equality between generated data Y and observed data Y^* can be achieved in situations with low-dimensional sufficient statistics, because only those statistics have to match. But this idea of generated data being "close to" observed data Y^* is the basis of all work using this description of the posterior distribution to conduct "ABC" – Approximate Bayesian Computation, apparently first described in the paragraph in Figure 5.1 (https://en.wikipedia.org/wiki/Approximate_Bayesian_computation, Tavare, Balding, Griffiths and Donnelly, 1997). We simply assume at this point that we have chosen some such metric to define the function M_k, and use it to define the ability of Truth T_k to generate data sets that match the observed data, Y^*.

6 The conditionally calibrated statistician's evaluation of procedure *P*

With the basic definitions of Sections 4 and 5 for C_k (the local calibration of P for Q_k under T_k) and the M_k^* (the matching ability of T_k for Y^*) now established, we are prepared to consider the concept of Conditional Calibration (CC) and to make the connection to being a sage statistician after having observed Y^*.

Conditional Calibration starts at the same place as Neymanian unconditional calibration, but is sensitive to the Bayesian and Fiducial arguments by discounting results from possible truths whose drawn data sets Y_{jk} are not close to Y^* as assessed by their match rates M_k^*. The point of doing this is: Why should we care about calibration for a priori possible Truths that could not have generated the observed Y^*, i.e., truths that are a posteriori implausible? But this CC perspective does not go to the full Bayesian extreme, which, first, ignores all aspects of the simulation except the Truths that generated data sets exactly matching the observed data set, and second, explicitly uses the often weakly justified prior distribution, W_k, to weight the local matching rates.

Hence, we are left summarizing our simulation, which evaluates a procedure P for inference about the estimand Q from observed data Y^*, using a two dimensional criterion: The local calibration of P for Q_k under truth T_k, that is C_k; and the local match rate for Truth T_k, the proportion of generated Y_{jk} under truth T_k, that are accepted as matching Y^*, M_k^* = the fraction of Y_{jk} that are considered equal to Y^*. Those possible truths with M_k^* near one are clearly more relevant to the situation with observed data Y^* than those truths with values of M_k^* near zero.

Thus procedure P applied to each possible truth with observed data Y^* can be displayed as K points in a two-dimensional graph where the horizontal axis is the average match to Y^* of the data sets generated by truth T_k, M_k^*, and the vertical axis is the local calibration of P for the data sets generated by Truth T_k, C_k. We call this the "conditional calibration plot" and it is illustrated and discussed in the Section 7. Of major relevance to drawing sage inferences for possible truths from observed data Y^*, many different

procedures can be displayed on the same conditional calibration plot for a fixed data set Y^* and a fixed set of possible truths.

7 The conditional calibration plot and its use for sage selecting statistician with observed data Y^*

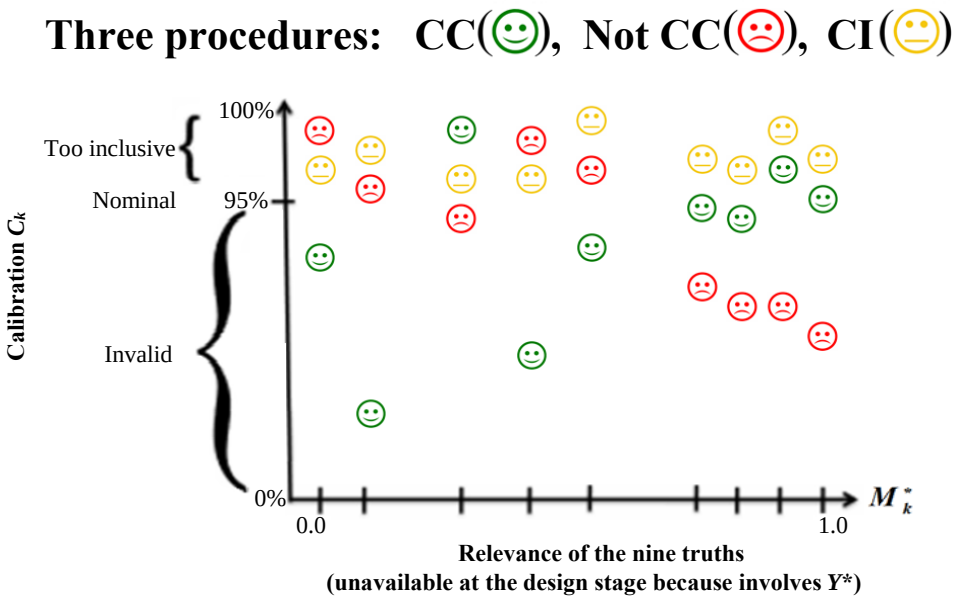


Figure 7.1 C_k versus M_k^* Plots for a fixed data set, with $K = 9$ Truths (columns).

The conditionally calibrated (CC) statistician faced with estimating Q using procedure P from data set Y^* cares about being approximately calibrated, i.e., close C_k to 95% especially for Truths with large values of M_k^*, indicating that such Truths could have plausibly generated Y^*. In other words, when comparing procedures for estimating Q from Y^*, the sage statistician, in addition to conservative unconditional calibration (i.e., confidence coverage), especially cares about accurate calibration for Truths that are plausible, and therefore implicitly ignores the calibration of procedures for Truths that are implausible given Y^*.

Figure 7.1 presents hypothetical simulation results with a fixed data set Y^* and a fixed set of nine possible Truths (with nine associated local match rates to Y^*) for three procedures, indicated by faces. The vertical axis is not linear in C_k but expanded for values of C_k closer to unity, which is where our interest is focused. One procedure is labeled "Smile" because it is approximately calibrated (C_k close to 95%) for possible Truths that could have generated Y^* (M_k^* close to 1), even though poorly calibrated (C_k well below 95%) for a priori possible Truths that are implausible given the observed Y^* (M_k^* much lower than 1). A second procedure is labeled "Frown" because it is not CC, being invalid (meaning its local calibration is substantially less than 95%), including for truths that are plausible given Y^*. The third procedure is

labeled as “Neutral [CI]” because, although it is a valid confidence interval in Neyman’s sense of having its minimum local calibration at least 95%, it is not approximately calibrated for Truths that are plausible given the observed data set, Y^*. This procedure could, for me, be described by a mild frown, but maybe not for Neyman, based on our 1970’s conversation.

That is, to repeat, Neymanian (conservative = confidence) calibration for each procedure formally just cares about the procedures’ minimum C_k across the entire ensemble of a priori possible truths. Also, the rigid Bayesian just cares about the weighted average of the M_k^* across the possible truths, weighted by the prior possibly unreliable distribution for the truths, W_k. The sage CC statistician cares about approximate local calibration of procedures for those Truths that are plausible; if a confidence-valid 95% procedure P displays C_k values substantially bigger than 95% for plausible Truths, this suggests that there exist better CC procedures for this situation with data set Y^*; that is, calibrated procedures that are more efficient and so result in shorter intervals. Notice for example, that the confidence-valid procedure in Figure 7.1 (Neutral face) has worse CC than Smile, and thus although a plausible competitor to Smile at the design stage should be seen as inferior to Smile after seeing data Y^* because it is too conservative for some of the relevant Truths.

8 Implementing this idea in practice

To implement this idea in practice would require work, certainly more intellectual effort than is currently expended in many statistical investigations. The implementation would begin at the same place as is standard in current carefully constructed studies. We would begin by considering a set of procedures, each of which is usually conservatively calibrated in the traditional sense, for the problem at hand. Then we would collect opinions from experts about the generally plausible Truths in the specific situation we are facing; this step is executed in some current problems, although typically informally.

If possible, then we should gather some information for W_k, the prior weights on the possible truths; these could be useful for later consideration of the construction of the matching averages M_k (no asterisk yet, because the data, Y^*, are not yet observed). We should obtain agreement on how to define M_k and whether to use the prior weights W_k. This is the ABC task. Finally, agreement is needed on how to use the CC plot to compare the various procedures being considered.

All of this effort should be conducted before the actual data set Y^* is observed. For this reason, alone, the implementation of this idea is more intellectually demanding than standard practice, but it is a component of being a sage statistician.

Acknowledgements

This is the written version of DB Rubin’s Waksberg Award address delivered 8 November 2018 in Ottawa, Canada. The author was a friend of Joseph Waksberg and was always impressed with his wise application of statistics, and hopes that this contribution continues that tradition. Other versions of this talk

have been given over the past five years, most recently as the SN Roy Invited Lecture in Kolkata India 27 Decmber 2018. The author acknowledges very helpful comments from Roderick Little, Tommy Wright, as well as from Wesley Yung and other members of the *Survey Methodology* editorial board, and recently from Hal Stern and Yannis Yatracos.

References

Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4), 383-430.

Cochran, W.G. (1963). *Sampling Techniques*, *2nd Edition*. New York: John Wiley & Sons, Inc.

De Finetti, B. (1972). *Probability, Induction, and Statistics*. New York: John Wiley & Sons, Inc.

Dempster, A.P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2), 325-339.

Ferris, T. (2018). *Topics in Casual Inference and the Law*. Senior Data Scientist, Google.

Fisher, R.A. (1934). Contribution to a discussion of J. Neyman's paper on the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97, 614-619.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

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

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley & Sons, Inc.

Lindley, D.V. (1971). *Bayesian Statistics: A Review*, SIAM.

Little, R.J. (2008). Weighting and prediction in sample surveys. *Calcutta Statistical Association Bulletin*, 60, 3-4, 147-167.

Martin, R., and Liu, C. (2016). *Inferential Models*. New York: Chapman and Hall/CRC.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Translated into English in *Statistical Science*, 1990, 5(4), 463-472.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

Rosenbaum, P.R., and Rubin, D.B. (1984). Sensitivity of Bayes inference with data dependent stopping rules. *The American Statistician*, 38, 106-109.

Rubin, D.B. (1978). Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

Rubin, D.B. (1983). A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. *Scientific Inference, Data Analysis and Robustness*. New York: Academic Press, Inc., 213-244.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 434, 473-489.

Rubin, D.B. (2008). Discussion of "Weighting and prediction in sample surveys" by R.J. Little. *Calcutta Statistical Association Bulletin*, 60, 185-190.

Rubin, D.B. (2016). Fisher, Neyman, and Bayes at FDA. *Journal of Biopharmaceutical Sciences*, 26, 1020-1024.

Savage, L.J. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Tavare, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2), 505-518.

Von Neumann, J. (1947). The mathematician. In *Works of the Mind*, (Ed., R.B. Haywood), University of Chicago Press, 180-196.

Wald, A. (1950). *Statistical Decision Functions*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 199-216
Statistics Canada, Catalogue No. 12-001-X

A bivariate hierarchical Bayesian model for estimating cropland cash rental rates at the county level

Andreea Erciulescu, Emily Berg, Will Cecere and Malay Ghosh[1]

Abstract

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) is responsible for estimating average cash rental rates at the county level. A cash rental rate refers to the market value of land rented on a per acre basis for cash only. Estimates of cash rental rates are useful to farmers, economists, and policy makers. NASS collects data on cash rental rates using a Cash Rent Survey. Because realized sample sizes at the county level are often too small to support reliable direct estimators, predictors based on mixed models are investigated. We specify a bivariate model to obtain predictors of 2010 cash rental rates for non-irrigated cropland using data from the 2009 Cash Rent Survey and auxiliary variables from external sources such as the 2007 Census of Agriculture. We use Bayesian methods for inference and present results for Iowa, Kansas, and Texas. Incorporating the 2009 survey data through a bivariate model leads to predictors with smaller mean squared errors than predictors based on a univariate model.

Key Words: Hierarchical Bayes; Bivariate mixed model; Benchmarking.

1 Introduction

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) conducts hundreds of surveys each year to obtain estimates related to diverse aspects of US agriculture. Examples of parameters that NASS estimates include total production, harvested area, and crop yield. Estimation for sub-state domains, such as counties, is difficult due to small sample sizes. Our interest is in estimation of the county-level cash rental rate, the market value of land rented on a per acre basis for cash only.

Estimates of county-level cash rental rates serve many purposes. Farmers use the estimates for guidance in determining rental agreements (Dhuyvetter and Kastens, 2009). Agronomists use the estimates to study research questions related to the interplay between cash rental rates and other economic characteristics such as commodity prices and fuel costs (Woodard, Paulson, Baylis and Woddard, 2010). NASS's published estimates of mean cash rental rates at the county level have implications for the Conservation Reserve Program, a policy that encourages agricultural landowners to conserve their land. The 2008 and 2014 Farm Bills require NASS to collect data on cash rental rates for three land use categories – non-irrigated cropland, irrigated cropland, and permanent pasture – for counties with at least 20,000 acres of cropland or pastureland.

To satisfy the requirements of the 2008 and 2014 Farm Bills, NASS conducts a Cash Rent Survey. A concern is that direct estimators of county means from the Cash Rent Surveys may be unstable due to small

1. Andreea Erciulescu, National Institute of Statistical Sciences, Washington D.C., U.S.A.; Emily Berg, Department of Statistics, Iowa State University, Ames, IA, U.S.A. E-mail: emilyb@iastate.edu; Will Cecere, Westat, Rockville, MD, U.S.A.; Malay Ghosh, Department of Statistics, University of Florida, Gainesville, FL, U.S.A.

realized sample sizes. We investigate the use of mixed models (Rao and Molina, 2015) to stabilize the estimators of average cash rental rates at the county level. NASS publishes estimates of average cash rental rates at the state level before county level estimation from the Cash Rent Survey is complete. To maintain internal consistency, the county predictors must satisfy a benchmarking restriction.

In a frequentist framework, Berg, Cecere and Ghosh (2014) use area-level models to predict county-level cash rental rates for all states and for the three land use categories of non-irrigated cropland, irrigated cropland, and permanent pasture. For each combination of land use category and state, the method of Berg et al. (2014) uses data from two years. An assumption that the variances for the two years are the same motivates the Pitman-Morgan transformation, which converts the vector of observations for the two time points into an average and a difference. After separate univariate models are applied to the average and the difference, the predictor for each time point is obtained by adding the predictor of the average to half of the predictor of the difference. The method of Berg et al. (2014) is demonstrated to provide a practical approach to obtaining reasonable predictions across a diverse range of conditions. Nonetheless, the effects of simplifying assumptions warrant additional investigation. If the variances for the two time-points differ, then, as discussed in Berg et al. (2014), the mean squared error (MSE) estimator based on the Pitman-Morgan transformtion can have a negative bias. Further, the Berg et al. (2014) method does not account for the effect of benchmarking when estimating the MSE.

This study addresses the issues of non-constant variances across time and the effect of benchmarking on efficiency in the context of the NASS Cash Rent Surveys through the use of a bivariate hierarchical Bayesian (HB) model for the unit-level data. The model is sufficiently flexible to allow the variances to differ between the two time-points. The use of Bayesian methods for inference facilitates estimation of the increase in posterior MSE due to benchmarking. Another innovation of the bivariate HB approach is that it incorporates the survey weights in the variance model. We also aim to improve the efficiency of the predictors for particular situations, relative to Berg et al. (2014), by allowing the covariates to differ across states. Datta, Day and Maiti (1998) examine HB bivariate models for the county crop acreage data of Battese, Harter and Fuller (1988). Our model extends the Datta et al. (1998) model to account for a relationship between the weight and the variance as well as an unbalanced data structure.

We focus on prediction of county level cash rental rates for non-irrigated cropland using the responses to the 2009 and 2010 Cash Rent Surveys as well as external sources of auxiliary information. In Section 2, we discuss the survey data and the auxiliary information in detail. We describe the bivariate HB model in Section 3. In Section 4, we summarize results for non-irrigated cropland in Iowa, Kansas, and Texas. In Section 5, we summarize and discuss possible future research applicable to both estimation of cropland cash rental rates and small area estimation more generally.

2 Data for modeling non-irrigated cropland cash rental rates

2.1 NASS Cash Rent Survey

NASS implemented a Cash Rent Survey in response to the 2008 Farm Bill. The specific objective of the Cash Rent Survey is to obtain county level estimates of average cash rental rates in three land use categories: non-irrigated cropland, irrigated cropland, and permanent pasture. The data for our study are from the 2009 and 2010 Cash Rent Surveys.

2.1.1 NASS Cash Rent Survey sample design

The 2009 and 2010 Cash Rent Surveys used a stratified sample design. To define the stratification, nine groups were formed on the basis of the dollars rented that an operation reported on previous surveys and censuses. The strata are the intersections of the nine groups and agricultural statistics districts. An agricultural statistics district is a group of contiguous counties within a state that are thought to have similar agricultural characteristics. The sampling fractions within strata are defined so that operations with higher dollars rented on previous surveys and censuses have greater probabilities of selection. The same sample was used for the 2009 and 2010 Cash Rent Surveys, which had a national sample size of approximately 224,000 operations. A unit may respond in only one year either because of nonresponse or because the operation only participated in a rental agreement in one of the two years.

2.1.2 Relationships between 2009 and 2010 non-irrigated cropland cash rents

A direct survey estimator for a particular land use category is a ratio of a weighted sum of the dollars rented to a weighted sum of acres rented. The weight associated with a respondent is the population size of the stratum containing the respondent divided by the number of responding units in that stratum. Berg et al. (2014) explore relationships between direct estimates for two years. For the states considered in Berg et al. (2014), the correlations between the direct estimates for the two years range from 0.20 to 0.99, where the correlation is across counties for a particular state. Because our emphasis is on unit level models, we focus on relationships over time at the unit level.

To measure the correlation between the reported 2009 and 2010 cash rental rates at the unit (farm operator) level, we compute differences between unit-level cash rental rates for non-irrigated cropland and the sample mean for a county. Only individuals that report a cash rental rate for non-irrigated cropland in both years are used to compute the differences. The difference for year t is $y_{ijt} - \bar{y}_{i.t}$, where y_{ijt} is the cash rent per acre for non-irrigated cropland reported by operator j in county i and year t, and $\bar{y}_{i.t}$ is the sample average of the y_{ijt} in county i that reported a non-irrigated cropland cash rental rate in both 2009 and 2010. The deviations between individual cash rental rates and the county means for Kansas are plotted in Figure 2.1. The deviations for 2009 and 2010 for Kansas are linearly related, and the correlation between the deviations for 2009 and the deviations for 2010 is 0.7. The extreme values in Figure 2.1 reflect the high variability among the non-irrigated cropland cash rental rates within a county in Kansas.

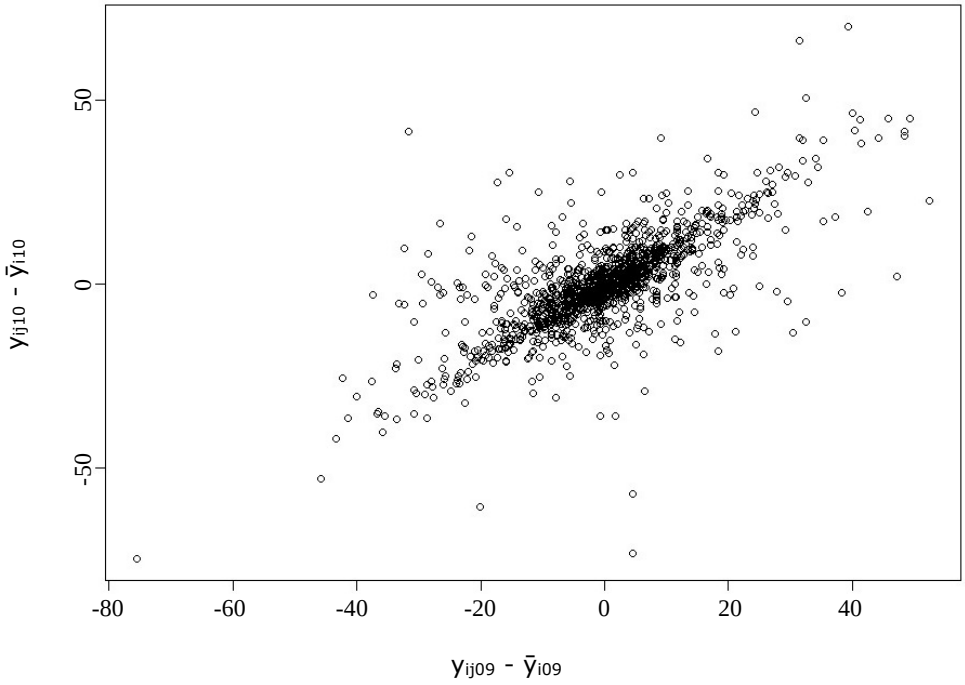


Figure 2.1 Deviations of unit-level cash rental rates from county means for 2009 (x-axis) and 2010 (y-axis) for units reporting non-irrigated cash rental rates in both years.

2.2 Auxiliary information

In an effort to improve the precision of the estimators of average cash rental rates at the county level, auxiliary variables were desired that would explain both the variability among the county means as well as the variability among units within a county. Auxiliary information for modeling cash rental rates is available from several sources external to the Cash Rent Survey. The potential covariates divide into three broad categories, depending on whether the covariate relates principally to land quality, the commodity value sold, or other farm characteristics. The list below summarizes the three categories of covariates, indicates whether each covariate is recorded at the county level or the unit level, and specifies if the covariate is only available for a particular state. Unit-level covariates are only available for units in the Cash Rent Survey sample, while area level covariates are treated as population means.

1. Land quality
    - Four National Commodity Crop Productivity Indexes (NCCPIs) are county-level covariates available for all states. Three climate-specific indexes called NCCPI-corn, NCCPI-wheat, and NCCPI-cotton reflect the quality of the soil for growing non-irrigated crops in three different climate conditions (Dobos, Sinclair and Robotham, 2012). The fourth index, Max-NCCPI, is the maximum of the three climate-specific indexes. The indexes are originally constructed at the level of a “mapunit,” an area that has relatively homogeneous soil

properties. The county-level covariates are averages of the indexes across all mapunits in a county.

- An average corn yield across years 2005-2009 is available at the county level for Iowa only. All counties in Iowa have a corn yield estimate available for at least one of the years between 2005 and 2009, and years for which a yield estimate is missing for a county are excluded from the average for that county.
- Because Kansas is more agriculturally diverse than Iowa, no single crop yield is published in at least one year between 2005 and 2009 for all counties of interest. To obtain a covariate that is measured for all counties, we constructed a non-irrigated yield index for Kansas. We first averaged NASS published yields for corn, wheat, and sorghum using the method described for the Iowa corn yields. The average yields were then standardized to have mean zero and variance one. The non-irrigated yield index for a county is defined as the largest of the three standardized yields. (For Texas, availability of crop yield information was too sparse to use to define a covariate).

2. Value of the commodity sold
   - Total value of production for a county based on the 2007 Census of Agriculture is available for all states.
   - Expected sales for an operation (unit) recorded on the NASS list frame are available for all states at the unit-level.

3. Other farm characteristics
   - Farm type is a unit level categorical covariate, available for all states. Farms are partitioned into 17 farm types on the NASS list frame. To define a covariate, the farm types are aggregated into two groups: (1) grains/oilseeds, and (2) other.
   - Acres rented for non-irrigated cropland recorded on the NASS Cash Rent Survey are available at the unit level for all states.

3 Bivariate hierarchical Bayesian model

The correlation between the 2009 and 2010 cash rental rates observed in Section 2.1.1 suggests that using the information in the data from 2009 has the potential to improve the predictions for 2010. A bivariate hierarchical model for a state is specified as a way to incorporate the data for both years. Let $a_{ij,t}$ and $y_{ij,t}$ be the acres and dollars per acre, respectively, rented by operator j in county i and year $t\,(t = 09, 10)$, and let $\mathbf{x}_{ij,t}$ be the associated column vector of auxiliary variables with dimension p_t. For covariates that are constant across years and individuals, $\mathbf{x}_{ij,t} = \mathbf{x}_{i109}$. Let $w_{ij,t} = a_{ij,t} N_{g(ijt)} n_{g(ijt)}^{-1}$, where $N_{g(ijt)}$ and $n_{g(ijt)}$

are the population size and number of respondents, respectively, in year t for the stratum g that contains unit (ij).

To specify the model, we divide the respondents into three sets:

- Set 1 consists of units (ij) that report a non-irrigated cash rental rate in both 2009 and 2010.
- Set 2 consists of units (ij) that only report a non-irrigated cash rental rate in 2009.
- Set 3 consists of units (ij) that only report a non-irrigated cash rental rate in 2010.

We assume that observations in set 1 satisfy the bivariate model

$$\begin{pmatrix} y_{ij,09} \\ y_{ij,10} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_{ij,09}\boldsymbol{\beta}_{09} + \nu_{i,09} + e_{ij,09} \\ \mathbf{x}'_{ij,10}\boldsymbol{\beta}_{10} + \nu_{i,10} + e_{ij,10} \end{pmatrix}, \tag{3.1}$$

where

$$\begin{pmatrix} e_{ij,09} \\ e_{ij,10} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0},\ \mathbf{D}_{wij}^{-0.5}\boldsymbol{\Sigma}_{ee}\mathbf{D}_{wij}^{-0.5}\right), \tag{3.2}$$

$\mathbf{D}_{wij} = \mathrm{diag}\left(w_{ij,09},\ w_{ij,10}\right)$, and

$$\begin{pmatrix} \nu_{i,09} \\ \nu_{i10} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\nu\nu}\right). \tag{3.3}$$

We denote the diagonal elements of $\boldsymbol{\Sigma}_{ee}$ corresponding to 2009 and 2010 by σ_{ee09} and σ_{ee10}, respectively. For units (ij) in set 2 or 3, we assume

$$y_{ij,t} = \mathbf{x}'_{ij,t}\boldsymbol{\beta}_t + \nu_{i,t} + e^*_{ij,t}, \tag{3.4}$$

where $e^*_{ij,t} \sim \mathrm{N}\left(0,\ w_{ij,t}^{-1}\,\tau^2_{e,t}\right)$, $t = 09$ for set 2, and $t = 10$ for set 3. The model not only allows the variances for the unit-level errors to differ across time points but also allows the variances of unit-level errors for units that respond in both time points to differ from the variances for units that only respond in one time-point. The quantity to predict for 2010 is

$$\theta_{i,10} = \bar{\mathbf{x}}'_{N_i,10}\boldsymbol{\beta}_{10} + \nu_{i,10}, \tag{3.5}$$

where $\bar{\mathbf{x}}_{N_i,10}$ is the population mean of the covariates for county i.

The variances of the unit-level errors, $e_{ij,t}$ and $e^*_{ij,t}$, are assumed to be inversely proportional to the weight, $w_{ij,t}$, for two reasons. First, incorporating the weights in the model aims to reduce bias that could arise if the design is informative for the model. As explained in Section 2, the weights depend on the dollar value of the land rented from the previous year. Therefore, the possibility that the sample design may be informative for a model without the weights is plausible. If $\boldsymbol{\Sigma}_{ee}$ and $\boldsymbol{\Sigma}_{\nu\nu}$ are diagonal, and if $\tau^2_{e,t} = \sigma_{eet}$, then in a frequentist framework, an empirical best linear unbiased predictor for the county i mean in year

t is the design-consistent pseudo-eblup of You and Rao (2002). The second reason to incorporate the weights is that the variances of residuals from preliminary analyses decrease as the acres increase.

Diffuse, proper priors are specified for the unknown regression coefficients and variances. Specifically, $\boldsymbol{\beta}_t \sim \mathrm{N}(\mathbf{0}, 10^6\mathbf{I})$, and $\tau_{e,t}^2 \sim \text{inverse} - \text{gamma}(0.001, 0.001)$. The covariance matrices, $\boldsymbol{\Sigma}_{ee}$ and $\boldsymbol{\Sigma}_{vv}$ have inverse-Wishart prior distributions with shape parameter 0.01 and a diagonal scale matrix with diagonal elements 0.001. The parameterizations for the inverse-gamma and inverse-Wishart distributions are from Gelman, Carlin, Stern and Rubin (2009). We choose priors with conjugate forms for computational simplicity. The choices of the hyperparameters are selected to be un-informative relative to the data for the Cash Rents Survey application.

3.1 Gibbs sampling and posteriors

We use Gibbs sampling to obtain a Monte Carlo approximation to the posterior distribution. An analysis of BGR statistics (Gelman et al., 2009) based on three MCMC chains, each with 20,000 iterations, indicated that 1,000 iterations is sufficient for burn-in. The analyses in Section 4 are based on one chain of length 20,000 for each of the three states, Iowa, Kansas and Texas, where the first 1,000 iterations are discarded for burn-in. By the choices of the likelihood and the priors, the full conditional distributions are known distributions. See Appendix A.

3.2 Prediction and MSE estimation

If $\bar{\mathbf{x}}_{N_i,10}$ is known, the Bayes predictor of $\theta_{i,10}$ for squared error loss is

$$\tilde{\theta}_{i,10}^{B} = E\left[\theta_{i,10} \,\middle|\, (\mathbf{y}, \mathbf{x}), \bar{\mathbf{x}}_{N_i,10}\right] = \bar{\mathbf{x}}'_{N_i,10}\hat{\boldsymbol{\beta}}_{10} + E\left[v_{i,10} \,\middle|\, (\mathbf{y}, \mathbf{x})\right], \tag{3.6}$$

where $\hat{\boldsymbol{\beta}}_{10} = E[\boldsymbol{\beta}_{10} \,|\, (\mathbf{y}, \mathbf{x})], (\mathbf{y}, \mathbf{x})$ denotes the observed cash rental rates and covariates for the two years, and the second equality in (3.6) follows from (3.5) and linearity of expectation. The posterior mean squared error of $\tilde{\theta}_{i,10}^{B}$ is

$$E\left[\left(\tilde{\theta}_{i,10}^{B} - \theta_{i,10}\right)^2 \,\middle|\, (\mathbf{y}, \mathbf{x}), \bar{\mathbf{x}}_{N_i,10}\right] = V\left\{\theta_{i,10} \,\middle|\, (\mathbf{y}, \mathbf{x}), \bar{\mathbf{x}}_{N_i,10}\right\}. \tag{3.7}$$

As discussed in Section 2, the population mean of the covariates, $\bar{\mathbf{x}}_{N_i,10}$, is not available for unit-level covariates in the Cash Rent Survey application. To define a predictor, we add a model for the covariate mean. See Lohr and Prasad (2003) for an approach that begins with a model specification for the unit level covariates. Partition $\mathbf{x}_{ij,10}$ into two sub-vectors, $\mathbf{x}_{ij,10}^{(1)}$ and $\mathbf{x}_{ij,10}^{(2)}$, where $\mathbf{x}_{ij,10}^{(1)}$ contains county-level covariates, and $\mathbf{x}_{ij,10}^{(2)}$ contains unit-level covariates. Assume $\bar{\mathbf{x}}_{wi10} \,|\, \bar{\mathbf{x}}_{N_i,10} \sim \mathrm{N}\left(\bar{\mathbf{x}}_{N_i,10}, \mathbf{V}_{xxi,10}\right)$, where $\bar{\mathbf{x}}_{wi10} = \left(\sum_{j=1}^{n_{i10}} w_{ij,10}\right)^{-1}\left(\sum_{j=1}^{n_{i10}} w_{ij,10}\mathbf{x}_{ij,10}\right)$, n_{i10} is the sum of the number of units in set 1 and in set 3, and $\mathbf{V}_{xxi,10}$ is known. The elements of $\mathbf{V}_{xxi,10}$ corresponding to $\mathbf{x}_{ij,10}^{(1)}$ are 0, and we explain how we obtain the elements of $\mathbf{V}_{xxi,10}$ corresponding to unit-level covariates in Appendix B. The Central Limit Theorem supports the assumption of normality for $\bar{\mathbf{x}}_{wi10}$ even if the distribution of the unit-level covariate values is not normal

(Kim, Park and Lee, 2017). Assuming $\bar{\mathbf{x}}_{N_i,10}$ has a flat prior, $\bar{\mathbf{x}}_{N_i,10} \,|\, \bar{\mathbf{x}}_{wi10} \sim \mathrm{N}\left(\bar{\mathbf{x}}_{wi10}, \mathbf{V}_{xxi,10}\right)$. The Bayes predictor of $\theta_{i,10}$ for squared error loss under the extended model in which the population mean of the covariates is unknown is

$$\hat{\theta}^{B}_{i,10} = \bar{\mathbf{x}}'_{wi10}\hat{\boldsymbol{\beta}}_{10} + E\left[v_{i,10} \,|\, (\mathbf{y}, \mathbf{x})\right]. \tag{3.8}$$

The posterior mean squared error of $\hat{\theta}^{B}_{i,10}$ is

$$\begin{aligned}
E\left[\left(\hat{\theta}^{B}_{i,10} - \theta_{i,10}\right)^2 \,\middle|\, (\mathbf{y}, \mathbf{x})\right] &= E\left\{\left(\hat{\theta}^{B}_{i,10} - \tilde{\theta}^{B}_{i,10} + \tilde{\theta}^{B}_{i,10} - \theta_{i,10}\right)^2 \middle|\, (\mathbf{y}, \mathbf{x})\right\} \\
&= E\left\{\left(\hat{\theta}^{B}_{i,10} - \tilde{\theta}^{B}_{i,10}\right)^2 \middle|\, (\mathbf{y}, \mathbf{x})\right\} \\
&\quad + 2E\left\{E\left[\left(\hat{\theta}^{B}_{i,10} - \tilde{\theta}^{B}_{i,10}\right)\left(\tilde{\theta}^{B}_{i,10} - \theta_{i,10}\right) \middle|\, (\mathbf{y}, \mathbf{x}), \bar{\mathbf{x}}_{N_i,10}\right] \middle|\, (\mathbf{y}, \mathbf{x})\right\} \\
&\quad + V\left\{\theta_{i,10} \,|\, (\mathbf{y}, \mathbf{x})\right\} \\
&= \hat{\boldsymbol{\beta}}'_{10} V\left\{\bar{\mathbf{x}}_{N_i,10} \,|\, \bar{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} + V\left\{\theta_{i,10} \,|\, (\mathbf{y}, \mathbf{x})\right\} \\
&= \hat{\boldsymbol{\beta}}'_{10} V\left\{\bar{\mathbf{x}}_{N_i,10} \,|\, \bar{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} \\
&\quad + V\left\{\bar{\mathbf{x}}'_{wi10}\boldsymbol{\beta}_{10} + v_{i,10} + \left(\bar{\mathbf{x}}_{N_i,10} - \bar{\mathbf{x}}_{wi10}\right)' \boldsymbol{\beta}_{10} \,|\, (\mathbf{y}, \mathbf{x})\right\} \\
&\approx \hat{\boldsymbol{\beta}}'_{10} V\left\{\bar{\mathbf{x}}_{N_i,10} \,|\, \bar{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10} + V\left\{\bar{\mathbf{x}}'_{wi10}\boldsymbol{\beta}_{10} + v_{i,10} \,|\, (\mathbf{y}, \mathbf{x})\right\},
\end{aligned} \tag{3.9}$$

where the final approximation assumes that the $\mathbf{Cov}\left\{\bar{\mathbf{x}}'_{wi10}\boldsymbol{\beta}_{10} + v_{i,10}, \left(\bar{\mathbf{x}}_{N_i,10} - \bar{\mathbf{x}}_{wi10}\right)' \boldsymbol{\beta}_{10} \,|\, (\mathbf{y}, \mathbf{x})\right\}$ is negligible. A comparison of (3.7) and (3.9) shows that the term $\hat{\boldsymbol{\beta}}'_{10} V\left\{\bar{\mathbf{x}}_{N_i,10} \,|\, \bar{\mathbf{x}}_{wi10}\right\}\hat{\boldsymbol{\beta}}_{10}$ accounts for the increase in posterior MSE due to replacing $\bar{\mathbf{x}}_{N_i,10}$ in (3.6) with $\bar{\mathbf{x}}_{wi10}$ in (3.8). To quantify the posterior MSE of $\hat{\theta}^{B}_{i,10}$, we use

$$\mathrm{MSE}\left(\hat{\theta}^{B}_{i,10}\right) = \widehat{\mathrm{MSE}}_{1i} + \widehat{\mathrm{MSE}}_{2i}, \tag{3.10}$$

where $\widehat{\mathrm{MSE}}_{1i} = V\left\{\bar{\mathbf{x}}'_{wi10}\boldsymbol{\beta}_{10} + v_{i,10} \,|\, (\mathbf{y}, \mathbf{x})\right\}$, and $\widehat{\mathrm{MSE}}_{2i} = \hat{\boldsymbol{\beta}}'_{10}\mathbf{V}_{xxi,10}\hat{\boldsymbol{\beta}}_{10}$. In the application of Section 4, we evaluate the effect of including the term $\widehat{\mathrm{MSE}}_{2i}$, which accounts for the increase in posterior MSE due to use of the sample mean of the covariate instead of the population mean, on the posterior MSE of the predictor.

3.3 Two-stage benchmarking

NASS obtains estimates of cash rental rates at the state level using data from a national survey conducted in June (the June Area Survey) in addition to the Cash Rent Survey. The state estimates are published before the county-level data from the Cash Rent Survey are fully processed. NASS also establishes estimates of cash rental rates for agricultural statistics districts. To retain internal consistency, appropriately weighted sums of county estimates must equal the district estimates and appropriately weighted sums of district estimates must equal the previously published state estimate. Letting $\hat{\theta}_{i10}$ be the benchmarked predictor for 2010, the benchmarking restrictions for a single time-point are defined by

$$\sum_{i \in d_k} w_{i10} \hat{\theta}_{i10} = \hat{\lambda}_{k10}, \tag{3.11}$$

and

$$\sum_{k=1}^{K} \eta_{k10} \hat{\lambda}_{k10} = \theta_{\text{pub10}}, \tag{3.12}$$

where $k = 1, \ldots, K$ index the districts, $w_{i10} = \left(\sum_{i \in d_k} z_{i10} \right)^{-1} z_{i10}$,

$$\eta_{k10} = \left(\sum_{k=1}^{K} \sum_{i \in d_k} z_{i10} \right)^{-1} \sum_{i \in d_k} z_{i10},$$

z_{i10} is the direct estimator of the acres rented in county i in year 2010, d_k is the index set for the counties in district k, $\hat{\lambda}_{k10}$ is the final estimate of the average cash rental rate for district k, and θ_{pub10} is the published estimate of the state-level cash rent per acre. We consider estimates for the year 2010 in (3.11) and (3.12) because we focus on estimation for 2010 in the analysis of Section 4.

We use the two-stage benchmarking procedure proposed by Ghosh and Steorts (2013) to define benchmarked estimates. The benchmarked estimates minimize the quadratic form

$$g(\mathbf{c}, \mathbf{b}) = \sum_{k=1}^{K} \sum_{i \in d_k} \xi_i \left(\hat{\theta}_{i10}^{B} - c_i \right)^2 + \sum_{k=1}^{K} \rho_k \left(\hat{\theta}_{k10,w}^{B} - b_k \right)^2 \tag{3.13}$$

subject to the constraints in (3.11) and (3.12), where $\mathbf{c} = (c_1, \ldots, c_D)$, D denotes the total number of counties, $\mathbf{b} = (b_1, \ldots, b_K)$, $\hat{\theta}_{k10,w}^{B} = \sum_{i \in d_k} w_{i10} \hat{\theta}_{i10}^{B}$, and (ρ_k, ξ_i) are constants selected by the analyst. We set $\xi_i = w_{i10}$ and $\rho_k = \eta_{k10}$, which gives the benchmarked estimates

$$\hat{\theta}_{i10} = \hat{\theta}_{i10}^{B} + \hat{\lambda}_{k(i)10} - \hat{\theta}_{k(i)10,w}^{B}, \tag{3.14}$$

with

$$\hat{\lambda}_{k(i)10} = \hat{\theta}_{k(i)10,w}^{B} + \frac{\left(\theta_{\text{pub10}} - \hat{\theta}_{w10}^{B} \right) \eta_{k(i)10} \left(1 + \eta_{k(i)10} \right)^{-1}}{\sum_{i \in d_{k(i)10}} \eta_{k(i)10}^2 \left(1 + \eta_{k(i)10} \right)^{-1}}, \tag{3.15}$$

for county i and district $k(i)$, respectively, where $k(i)$ is the district containing county i. In (3.15), $\hat{\theta}_{w10}^{B} = \sum_{k=1}^{K} \eta_{k10} \hat{\theta}_{k10,w}^{B}$. Each of the benchmarked estimates in (3.14) and (3.15) is a sum of the hierarchical Bayes predictor and an adjustment term. If the hierarchical Bayes predictor for the state is larger (smaller) than the previously published state total, then the adjustment is negative (positive), and the benchmarked county and district estimates are smaller (larger) than the hierarchical Bayes predictors. The posterior mean squared error of the benchmarked predictor for year t is

$$\text{MSE}_{i10}^{B\text{Bench}} = \text{MSE}\left(\hat{\theta}_{i10}^{B} \right) + \left(\hat{\theta}_{i10}^{B} - \hat{\theta}_{i10} \right)^2, \tag{3.16}$$

where $\text{MSE}\left(\hat{\theta}_{i10}^{B}\right)$ is defined in (3.10). See (You, Rao and Dick, 2004) for a derivation of the posterior MSE of a benchmarked predictor.

4 Results for non-irrigated cropland in Iowa, Kansas, and Texas

The model of Section 3 was fit to the non-irrigated cropland cash rental rates reported on the 2009 and 2010 Cash Rent Surveys for Iowa, Kansas, and Texas. These three states were chosen to reflect a range of situations. All counties in Iowa have estimates for corn yields, and cash renting is a relatively common way to rent non-irrigated cropland. Kansas is more agriculturally diverse than Iowa. According to agricultural specialists at NASS, share-renting is a more common way to rent land than cash renting in many parts of Texas, which may explain why realized sample sizes for some Texas counties are as small as zero or one report.

4.1 Covariate selection

The potential covariates for Iowa, Kansas, and Texas are listed in Section 2.2. For each state, the covariates include four variables related to the NCCPI, the total value of production for a county based on the 2007 Census of Agriculture, the expected sales for an operation recorded on the NASS list frame, the farm type recorded on the NASS list frame, and the acres rented for non-irrigated cropland recorded on the NASS Cash Rent Survey. For Iowa, an additional covariate is the corn yield for the county. For Kansas, an additional covariate is the non-irrigated yield index.

The covariates for each state were selected according to the following procedure. First, univariate models were fit to the data for 2009 and 2010 separately using maximum likelihood estimation. The univariate model used for covariate selection is of the form

$$y_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\alpha}_t + \nu_{it} + \epsilon_{ijt}, \tag{4.1}$$

where $\epsilon_{ijt} \sim \text{N}\left(0, \sigma^2_{\epsilon,t}\right)$, and $\nu_{it} \sim \text{N}\left(0, \sigma^2_{\nu,t}\right)$. The data for each farm operator who reported a non-irrigated cropland cash rental rate in year t were used to fit the univariate model for year t, regardless of whether or not the unit also reported a cash rental rate in year $s\,(s \neq t)$. The `R` function `lmer` in the package `nlme` is used for maximum likelihood estimation. For each year, step-wise selection using the `R` function `stepAIC` is performed using the BIC measure. The selected covariates are the variables that are in the minimum BIC models for both the 2009 and 2010 univariate models. We acknowledge that the minimum BIC model is a local minimum identified by the `stepAIC` procedure rather than a global minimum. The selected covariates for Iowa, Kansas, and Texas are as follows:

- Iowa: corn yield, expected sales, non-irrigated acres rented for cash.
- Kansas: non-irrigated yield index, expected sales, farm type.
- Texas: max-NCCPI, expected sales, farm type.

4.2 Estimates of correlation parameters

The exploratory analysis of Section 2.1 suggests a substantial correlation between the non-irrigated cropland cash rental rates for 2009 and 2010. Table 4.1 contains summaries of the posterior distributions of the correlations in the bivariate HB model defined in Section 3.1. The columns labeled "Median" are the posterior medians of the correlations, and lower and upper endpoints of the 95% credible intervals are the 2.5 and 97.5 percentiles of the posterior distributions of the correlations. Even though the variances of e_{ij09} and e_{ij10} are proportional to the inverses of the weights, the correlation is a constant because the weights cancel in the definition of the correlation.

Table 4.1
Posterior distributions of correlations between 2009 and 2010

	$\text{Cor}\{\nu_{i09}, \nu_{i10}\}$		$\text{Cor}\{e_{ij09}, e_{ij10}\}$	
State	**Median**	**95% Credible Interval**	**Median**	**95% Credible Interval**
Iowa	0.746	[0.611, 0.839]	0.570	[0.548, 0.592]
Kansas	0.919	[0.870, 0.950]	0.727	[0.701, 0.751]
Texas	0.884	[0.831, 0.921]	0.691	[0.667, 0.714]

The posterior medians of the county-level and unit-level correlations exceed 0.74 and 0.57, respectively. The lower endpoints of the 95% credible intervals exceed 0.61 and 0.54 for the county-level and unit-level correlations, respectively. For each state, the correlations at the level of the county are larger than the correlations for individual units. The significant correlations suggest the potential for an efficiency gain for the predictors relative to a univariate model.

4.3 Comparison of 2010 predictors for bivariate and univariate models

To demonstrate the gain in efficiency due to the use of the bivariate model relative to a univariate model, we compare the posterior mean squared errors of the predictors from the bivariate model to the posterior mean squared errors of the predictors from a corresponding univariate model. The assumptions of the univariate models are the same as the assumptions of the bivariate models except that the covariance parameters in $\boldsymbol{\Sigma}_{ee}$ and $\boldsymbol{\Sigma}_{\nu\nu}$ are assumed to equal zero. To fit the univariate models, we use inverse-gamma prior distributions for σ_{eet} and $\sigma_{\nu\nu t}$ $(t = 09, 10)$.

To compare the bivariate and univariate models, we define the relative posterior MSE (RelMSE) for county i by

$$\text{RelMSE}_{i,10} = \frac{\text{MSE}_{i10}^{B\text{Bench}}}{\text{MSE}_{i10}^{\text{UNIBench}}}, \tag{4.2}$$

where $\text{MSE}_{i10}^{B\text{Bench}}$ is defined in (3.16) and $\text{MSE}_{i10}^{\text{UNIBench}}$ is the posterior MSE based on the corresponding univariate model. The average relative MSEs for Iowa, Kansas, and Texas are 88.71%, 97.27%, and 88.65%, respectively, where the average relative mean squared error for a state is $D^{-1}\sum_{i=1}^{D}\text{RelMSE}_{i,10}$. Note that the effects of both estimating the covariate mean and benchmarking are incorporated in the forms for the

posterior MSE for both the bivariate and univariate models. Because of the significant correlations in the model errors for the two time points, the posterior MSE from a bivariate model is smaller than the posterior MSE from the corresponding univariate model, and the average relative efficiencies are less than one.

To assess the effect of estimating the covariate population mean on the MSE of the predictor, we calculate the average of the ratios $\widehat{\text{MSE}}_{2i}\widehat{\text{MSE}}_{1i}^{-1}$ for $i = 1, \ldots, D$, where $\widehat{\text{MSE}}_{2i}$ and $\widehat{\text{MSE}}_{1i}$ are defined following (3.10). The ratios are 18.21%, 28.20%, and 21.07% for Iowa, Kansas, and Texas, respectively. Compared to Iowa and Texas, the contribution to the prediction MSE due to using the sample covariate mean instead of the population covariate mean is higher in Kansas, and this makes sense since Kansas is more agriculturally diverse. The relatively large average relative MSE for Kansas (97.27%) reflects the relatively large increase in posterior MSE due to estimating the covariate mean.

4.4 Model assessment

To assess model fit, we use the posterior predictive $\text{p}-$value, which measures departures between the observed data and the model. The posterior predictive $\text{p}-$value compares the posterior predictive distribution of selected summary statistics to the corresponding values obtained using the original sample. For the analysis below, we use only the elements observed in both 2009 and 2010 (set 1).

We consider two summary statistics: the mean for each year and the multivariate skewness. The mean for year t is the mean of the observations in set 1 for year t and is defined

$$\bar{y}_t = \left(\sum_{i=1}^{D} |A_i|\right)^{-1} \sum_{i=1}^{D} \sum_{j \in A_i} y_{ijt},$$

where A_i denotes the elements in set 1 for county i. The multivariate skewness is defined by

$$\hat{\gamma}_{1,p} = \left(\sum_{i=1}^{D} |A_i|\right)^{-1} \sum_{i=1}^{D} \sum_{k=1}^{D} \sum_{j \in A_i} \sum_{\ell \in A_i} m_{ijk\ell}^3,$$

where $m_{ijk\ell} = (\mathbf{y}_{ij} - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_{k\ell} - \bar{\mathbf{y}})$, $\mathbf{y}_{ij} = (y_{ij,09}, y_{ij,10})'$, $\bar{\mathbf{y}} = (\bar{y}_{09}, \bar{y}_{10})'$, and $\mathbf{S} = \left(\sum_{i=1}^{D} |A_i| - 1\right)^{-1} \sum_{i=1}^{D} \sum_{j \in A_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})'$.

The posterior predictive $\text{p}-$value is defined as the proportion of summary statistics calculated with samples generated from the posterior predictive distribution that exceed the corresponding value based on the original sample. To be specific, let $T(\mathbf{y}^{(r)})$ be the summary statistic based on the r^{th} data set generated from the posterior predictive distribution, where the procedure to generate data from the posterior predictive distribution is defined in Appendix C. Let $T(\mathbf{y})$ be the corresponding statistic based on the original sample. The posterior predictive $\text{p}-$value is $R^{-1}\sum_{r=1}^{R} I[T(\mathbf{y}^{(r)}) > T(\mathbf{y})]$. A $\text{p}-$value close to 0.5 indicates that the model provides a reasonable fit to the sample data.

Table 4.2 contains the posterior predictive $\text{p}-$values for Iowa, Kansas, and Texas. For Kansas, the posterior predictive values indicate that the model is a good fit to the data. For Iowa and Texas, the posterior

predictive $p-$values indicate lack of fit. A further analysis of residuals suggests that the lack of fit may result from outliers. The posterior predictive $p-$values far from 0.5 may also arise because we only use the observations sampled in both 2009 and 2010 to calculate the posterior predictive $p-$values, while we use the full data set to fit the model.

Table 4.2
Posterior predictive P – values

State	Statistic	P – value
IA	Mean t = 09	1.000
	Mean t = 10	1.000
	Skewness	0.931
KS	Mean t = 09	0.291
	Mean t = 10	0.507
	Skewness	0.371
TX	Mean t = 09	0.025
	Mean t = 10	0.039
	Skewness	0.004

5 Conclusions and future work

We use a bivariate HB model to obtain predictors of county-level cash rental rates for non-irrigated cropland in Iowa, Kansas, and Texas. The model incorporates auxiliary information related to land quality, commodity values, and farm characteristics. Significant correlations exist between the 2009 and 2010 model random effects at both the unit and county levels. As a consequence, using the information in the 2009 cash rent estimates reduces the posterior MSE relative to a univariate model. The analysis of the bivariate HB model provides support that a more refined approach than that of Berg et al. (2014) is possible. To incorporate unit-level covariates with unknown population means, we add a level to the hierarchical model that justifies adding a term to the posterior mean squared error to account for uncertainty in the unknown population means of the unit-level covariates. Unlike Berg et al. (2014), the proposed bivariate HB model allows variability to change over time and accounts for effects of benchmarking on the MSE.

The analysis of the residuals and the posterior predictive $p-$values suggests that accounting for outliers may be an important way to substantially improve the model fit. One option is to consider a heavy-tailed distribution, such as a $t-$distribution or a mixture of normal distributions, that may represent the observed responses more appropriately than the assumed normal distribution. An extension of Gershunskaya (2010) to bivariate framework and Bayesian estimation is one possible way to approach the issue of outliers.

Acknowledgements

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) supported this work. The authors are grateful to Wendy Barboza, Dan Beckler, Angie Considine,

Mark Harris, Sharyn Lavender, Joe Parsons, Scot Rumberg, Scott Shimmin, Curt Stock, and Linda Young from the National Agriculture Statistics Service. Further, the authors thank Bob Dobos from the National Resource Conservation Service and Rich Iovanna from the Farm Service Agency for their assistance in acquiring the National Commodity Crop Productivity Index. Without the generous assistance from these individuals, this research would have been impossible. The views expressed in this paper are those of the authors and do not necessarily represent the views of NASS or the USDA.

Appendix A

To specify the full conditional distributions for Gibbs sampling, we introduce notation. Let Θ_γ be the set of parameters except for the parameter denoted by γ. Let $\mathbf{X}_{ij} = \left(\mathbf{z}_{ij,09}, \mathbf{z}_{ij,10}\right)'$, where $\mathbf{z}_{ij,09} = \left(\mathbf{x}'_{ij,09}, \mathbf{0}'_{p_{10}}\right)'$, and $\mathbf{z}_{ij,10} = \left(\mathbf{0}'_{p_{09}}, \mathbf{x}'_{ij,10}\right)'$. Let $\mathbf{y}'_{ij} = \left(y_{ij,09}, y_{ij,10}\right)$. Let A_i be the set of units (farm operators) in county i that are in set 1, $B_{i,09}$ be the set of units in county i that are in set 2, and $B_{i,10}$ be the set of units in county i that are in set 3, where set 1, set 2, and set 3 are defined in Section 3. Full conditionals are as follows.

1. $\boldsymbol{\beta}\,|\left(\Theta_{\boldsymbol{\beta}}, \mathbf{y}\right) \sim \mathrm{N}\left(\boldsymbol{\Sigma}_{\beta\beta}\mathbf{r}_\beta, \boldsymbol{\Sigma}_{\beta\beta}\right)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\beta\beta} &= \left[\sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{X}'_{ij}\mathbf{D}_{wij}^{0.5}\,\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{D}_{wij}^{0.5}\mathbf{X}_{ij} + 10^{-6}\mathbf{I}_{p_{09}+p_{10}} + \boldsymbol{\Omega}\right]^{-1} \\
\boldsymbol{\Omega} &= \text{block-diag}\left(\tau_{e,09}^{-2}\sum_{i=1}^{D}\sum_{j\in B_{i,09}} w_{ij,09}\mathbf{x}_{ij,09}\mathbf{x}'_{ij,09}, \tau_{e,10}^{-2}\sum_{i=1}^{D}\sum_{j\in B_{i,10}} w_{ij,10}\mathbf{x}_{ij,10}\mathbf{x}'_{ij,10}\right) \\
\mathbf{r}_\beta &= \sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{X}'_{ij}\mathbf{D}_{wij}^{0.5}\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{D}_{wij}^{0.5}\left(\mathbf{y}_{ij} - \mathbf{v}_i + \mathbf{r}_{\beta 2}\right),
\end{aligned} \tag{A.1}$$

and

$$\mathbf{r}_{\beta 2} = \begin{pmatrix} \sum_{i=1}^{D}\sum_{j\in B_{i,09}} \tau_{e,09}^{-2} w_{ij,09}\mathbf{x}_{ij,09}\left(y_{ij,09} - \nu_{i,09}\right) \\ \sum_{i=1}^{D}\sum_{j\in B_{i,10}} \tau_{e,10}^{-2} w_{ij,10}\mathbf{x}_{ij,10}\left(y_{ij,10} - \nu_{i,10}\right) \end{pmatrix}. \tag{A.2}$$

2. $\boldsymbol{\Sigma}_{ee}\,|\left(\boldsymbol{\Theta}_{\boldsymbol{\Sigma}_{ee}}, \mathbf{y}\right) \sim \text{Inverse-Wishart}\left(\mathbf{A}_e, d_e\right)$, where $d_e = \sum_{i=1}^{D}|A_i| + 0.001$, and

$$\mathbf{A}_e = \sum_{i=1}^{D}\sum_{j\in A_i}\mathbf{D}_{w_{ij}}^{0.5}\left(\mathbf{y}_{ij} - \mathbf{v}_i - \mathbf{X}_{ij}\boldsymbol{\beta}\right)\left(\mathbf{y}_{ij} - \mathbf{v}_i - \mathbf{X}_{ij}\boldsymbol{\beta}\right)'\mathbf{D}_{w_{ij}}^{0.5}. \tag{A.3}$$

3. $\boldsymbol{\Sigma}_{vv}\,|\left(\boldsymbol{\Theta}\mathbf{Q}_{\boldsymbol{\Sigma}_{vv}}, \mathbf{y}\right) \sim \text{Inverse-Wishart}\left(\mathbf{A}_v, d_v\right)$, where

$$d_v = D + 0.001, \tag{A.4}$$

and

$$\mathbf{A}_{v} = \sum_{i=1}^{D} \mathbf{v}_i \mathbf{v}_i'. \tag{A.5}$$

$\tau_{e,t}^2 \left| \left(\boldsymbol{\Theta}_{\tau_{e,t}^2}, \mathbf{y} \right) \sim \text{Inverse-Gamma}\left(a_{et}, d_{et}\right), \right.$ where

$$d_{et} = \sum_{i=1}^{D} \left| B_{i,t} \right| + 0.001, \tag{A.6}$$

and

$$a_{et} = \sum_{i=1}^{D} \sum_{j \in B_{it}} \mathbf{D}_{wi_j} \left(y_{ij,t} - v_{i,t} - \mathbf{x}_{ij,t} \boldsymbol{\beta}_t \right)^2. \tag{A.7}$$

4. $\mathbf{v}_i \left| \left(\boldsymbol{\Theta}_{v_i}, \mathbf{y} \right) \sim \text{N}\left(\boldsymbol{\mu}_{vv}, \mathbf{M}_i^{-1}\right), \right.$ where

$$\mathbf{M}_i = \left(\boldsymbol{\Sigma}_{vv}^{-1} + \boldsymbol{\Sigma}_{ee,wi}^{-1} + \boldsymbol{\Omega}_{ee,wi}^{-1} \right)^{-1}, \tag{A.8}$$

$\boldsymbol{\mu}_{vv} = \mathbf{M}_i^{-1} \left(\mathbf{r}_{i_1} + \mathbf{r}_{i_2} \right)$, $\boldsymbol{\Sigma}_{ee,wi} = \sum_{j \in A_i} \mathbf{D}_{w_{ij}}^{0.5} \boldsymbol{\Sigma}_{ee}^{-1} \mathbf{D}_{w_{ij}}^{0.5}$,

$$\begin{aligned} \mathbf{W}_{ee,wi} &= \text{diag}\left(\tau_{e,09}^{-2} \sum_{j \in B_{i09}} w_{ij,09}, \tau_{e,10}^{-2} \sum_{j \in B_{i10}} w_{ij,10} \right), \\ \mathbf{r}_{i_1} &= \sum_{j \in A_i} \mathbf{D}_{w_{ij}}^{0.5} \boldsymbol{\Sigma}_{ee}^{-1} \mathbf{D}_{w_{ij}}^{0.5} \left(\mathbf{y}_{ij} - \mathbf{X}_{ij}' \boldsymbol{\beta} \right), \end{aligned} \tag{A.9}$$

and

$$\mathbf{r}_{i_2} = \begin{pmatrix} \sum_{j \in B_{i,09}} w_{ij,09} \left(y_{ij,09} - \mathbf{x}_{ij,09}' \boldsymbol{\beta}_{09} \right) \tau_{e,09}^{-2} \\ \sum_{j \in B_{i,10}} w_{ij,10} \left(y_{ij,10} - \mathbf{x}_{ij,10}' \boldsymbol{\beta}_{10} \right) \tau_{e,10}^{-2} \end{pmatrix}. \tag{A.10}$$

Appendix B

We define an estimator of the diagonal elements of $V\left\{ \overline{\mathbf{x}}_{wi10} \left| \overline{\mathbf{x}}_{N_i,10} \right. \right\} := \mathbf{V}_{xxi,10}$ corresponding to unit-level covariates, $\mathbf{x}_{ijk10}$ for $k = 1, \ldots, p_{10}$. The variance estimator is based on a working assumption that a probability proportional to size with replacement (PPSWR) sample is a reasonable approximation for the cash rent survey design. As discussed in Cochran (1977), use of a PPSWR approximation is often reasonable if the sampling fraction is less than 10%. Suppose the draw probability for element j in area i for the PPSWR design is $p_{ij} = n_{i10}^{-1} w_{ij,10}^{-1}$. Because $n_{i10} \leq 1$ for some counties, we define the estimator of the diagonal elements of $\mathbf{V}_{xxi10}$ corresponding to unit level covariates as a convex combination of a direct estimator of the within-area variance and a variance estimator that pools information across all counties in

a state. For area i with $n_{i10} > 1$, the estimate of the within-area variance of x_{ijk10} under the assumed PPSWR design (Särndal, Swensson and Wretman, 1992) is given by

$$S^2_{ik10} = \frac{n^2_{i10}}{\left(\sum_{j=1}^{n_{i10}} w_{ij,10}\right)^2 (n_{i10} - 1)} \sum_{j=1}^{n_{i10}} w^2_{ij,10} \left(x_{ijk10} - \bar{x}_{wik10}\right)^2,$$

where $\bar{x}_{wik10}$ is the k^{th} element of $\bar{\mathbf{x}}_{wi10}$. The pooled estimator of the variance is defined by

$$S^2_{pk10} = \frac{1}{w^2_{..10}\left(\tilde{n}_{10} - \tilde{D}_{10}\right)} \sum_{i=1}^{D} \left(n^2_{i10} \sum_{j=1}^{n_{i10}} w^2_{ij,10} \left(x_{ijk10} - \bar{x}_{wik10}\right)^2 \right) I\left[n_{i10} > 1\right],$$

where $w_{..10} = \sum_{i=1}^{D} \left(\sum_{j=1}^{n_i} w_{ij,10}\right) I\left[n_i > 1\right]$, $\tilde{n}_{10} = \sum_{i=1}^{D} n_{i10} I\left[n_{i10} > 1\right]$, and $\tilde{D}_{10} = \sum_{i=1}^{D} I\left[n_{i10} > 1\right]$. The element of the diagonal covariance matrix $\mathbf{V}_{xxi,10}$ corresponding to the k^{th} unit level covariate is then given by

$$\hat{V}\left\{\bar{x}_{wik10}\right\} = n^{-1}_{i10} \hat{S}^2_{ik10} I\left[n_{i10} \neq 1\right] + n^{-1}_{i10} S^2_{pk10} I\left[n_{i10} = 1\right], \tag{B.1}$$

where

$$\hat{S}^2_{ik10} = \frac{n_{i10}}{n_{i10} + 1} S^2_{ik10} + \frac{1}{n_{i10} + 1} S^2_{pk10}. \tag{B.2}$$

We provide a heuristic justification for the combination in (B.2), which is related to Haff (1980). Let $S^2 = n^{-1} \sum_{i=1}^{n} X^2_i$, where $X_i \sim \mathrm{N}(0, \sigma^2)$. Assume $\sigma^2 \sim$ Inverse-Gamma(α, β), where $E\left[\sigma^2\right] := v = \beta(\alpha - 1)^{-1}$. Then,

$$E\left[\sigma^2 \mid S^2\right] = \frac{2(\alpha - 1)v}{n + 2(\alpha - 1)} + \frac{nS^2}{n + 2(\alpha - 1)}.$$

In application to estimation of county-level cash rental rates, S^2_{ik10} plays the role of S^2 and S^2_{pk10} plays the role of v. Taking $\alpha = 1.5$ gives the desired multiplier.

Appendix C

Data simulation from the posterior distributions

Consider the posterior samples for $\boldsymbol{\beta}_{09}$, $\boldsymbol{\beta}_{10}$, $\boldsymbol{\Sigma}_{vv}$ and $\boldsymbol{\Sigma}_{ee}$, denoted by $\boldsymbol{\beta}^s_{09}$, $\boldsymbol{\beta}^s_{10}$, $\boldsymbol{\Sigma}^s_{vv}$ and $\boldsymbol{\Sigma}^s_{ee}$, respectively, for $s = 1, \ldots, S$. Define

$$\boldsymbol{\Sigma}^s_{eeij} := \mathbf{D}^{-0.5}_{wij} \boldsymbol{\Sigma}^s_{ee} \mathbf{D}^{-0.5}_{wij},$$

for $s = 1, \ldots, S$. Draw replicates ν_{i09}^r, ν_{i10}^r, y_{ij09}^r and y_{ij10}^r, for $r = 1, \ldots, R$, following model (1-3) and properties of the multivariate conditional normal distribution as follows:

$$
\begin{aligned}
\nu_{i09}^r &\sim N\left(\mathbf{0},\, \boldsymbol{\Sigma}_{\nu\nu,(11)}^r\right) \\
\nu_{i10}^r &\sim N\left(\left(\boldsymbol{\Sigma}_{\nu\nu,(11)}^r\right)^{-1} \boldsymbol{\Sigma}_{\nu\nu,(12)}^r \nu_{i09}^r,\, \left(\boldsymbol{\Sigma}_{\nu\nu,(11)}^r\right)^{-1} \boldsymbol{\Sigma}_{\nu\nu,(11)}^r \boldsymbol{\Sigma}_{\nu\nu,(22)}^r - \left(\boldsymbol{\Sigma}_{\nu\nu,(12)}^r\right)^2\right), \\
\mu_{i09}^r &= \mathbf{x}_{ij09}' \boldsymbol{\beta}_{09}^r, \\
y_{ij09}^r &\sim N\left(\mu_{i09}^r + \nu_{i09}^r,\, \boldsymbol{\Sigma}_{eeij,(11)}^r\right) \\
\mu_{i10}^r &= \mathbf{x}_{ij10}' \boldsymbol{\beta}_{10}^r, \\
y_{ij10}^r &\sim N\left(\mu_{i10}^r + \nu_{i10}^r + \left(\boldsymbol{\Sigma}_{eeij,(11)}^r\right)^{-1} \boldsymbol{\Sigma}_{eeij,(12)}^r \left(y_{ij09}^r - \mu_{i09}^r - \nu_{i09}^r\right),\right. \\
&\qquad \left. \left(\boldsymbol{\Sigma}_{eeij,(11)}^r\right)^{-1} \left(\boldsymbol{\Sigma}_{eeij,(11)}^r \boldsymbol{\Sigma}_{eeij,(22)}^r - \left(\boldsymbol{\Sigma}_{eeij,(12)}^r\right)^2\right)\right).
\end{aligned}
$$

Although the number of posterior samples is $S = 20{,}000$, we construct $R = 1{,}901$ replicates, where r is selected from the sequence 1,000 to T by skipping every 10 samples.

References

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

Berg, E., Cecere, W. and Ghosh, M. (2014). Small area estimation for county-level farmland cash rental rates. *Journal of Survey Statistics and Methodology*, 2, 1-37.

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

Datta, G.S., Day, B. and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā: The Indian Journal of Statistics*, 60, 344-362.

Dhuyvetter, D., and Kastens, T. (2009). *Kansas Land Values and Cash Rents at the County Level*.

Dobos, R.R., Sinclair, H.R. and Robotham, M.P. (2012). *User Guide for the National Commodity Crop Productivity Index (NCCPI), Version 2.0*, NRCS/USDA publication.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2009). *Bayesian Data Analysis: Second Edition*, CRC Press.

Gershunskaya, J. (2010). Robust small area estimation using a mixture model. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Ghosh, M., and Steorts, R. (2013). Two-stage Bayesian benchmarking as applied to small area estimation. *TEST*, 22, 670-687.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 586-597.

Kim, J.K., Park, S. and Lee, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling. *Canadian Journal of Statistics*, DOI: 10.1002/cjs.11339.

Lohr, S.L., and Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *Canadian Journal of Statistics*, 31, 383-396.

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

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

Woodard, S., Paulson, N., Baylis, K. and Woddard, J. (2010). Spatial analysis of Illinois agricultural cash rents. *The Selected Works of Kathy Baylis*. http://works.bepress.com/kathy_baylis/29.

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

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchicial Bayes small area estimators with applications in census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 217-247
Statistics Canada, Catalogue No. 12-001-X

Estimation of response propensities and indicators of representative response using population-level information

Annamaria Bianchi, Natalie Shlomo, Barry Schouten, Damião N. Da Silva and Chris Skinner[1]

Abstract

In recent years, there has been a strong interest in indirect measures of nonresponse bias in surveys or other forms of data collection. This interest originates from gradually decreasing propensities to respond to surveys parallel to pressures on survey budgets. These developments led to a growing focus on the representativeness or balance of the responding sample units with respect to relevant auxiliary variables. One example of a measure is the representativeness indicator, or R-indicator. The R-indicator is based on the design-weighted sample variation of estimated response propensities. It pre-supposes linked auxiliary data. One of the criticisms of the indicator is that it cannot be used in settings where auxiliary information is available only at the population level. In this paper, we propose a new method for estimating response propensities that does not need auxiliary information for non-respondents to the survey and is based on population auxiliary information. These population-based response propensities can then be used to develop R-indicators that employ population contingency tables or population frequency counts. We discuss the statistical properties of the indicators, and evaluate their performance using an evaluation study based on real census data and an application from the Dutch Health Survey.

Key Words: Nonresponse; Missing data; Nonresponse bias; Balanced response.

1 Introduction

Nonresponse bias in surveys is of increasing concern with declining response rates and tighter budgets. National Statistics Institutes (NSIs) charged with conducting national surveys to convey the state of their country's economic, social and demographic characteristics are facing increasing challenges in maintaining the quality of their survey response. In this paper, we focus on one particular survey conducted since 1998 by Statistics Netherlands, The Dutch Health Survey, which up until 2010 was a face-to-face survey. In 2010, online data collection was added as a sequential mode before the face-to-face interviews. The response rates have gradually declined from values close to 70% to values around 60%. Other NSIs and survey organizations have reported declining response rates, particularly when moving to mixed modes of data collection in order to reduce budgets, with respondents pushed towards cheaper modes. However, response rates alone are not enough to judge the quality of the survey response, as nonresponse bias results from the contrast between those responding and not responding to the surveys. Nonresponse bias in the Dutch Health Survey is conjectured to arise from persons with weaker health, certain habits like smoking or fewer dentist visits, and poorer living conditions. Important predictors are age, marital status, income and ethnicity.

A number of indirect measures of nonresponse bias have been developed recently to supplement the traditional response rate. Wagner (2012) provides a taxonomy of such measures: indicators that include only observed auxiliary variables and indicators that also include observed survey variables which may or may

1. Annamaria Bianchi, University of Bergamo, Italy. E-mail: annamaria.bianchi@unibg.it; Natalie Shlomo, University of Manchester, United Kingdom and Social Statistics, School of Social Sciences, Humanities Bridgeford Street Room G17A, University of Manchester M13 9PL United Kingdom. E-mail: natalie.shlomo@manchester.ac.uk; Barry Schouten, Statistics Netherlands and Utrecht University. E-mail: jg.schouten@cbs.nl; Damião N. Da Silva, Universidade Federal do Rio Grande do Norte, Brazil. E-mail: damiao@ccet.ufrn.br; Chris Skinner, London School of Economics and Political Science, United Kingdom. E-mail: c.j.skinner@lse.ac.uk.

not account for nonresponse weighting. The most prominent indicators that only use observed auxiliary variables are R-indicators (Schouten, Cobben and Bethlehem, 2009; Schouten, Shlomo and Skinner, 2011) and balance indicators (Särndal, 2011; Lundquist and Särndal, 2013).

The development of these measures comes at a time when there is an increased interest in adapting data collection (Schouten, Calinescu and Luiten, 2013; Wagner, 2013; Wagner and Hubbard, 2014; Beaumont, Bocci and Haziza, 2014) so that the level of effort targeted at different subgroups as defined by auxiliary variables may be varied over time, possibly through a change of strategy, according to patterns of response (Schouten, Bethlehem, Beulens, Kleven, Loosveldt, Rutar, Shlomo and Skinner, 2012; Särndal and Lundquist, 2014). Both R-indicators and balance indicators must be viewed in conjunction with the auxiliary data that is employed. Different auxiliary variables may lead to different values of the indicators.

In addition, Schouten, Cobben, Lundquist and Wagner (2016) present empirical evidence that it is beneficial for samples to be more balanced with respect to auxiliary variables, even when these variables are used in nonresponse adjustment afterwards. Based on 14 survey data sets they show that, on average, a design with a more representative response has smaller nonresponse biases, even after adjustments on the characteristics for which representativeness was evaluated. Särndal and Lundquist (2014) also found gains in balancing the respondents set, over and beyond those obtained by calibrating the sample. Further, it is worth noting that a more balanced sample leads to less variability in adjustment weights, which is a desirable property as large variation in adjustment weights may inflate standard errors of estimates. Of course, nonresponse adjustment weighting will still be necessary as there will always be some imbalance remaining in the final response dataset.

The auxiliary data used for the response indicator measures may stem from sampling frame data, administrative data and data about the data collection process, called paradata (Kreuter, 2013). Balance indicators and R-indicators are very similar and are often proportional in size. In this paper, we focus on R-indicators. However, much of the discussion and results can easily be translated to balance indicators.

R-indicators presume the availability of auxiliary variables obtained by linking data from, for example, sample frames or registers, to the survey sample. This presumption of linked survey samples may be infeasible in many settings and hampers application. While national statistical institutes often have access to government registrations, university and market researchers usually do not. For indicators to become useful for these researchers, they must be based on different forms of auxiliary information. The only form of auxiliary information that is generally accessible are the sets of statistics produced by the national statistical institutes. These institutes disseminate tables on a wide range of population statistics. This paper develops R-indicators that are based solely on such population statistics and that can be computed without any knowledge about the non-respondents. As an example, market research companies compare the response distributions of a fixed, prescribed set of auxiliary variables to national statistics, termed the gold standard. The R-indicator estimators proposed here allow for monitoring and evaluating gold standard variables during and after data collection.

Although the R-indicators based on population auxiliary information are motivated in this paper from survey data collection practice, they can be applied to any setting with missing data on variables of interest

and (almost) complete auxiliary data. They can for instance, be used to monitor and evaluate the completion of administrative data, which is useful if the data is streamed and gradually accumulated over time. In this case, population based R-indicators would provide an assessment of the representativeness of the streamed administration data. Another useful application for such indicators is to assess the representativeness of linked records. Van der Laan and Bakker (2015) proposed a Linkage Representativeness Indicator (LR-indicator) which examines the similarity of linked records to the target population under investigation.

R-indicators and their statistical properties, as discussed in Shlomo, Skinner and Schouten (2012), relate to the case where we have linked sample level auxiliary information for non-respondents. To develop R-indicators based on population statistics, we propose a new method for estimating response propensities that does not require auxiliary information for non-respondents to the survey. They will be called population-based response propensities. To our knowledge, there is no record in the literature about models for response propensities that employ population information only. In this respect, the current paper is innovative and may be valuable and relevant to other statistical areas as well. In this paper, we concentrate on the use of population-based response propensities in the computation of R-indicators.

With respect to adapting data collection, it is clear that settings where population-based R-indicators are needed are harder for the implementation of these types of adaptive designs as we do not know the values of the covariates for nonrespondents. However, using these types of R-indicators based on population-based auxiliary information, we can make design features more salient to those that are lagging behind in terms of response. So, for example, if young people have lower response rates, we can send a general reminder with more focus on young persons or alternatively instruct interviewers to monitor more carefully those addresses where they expect younger persons.

The auxiliary information for population-based response propensities is obtained from population tables and population counts. In order to do so, we first propose estimating response propensities based on population values, by replacing sample covariance matrices and sample means by known population covariances and population means. Next, using population-based response propensities, we compute estimates for the R-indicator. We call the resulting indicator a population-based R-indicator, and we call the traditional R-indicator a sample-based R-indicator. We focus on three research questions:

- How to extend sample-based response propensities and R-indicators to population-based response propensities and R-indicators?
- What are the statistical properties of population-based R-indicators?
- Are the population-based R-indicators practicable in real survey settings?

In Section 2, we propose a new method for estimating population-based response propensities. In Section 3, we briefly review the definitions and methodology behind R-indicators and then consider their estimation in the population-based setting. In Section 4, we present an evaluation study that is based on drawing samples from real Census data under realistic assumptions about nonresponse in social surveys and evaluate the properties of the population-based R-indicators. In Section 5, we demonstrate the proposed

R-indicators on an application from the Dutch Health Survey of the Netherlands. In Section 6, we end with a discussion and present some caveats related to the proposed indicators and future work.

2 Population-based response propensities

2.1 General notation

We suppose that a sample survey is undertaken, where a sample s is selected from a finite population U. The sizes of s and U are denoted by n and N, respectively. The units in U are labelled $i = 1, 2, \ldots, N$. The sample is assumed to be drawn by a probability sampling design $p(.)$, where the sample s is selected with probability $p(s)$. The first order inclusion probability of unit i is denoted π_i and $d_i = \pi_i^{-1}$ is the design weight. The evaluation study is based on simple random sampling without replacement. Although large-scale national surveys may use more complex two-stage designs, many are generally planned so that all survey units have an equal inclusion probability. We also provide theoretical expressions under more general complex survey designs.

We suppose that the survey is subject to unit nonresponse. The set of responding units is denoted by r, so $r \subset s \subset U$. We denote summation over the respondents, sample and population by Σ_r, Σ_s and Σ_U, respectively. Let r_i be the response indicator variable so that $r_i = 1$ if unit i responds and $r_i = 0$, otherwise. Hence, $r = \{i \in s;\ r_i = 1\}$. We shall suppose that the typical target of inference is a population mean $\bar{Y} = N^{-1}\sum_U y_i$ of a survey variable, taking value y_i for unit i.

We suppose that the data available for estimation purposes consists first of the values $\{y_i;\ i \in r\}$ of the survey variable, observed only for respondents. Secondly, we suppose that information is available on the values $\mathbf{x}_i = \left(x_{1,i}, x_{2,i}, \ldots, x_{K,i}\right)^T$ of a vector of auxiliary variables $\mathbf{X}$. We shall usually suppose each $x_{k,i}$ is a binary indicator variable, where $\mathbf{x}_i$ represents one or more categorical variables, since this will be the case in the applications we consider, but our presentation allows for general $x_{k,i}$ values. We assume that values of $\mathbf{x}_i$ are observed for all respondents so that $\{y_i, x_i;\ i \in r\}$ is observed.

We distinguish two settings: one in which $\mathbf{x}_i$ is known for all sample units, i.e., for both respondents and non-respondents, and one in which $\mathbf{x}_i$ is known only at the aggregate level: the population total $\sum_U \mathbf{x}_i$ and/or the population cross-products $\sum_U \mathbf{x}_i \mathbf{x}_i^T$. We refer to the two types of information as *sample-based auxiliary information* and *aggregate population-based auxiliary information*. The first setting is relevant if the variables making up $\mathbf{X}$ are available on a register. However, as outlined in the introduction, in many countries and surveys the availability of auxiliary information on non-respondents may be limited and the second setting using population-based auxiliary information may be more useful.

2.2 Definition of response propensities

The theory of propensity scores was introduced by Rosenbaum and Rubin (1983) and discussed in the context of survey nonresponse by Little (1986; 1988). Response propensities are defined as the conditional

expectation of the response indicator variable r_i given the values of specified variables and survey conditions: $\rho_X(\mathbf{x}_i) = E_m(r_i \mid \mathbf{x}_i)$, where the vector of auxiliary variables is defined as in Section 2.1. For simplicity, we shall write $\rho_i = \rho_X(\mathbf{x}_i)$ and hence denote the response propensity just by ρ_i. $E_m(.)$ denotes expectation with respect to the model underlying the response mechanism. A detailed discussion of response propensities and their properties is presented in Shlomo et al. (2012). They argue that it is desirable to select auxiliary variables constituting $\mathbf{x}_i$ in such a way that the missing at random assumption, denoted MAR (Little and Rubin, 2002), holds as closely as possible.

2.3 Estimation of response propensities using population-level information

In the case of sample-based auxiliary information, it is possible to estimate response propensities for all sampled units by means of regression models $g(\rho_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, where $g(.)$ is a link function, r_i is the dependent variable, and $\mathbf{x}_i$ is a vector of explanatory variables. Generally, the response propensities are modelled by generalized linear models. Shlomo et al. (2012) use a logistic link function.

In the population-based setting, it is convenient to consider the identity link function. The identity link function is a good approximation to the more widely used logistic link function when response rates are mid-range, between 30% and 70%, which is the typical response rate obtained in national and other surveys. We demonstrate this fact in the evaluation study presented in Section 4 where three ranges of response rates are investigated: low, medium and high. The identity link function also forms the basis for other representativeness indicators in the literature, such as the imbalance and distance indicators proposed by Särndal (2011) some of which are similar to the g-weights calculated in the Generalized Regression Estimators (GREG).

Under the identity link function we assume that the true response propensities satisfy the "linear probability model"

$$\rho_i = \mathbf{x}_i^T\boldsymbol{\beta}, \quad i \in U. \tag{2.1}$$

The linear probability model in (2.1) can be estimated by weighted least squares, where d_i is the design weight. The implied estimator of ρ_i is given by

$$\hat{\rho}_i^{\mathrm{OLS}} = \mathbf{x}_i^T\left(\sum_s d_i\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\sum_s d_i\mathbf{x}_i r_i, \quad i \in s. \tag{2.2}$$

In the case of population-based auxiliary information, we first note that $\sum_s d_i\mathbf{x}_i$ and $\sum_s d_i\mathbf{x}_i\mathbf{x}_i^T$ are unbiased for $\sum_U \mathbf{x}_i$ and $\sum_U \mathbf{x}_i\mathbf{x}_i^T$, respectively and that in large samples we may expect that $\sum_s d_i\mathbf{x}_i \approx \sum_U \mathbf{x}_i$ and $\sum_s d_i\mathbf{x}_i\mathbf{x}_i^T \approx \sum_U \mathbf{x}_i\mathbf{x}_i^T$. It follows from (2.2) that, in the population-based setting, we may approximate $\hat{\rho}_i^{\mathrm{OLS}}$ by

$$\tilde{\rho}_{i,T1} = \mathbf{x}_i^T\mathbf{T}_1^{-1}\sum_r d_k\mathbf{x}_k, \quad i \in r \tag{2.3}$$

where $\mathbf{T}_1 = \sum_U \mathbf{x}_j\mathbf{x}_j^T$. We note that $\tilde{\rho}_{i,T1}$ is computed only on the set of responding units.

The estimator in (2.3) requires knowledge of the population sums of squares and cross-products $\sum_U \mathbf{x}_i\mathbf{x}_i^T$ of the elements of $\mathbf{x}_i$. However the cross-products might be unknown. In that case, we can estimate $\sum_s d_i\mathbf{x}_i\mathbf{x}_i^T$ in (2.2) by rewriting

$$\sum_s d_i \, \mathbf{x}_i \, \mathbf{x}_i^T = \sum_s d_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_s)(\mathbf{x}_i - \overline{\mathbf{x}}_s)^T + N \overline{\mathbf{x}}_s \overline{\mathbf{x}}_s^T, \tag{2.4}$$

where $\overline{\mathbf{x}}_s = \sum_s d_i \mathbf{x}_i / N$. $\overline{\mathbf{x}}_s$ may be replaced by $\overline{\mathbf{x}}_U$ and the covariance matrix

$$\mathbf{S}_{xx} = N^{-1} \sum_s d_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_s)(\mathbf{x}_i - \overline{\mathbf{x}}_s)^T \tag{2.5}$$

may be replaced by its estimate using the response set

$$\hat{\mathbf{S}}_{xx} = \left(\sum_s d_j r_j \right)^{-1} \sum_s d_i r_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_U)(\mathbf{x}_i - \overline{\mathbf{x}}_U)^T. \tag{2.6}$$

We can also estimate (2.6) using propensity weighting by $\tilde{\rho}_i^{-1}$ to adjust for nonresponse bias in the variance of the response propensities relative to a set of X variables.

Combining (2.3), (2.4) and (2.6), we obtain the following estimator:

$$\tilde{\rho}_{i,T2} = \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1} \sum_r d_k \mathbf{x}_k, \quad i \in r, \tag{2.7}$$

where $\hat{\mathbf{T}}_2 = N \hat{\mathbf{S}}_{xx} + N \overline{\mathbf{x}}_U \overline{\mathbf{x}}_U^T$.

We therefore distinguish between two types of aggregated population-based auxiliary information as denoted by the indices "T_1" in (2.3) and "T_2" in (2.7):

TYPE 1 *Full aggregate population-based auxiliary information*: the population cross products are available, i.e., $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ or $\sum_U (\mathbf{x}_i - \overline{\mathbf{x}}_U)(\mathbf{x}_i - \overline{\mathbf{x}}_U)^T$, where $\overline{\mathbf{x}}_U = \sum_U \mathbf{x}_i / N$;

TYPE 2 *Marginal aggregate population-based auxiliary information*: only the population marginal counts are available, i.e., $\sum_U \mathbf{x}_i$.

The first type implies that we have available all two-by-two tables, e.g., age times gender, age times marital status and gender times marital status. This information might be available to a national statistical institute which has access to population registers or detailed population demographics and wishes to use population-based information to monitor data collection due to a lack of sample-based information on the sample frames. The second type is more restrictive as we have only frequency counts, e.g., age, gender, marital status, without any knowledge about the interactions. This information would be routinely available through websites of national statistical institutes and therefore can be used by marketing and other data collection agencies to monitor their data collection.

3 Estimation of R-indicators based on population totals

In this section, we first briefly review the definition and concepts of R-indicators, and their estimation based on sample-level auxiliary information. Details can be found in Shlomo et al. (2012). Next, applying the theory introduced in Section 2.3, we adapt the sample-based R-indicator to the case where auxiliary information is obtained from population tables and population counts. Further, we investigate the statistical properties of this estimator.

3.1 R-indicators

Schouten et al. (2009) introduce the concept of representative response. A response to a survey is said to be *representative with respect to* $\mathbf{X}$ when response propensities are constant for $\mathbf{X}$, i.e.,

$$\rho_i = \rho_X(\mathbf{x}_i) = \bar{\rho}, \quad \forall \mathbf{x}_i,$$

where $\bar{\rho}$ denotes the average response propensity in the population.

The overall measure of representative response is the R-indicator. The R-indicator associated with a set of population response propensities $\{\rho_i: i \in U\}$ is defined as

$$R_\rho = 1 - 2S_\rho, \tag{3.1}$$

where S_ρ denotes the standard deviation of the individual response propensities

$$S_\rho^2 = \frac{1}{N-1}\sum_U (\rho_i - \bar{\rho}_U)^2 = \frac{N}{N-1}\left\{\frac{1}{N}\sum_U \rho_i^2 - \left[\frac{1}{N}\sum_U \rho_i\right]^2\right\}, \tag{3.2}$$

where $\bar{\rho}_U = \sum_U \rho_i / N$.

The *R-indicator* takes values on the interval $\left[1 - \sqrt{\frac{N}{N-1}}, 1\right]$ with the upper value 1 indicating the most representative response, where the ρ_i's display no variation, and the lower value $1 - \sqrt{\frac{N}{N-1}}$ (which is close to 0 for large surveys) indicating the least representative response, where the ρ_i's display maximum variation.

An important related measure of representativeness is the coefficient of variation of the response propensities

$$\mathrm{CV}_\rho = \frac{S_\rho}{\bar{\rho}_U}. \tag{3.3}$$

This is a relevant measure when considering population means or totals as parameters of interest. In those cases, it may be used instead of the R-indicator. For other types of parameters of interest, such as the median or a ratio, other indicators can be used (Brick and Jones, 2008).

The coefficient of variation in (3.3) bounds the absolute nonresponse bias of unadjusted response means for a variable Y divided by its standard deviation. Schouten et al. (2016) also used the coefficient of variation to assess "worst case" nonresponse bias intervals for standard nonresponse adjusted post-survey estimators, such as the generalized regression estimator (GREG) (Deville and Särndal, 1992) and inverse propensity weighting (IPW) (Little, 1988).

3.2 Sample-based R-indicators

In the case of sample-based auxiliary information, it is possible to estimate response propensities for all sampled units. In the following, let $\hat{\rho}_i$ be an estimator for ρ_i. The sample-based estimator for the R-indicator is

$$\hat{R}_{\hat{\rho}} = 1 - 2\hat{S}_{\hat{\rho}}, \tag{3.4}$$

where $\hat{S}^2_{\hat{\rho}}$ is the design-weighted sample variance of the estimated response propensities computed using the first expression in (3.2)

$$\hat{S}^2_{\hat{\rho}} = \frac{1}{N-1}\sum_s d_i \left(\hat{\rho}_i - \hat{\bar{\rho}}_U\right)^2,$$

where $\hat{\bar{\rho}}_U = \left(\sum_s d_i \hat{\rho}_i\right) \Big/ N$.

The sample-based R-indicator defined by (3.4) is a statistic with a certain precision and bias. Shlomo et al. (2012) discuss bias adjustments and confidence intervals for $\hat{R}_{\hat{\rho}}$. These are available in SAS and R code at www.risq-project.eu, and a manual is provided by De Heij, Schouten and Shlomo (2015). We return to the statistical properties in Section 3.4.

3.3 Population-based R-indicators

We demonstrate in Section 4 that the R-indicators depend only mildly on the type of link function when estimating response propensities if response rates are not in the tails, i.e., very high or very low. Furthermore, we obtain similar estimation of R-indicators when population-based response propensities are estimated according to the Type 1 or Type 2 types of information.

In the population-based setting, an estimator for the R-indicator is then

$$\tilde{R}_{\tilde{\rho}} = 1 - 2\tilde{S}_{\tilde{\rho}}, \tag{3.5}$$

where

$$\tilde{S}^2_{\tilde{\rho}} = \frac{N}{N-1}\left\{\frac{1}{N}\sum_r d_i \tilde{\rho}_i - \left(\frac{1}{N}\sum_r d_i\right)^2\right\}, \tag{3.6}$$

and $\tilde{\rho}_i$ denotes either response propensities computed under Type 1 information $\left(\tilde{\rho}_{i,T1}\right)$ or response propensities estimated under Type 2 information $\left(\tilde{\rho}_{i,T2}\right)$.

Notice that the estimation of the R-indicator is based on the second expression for S^2_ρ in (3.2). This choice indeed makes the estimator $\tilde{S}^2_{\tilde{\rho}}$ linear in $\tilde{\rho}_i$, which provides an advantage for bias computations as described in Section 3.4. The evaluation study in Section 4 empirically demonstrates that the two expressions for S^2_ρ are similar for the types of large-scale national surveys under consideration. Furthermore, we use propensity-weighting by $\tilde{\rho}_i^{-1}$ to adjust for nonresponse bias. As for standard nonresponse weighting, the validity of this correction depends on the validity of the estimates $\tilde{\rho}_i$.

We remark that any adjustment technique for nonresponse can be applied to construct estimators for R_ρ, e.g., calibration estimators such as linear or multiplicative weighting (Särndal and Lundström, 2005) or weighting class adjustments (Little, 1986). It is generally known that propensity weighting may lead to larger standard errors. It may, therefore, be more efficient to use parsimonious models to estimate the

R-indicator. For instance, this can be done by stratifying on response propensity classes. However, we did not explore such estimators, and restricted ourselves to the propensity-weighted estimator (3.5). This is a topic for future research.

The estimation of the coefficient of variation (3.3) in the population-based setting is straightforward

$$\mathrm{CV}_{\tilde{\rho}} = \frac{\tilde{S}_{\tilde{\rho}}}{\tilde{\bar{\rho}}_U},$$

where $\tilde{\bar{\rho}}_U = \sum_r d_i / N$.

Despite being straightforward estimators, the population-based R-indicators based on (2.3) and (2.7) are problematic. Their standard errors and biases increase with higher response rates. We will demonstrate this tendency in the evaluation study in Section 4.2. Clearly, more respondents should provide smaller standard errors and reduce bias since the auxiliary variables will not vary as much among the remaining non-respondents. The reason that (2.3) and (2.7) have these properties is that they are natural but naïve estimators that ignore the sampling which causes sample covariances in the denominator of the estimated response propensities to vary along with the numerator. By "plugging" in a fixed population covariance in the denominator, variation from sampling is avoided.

One way to moderate this effect would be to use a composite estimator, i.e., to employ a linear combination of the estimated propensity and the response rate,

$$\tilde{\rho}^C_{i,T1} = (1-\lambda)\,\tilde{\rho}_{i,T1} + \lambda\tilde{\bar{\rho}}_U, \tag{3.7}$$

with $\tilde{\bar{\rho}}_U = \sum_r d_i / N$, and similarly for Type 2. The composite estimate in (3.7) is similar to a "shrinkage" estimator, e.g., Copas (1983 and 1993), for the variance of the response propensities $\tilde{S}^2_{\rho}$ given by (3.6). In that case, the optimal λ is usually chosen to minimize the MSE by solving the derivative of the MSE with respect to λ. We return to the choice of λ in Section 3.4 and note here that, given the observed bias and variance properties, λ should be an increasing function of the response rate and should converge to 1 with higher response rates. Estimated response propensities greater than 1 will be drawn closer to 1 by such a λ due to the use of the linear link function under high response rates.

We explored several other possible alternatives to the composite estimator in (3.7), for example, a composite estimator of the population covariance matrix and the response covariance matrix of the $\mathbf{x}_i$, and response propensities truncated to the interval [0, 1] for high response rates, but this gave worse results compared to the composite estimator in (3.7). In addition, we also investigated a Hájek-type estimate but this gave similar results to those provided by the proposed estimator in (3.6). Another advantage to using the composite estimator in (3.7) is that we can easily construct bias adjustments of the R-indicators similar to the bias adjustments constructed based on the propensities in (2.3) or (2.7).

A promising alternative may be to adopt an EM-algorithm approach in which the missing auxiliary variables for nonrespondents are imputed. Such an approach is, however, very different in nature and we leave this to future research.

3.4 Bias and standard error of the population-based R-indicators

Shlomo et al. (2012) derive analytic approximations for the bias and standard errors of the sample-based estimate of the R-indicator (3.4). The bias in this estimator arises mostly from "plugging in" estimated response propensities in the sample variances. This source of bias is referred to as small sample bias. A much smaller and usually negligible contribution to the bias originates from using sample means rather than population means. Even if the response is representative, i.e., has equal response propensities, some variation in estimated response propensities is found. The bias is inversely proportional to the sample size meaning that the larger the sample, the smaller the bias. Schouten et al. (2009) investigate the bias for different sample sizes. From their analyses, it follows that the bias is relatively small for typical sample sizes used in large-scale surveys in comparison to the standard error of the R-indicators. Also, the bias adjustment is successful in removing the bias.

For the estimated population-based R-indicators, we expect that statistical properties will be quite different from their sample-based counterparts. As these estimators use less information, the standard errors will be larger. The bias of the population-based estimators may also be larger since in addition to the bias that was evident for small sample sizes in the sample-based estimators, the population-based estimators will likely have bias arising from the estimation of the sample means and covariances and from the restriction to (propensity-weighted) response means.

To reduce the bias of the population-based estimators, we propose to adjust $\tilde{S}^2_{\tilde{\rho}_{T1}}$ and $\tilde{S}^2_{\tilde{\rho}_{T2}}$ for bias. This leads to the adjusted version of the estimator for the R-indicator under Type 1 information:

$$\tilde{R}^{\text{ADJ}}_{\tilde{\rho}_{T1}} = 1 - 2\left[\tilde{S}^2_{\tilde{\rho}_{T1}} - \tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)\right]^{1/2}. \tag{3.8}$$

Appendix A derives the general expression for $\tilde{B}_{\bar{\rho}T1}\left(\tilde{S}^2_{\bar{\rho}T1}\right)$ under both simple random sampling and a more general expression under complex sampling. From Appendix A, the response-set based estimator for the bias under simple random sampling is:

$$\tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n} - \frac{n_r}{n^2}\right], \tag{3.9}$$

where n_r denotes the size of the response set r.

In the case of Type 2 information, the adjusted version of the estimator for the R-indicator is as (3.8) with the Type 2 terms replacing the Type 1 information.

Appendix B derives the general expression for the bias of $\tilde{S}^2_{\tilde{\rho}_{T2}}$, $\tilde{B}_{\bar{\rho}T2}\left(\tilde{S}^2_{\bar{\rho}T2}\right)$, under simple random sampling and the more general case of complex sampling. From Appendix B, the response-set based estimator for the bias under simple random sampling is:

$$\tilde{B}^{\mathrm{SRS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^{2}_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\left\{\frac{1}{n_r^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\right.$$
$$-\frac{N}{nn_r}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{z}_i\mathbf{z}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}$$
$$+\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{x}_i$$
$$\left.+\frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T2}-\left(1-\frac{n}{N}\right)\frac{\tilde{S}^{2}_{\tilde{\rho}_{T2}}}{n}-\frac{n_r}{n^2}\right\},$$

where $\hat{\mathbf{F}} = Nn^{-1}\sum_r \mathbf{z}_k\mathbf{z}_k^T$, $\hat{\mathbf{t}} = Nn^{-1}\sum_r \mathbf{x}_k$, and $\mathbf{z}_i = (\mathbf{x}_i - \bar{\mathbf{x}}_U)$.

Turning to the composite estimator, it is straightforward to show that (3.7) can be rewritten as

$$\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}} = (1-\lambda)\,\tilde{S}^{2}_{\tilde{\rho}_{T1}}, \tag{3.10}$$

and its bias equals

$$B\left(\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}\right) = (1-\lambda)\,B\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right) - \lambda S^{2}_{\rho}. \tag{3.11}$$

A response-set based estimator for $B\left(\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}\right)$ is obtained using the response-set based estimator developed for $B\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right)$. For the Type 1 estimator and under simple random sampling:

$$\tilde{B}^{\mathrm{SRS}}_{\tilde{\rho}^{C}_{T1}}\left(\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}\right) = (1-\lambda)\,\tilde{B}^{\mathrm{SRS}}_{\tilde{\rho}^{C}_{T1}}\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right) - \lambda\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}$$
$$= (1-\lambda)\frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}^{C}_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i\right.$$
$$\left.+\frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}^{C}_{i,T1}-\left(1-\frac{n}{N}\right)\frac{\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}}{n}-\frac{n_r}{n^2}\right]-\lambda\tilde{S}^{2}_{\tilde{\rho}^{C}_{T1}}. \tag{3.12}$$

The same approach applies for Type 2 estimator.

The variance of (3.10) is equal to

$$V\left(\tilde{S}^{2}_{\tilde{\rho}^{C}_{T_1}}\right) = (1-\lambda)^2\,V\left(\tilde{S}^{2}_{\tilde{\rho}_{T_1}}\right). \tag{3.13}$$

To estimate the variance of $\tilde{R}^{\mathrm{ADJ}}_{\tilde{\rho}T1}$ in (3.8) as well as the variance of the composite estimator in (3.13) we need to estimate the variance of $\tilde{S}^{2}_{\tilde{\rho}_{T_1}}$ defined in (3.6) and denoted by $V\left(\tilde{S}^{2}_{\tilde{\rho}_{T_1}}\right)$. To estimate this variance we use resampling methods. More specifically, we employ bootstrap methods (see: Efron and Tibshirani, 1993; Booth, Butler and Hall, 1994 and Wolter, 2007 for the use of bootstrapping methods for finite populations) and assess their performance in the evaluation study in Section 4.

We return now to the choice of λ for the composite estimator in (3.7). The optimal λ can be derived by combining (3.11) and (3.13), and then taking derivatives. Letting B and V denote $B\left(\tilde{S}^{2}_{\tilde{\rho}_{T_1}}\right)$ and $V\left(\tilde{S}^{2}_{\tilde{\rho}_{T_1}}\right)$, respectively, it follows that the optimal λ is

$$\lambda_{\text{opt}} = \frac{B\left(B + S_{\rho}^{2}\right) + V}{\left(B + S_{\rho}^{2}\right)^{2} + V}. \tag{3.14}$$

We note that as the sample size increases, both the B and V terms tend to zero and it is possible that λ_{opt} might be negative. However, based on the evaluation study for the types of large-scale national surveys under consideration, this problem does not arise in practice.

In order to estimate λ_{opt}, the quantities B, V and S_{ρ}^{2} need to be estimated. Under Type 1 information and simple random sampling, we propose to estimate B by $\tilde{B}_{\tilde{\rho}_{T1}}^{\text{SRS}}\left(\tilde{S}_{\tilde{\rho}_{T1}}^{2}\right)$ as in (3.9), S_{ρ}^{2} by $\tilde{S}_{\tilde{\rho}_{T1}}^{2}$, and V by the bootstrap variance estimator of $\tilde{S}_{\tilde{\rho}_{T1}}^{2}$. This leads to the population-based Type 1 estimator for λ_{opt}, denoted by $\tilde{\lambda}_{\text{opt},\,T1}$, and the population-based composite propensities

$$\tilde{\rho}_{i,T1}^{\text{PC}} = \left(1 - \tilde{\lambda}_{\text{opt},T1}\right)\tilde{\rho}_{i,T1} + \tilde{\lambda}_{\text{opt},T1}\tilde{\bar{\rho}}_{U}.$$

The corresponding population-based R-indicator is then computed as in (3.5) and its bias-adjusted version as in (3.8), where the bias adjustment is given by (3.12).

We propose to estimate the variance of the population-based composite estimator by linearization

$$\frac{\tilde{V}^{\text{BT}}\left(\tilde{S}_{\tilde{\rho}_{T1}}^{2}\right)\left(1 - \tilde{\lambda}_{\text{opt},T1}\right)^{2}}{\tilde{S}_{\tilde{\rho}_{T1}}^{2}},$$

where $\tilde{V}^{\text{BT}}\left(\tilde{S}_{\tilde{\rho}_{T1}}^{2}\right)$ is the bootstrap variance estimator for $V\left(\tilde{S}_{\tilde{\rho}_{T1}}^{2}\right)$.

The same approach applies for Type 2 information.

4 Evaluation study

In this section, we carry out an evaluation study on real census data from the 1995 Israel Census Sample to assess the sampling properties of the estimation procedures introduced in Section 3.

The aim of the evaluation is two-fold: a) to study the sampling properties of the unadjusted and bias adjusted population-based R-indicators, comparing them to those of their sample-based counterpart and assessing the effect of sample size, number of auxiliary variables in the model, and response rate; b) to investigate the performance of the bootstrap estimator for estimating the variance of the population-based R-indicator.

4.1 Data and design of evaluation study

The 1995 20% Israel Census Sample contains 753,711 individuals aged 15 and over in 322,411 households. The census sample design is a random systematic sample where every fifth household was delivered a long questionnaire covering a range of socio-economic questions. The sample units are households and all persons over the age of 15 in the sampled households are interviewed. Typically a proxy

questionnaire is used and therefore there is no individual nonresponse within the household. In this study, we assume that every household has an equal probability to be included in the sample. This evaluation study uses data at the household level $(N = 322{,}411)$.

We carried out a two-step design to define response propensities in the population (census) data. This procedure ensures that we have a known model generating the response propensities. Moreover, in order to explore the effect of varying response rates and the number of auxiliary variables in the model on the performance of the estimators, we considered six scenarios defined by the level of response rates (3 categories) and the type of model (2 categories).

A. First, probabilities of response were defined according to variables: Type of locality (4 categories defined by rural/urban and type of population), number of persons in household grouped to 3 categories (1-2, 3-5, 6+), children in the household indicator (yes, no), region (7 categories dividing the country from north to south), and density (3 categories: less than 1.5, 1.5-3.0, greater than 3.0). These variables define groups that are known to have differential response rates for social surveys in practice. To study the effect of response rates on the performance of the estimators, probabilities of response p were defined according to $p = p_1 p_2 p_3 p_4 p_5 + \alpha$ with three choices $\alpha = 0.15$ (RR1), $\alpha = 0.55$ (RR2), and $\alpha = 0.75$ (RR3), where the probabilities p_1, p_2, p_3, p_4, p_5 are given in Table 4.1. We generate three response indicator variables using the Bernoulli distribution for each of the response scenarios defined under RR1, RR2, and RR3.

B. For each of the response scenarios from step (A), we use the response indicator as the dependent variable and fit both a linear and a logistic regression model to the population to predict "true" response propensities for our evaluation study under both link functions. Two different models were considered for prediction of "true" response propensities. In Model 1, independent variables are exactly the explanatory variables used in step A for the definition of response probabilities (child indicator, number of persons in the household, region, type of locality, density). In Model 2, independent variables are type of locality, number of persons in household, child indicator. Notice that we use the same response indicator variables to fit the two models. This allows the effect of the model to be isolated, excluding differences due to random variability in the response indicator.

Response rates for the variables defining probabilities as well as the overall response rates and true population values of the R-indicator under the two models are shown in Table 4.1. For comparison purposes, we report population values of the R-indicator based on both linear and logistic regression models where the response rates range between 25.1% and 35.1% under RR1, between 64.7% and 75.4% under RR2 and between 84.7% and 94.6% under RR3. RR2 represents the type of response rate seen in large-scale national social surveys. As can be seen in Table 4.1, there is little difference in the population values of the R-indicators based on the linear and logistic link function for RR1 and RR2 and a slight difference for RR3

under both models where response rates are in the upper tail of the distribution. We also note that across the very different overall response rates, the population values of the R-indicator are generally high.

Table 4.1
Probabilities of response and percent response generated in the evaluation population dataset according to auxiliary variables

			Percentage response		
Variable	**Category**	**Probability of response**	**RR1**	**RR2**	**RR3**
Children in Household	None	0.6	25.7	65.6	85.7
	1+	0.8	35.1	75.4	94.6
Number of Persons in Household	1-2	0.5	24.6	64.5	84.7
	3-5	0.8	32.9	72.8	92.5
	6+	0.7	29.9	70.3	90.0
Type of Locality	Type 1	0.6	25.1	64.9	85.0
	Type 2	0.7	28.3	68.5	88.4
	Type 3	0.8	31.5	71.7	91.2
	Type 4	0.75	28.9	69.2	88.9
Region	1	0.6	25.1	65.1	84.7
	2	0.8	31.2	71.5	91.0
	3	0.7	28.1	67.6	87.8
	4	0.6	26.7	66.5	86.4
	5	0.6	24.8	64.7	84.9
	6	0.7	27.6	67.8	88.0
	7	0.8	30.3	70.4	90.9
Density	<=1.5	0.6	26.1	66.0	86.2
	1.5-3.0	0.8	28.9	68.9	88.8
	>3	0.7	24.7	64.7	84.7
Overall response rate			27.1	67.0	87.0
“True” Population R-indicator (logistic)	Model 1		0.9031	0.9005	0.9063
	Model 2		0.9103	0.9074	0.9137
“True” Population R-indicator (linear)	Model 1		0.9033	0.9006	0.9076
	Model 2		0.9104	0.9074	0.9145

When using Model 2, the true R-indicator is always around 0.007 points greater than the corresponding value under Model 1. This is due to the fact that Model 2 for estimating the response propensities is mis-specified. There are fewer auxiliary variables and hence smaller variation in the estimated response propensities which leads to a higher R-indicator. As a consequence we obtain a slightly higher R-indicator for Model 2 as some of the variation is not captured. For this reason, it is always important to report R-indicators together with the auxiliary information used to calculate them since their values depend on the nonresponse model. In addition, we should use covariates that correlate to the survey variables (Schouten et al., 2012).

For each response scenario, five hundred samples were drawn from the population under simple random sampling (SRS) at three different sampling rates 1% $(n = 3{,}224)$, 2% $(n = 6{,}448)$ and 4% $(n = 12{,}896)$. For each sample drawn, a sample response indicator was generated from the "true" population response probability based on the logistic link function. This determines the response set r. Response propensities and R-indicators were then estimated from each sample for both sample and population-based auxiliary variables. Response propensities are estimated in the sample using the "true" model (either Model 1 or Model 2, depending on the scenario).

In order to estimate the variance of population-based estimators, we employ a non-parametric bootstrap algorithm. From each response set, we drew $B = 500$ bootstrap samples using simple random sampling (SRS) with replacement. Subsequently, nonresponse was generated in the bootstrap sample by copying the 0-1 sample response indicator values. A replicate of the estimator was computed over each bootstrap sample.

4.2 Results

Table 4.2 presents results of the evaluation study for each response rate scenario, type of model and each sampling rate. We contrast the sample-based R-indicators (under both link functions to highlight any differences) with the population-based R-indicators. In the evaluation, we also investigate the performance of the population-based composite estimator (PC) as shown in (3.7).

For each estimator, Table 4.2 shows: a) the percentage Relative Bias (%RB) calculated as $100\left\{\left[\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j} - R_\rho\right)\Big/R_\rho\right]\Big/500\right\}$, where $\hat{R}_{\hat{\rho}j}$ is the value of the estimator computed for the j^{th} sample and R_ρ is the true R-indicator based on the linear regression model (from Table 4.1), and similarly for $\tilde{R}_{\tilde{\rho}_{T1}}$, $\tilde{R}_{\tilde{\rho}_{T2}}$, and the composite estimator; b) the Relative Root Mean Square Error (RRMSE) calculated as

$$100\left\{R_\rho^{-1}\sqrt{\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j} - R_\rho\right)^2\Big/500}\right\}.$$

Table 4.2 shows that differences between the sample-based estimators computed using the linear and the logistic link functions are very small in general, except when the response rates get very close to 1 (RR3).

For sample-based and population-based Type 1 and Type 2 estimators there is a general downward bias in the unadjusted R-indicators and this tends to decrease as the sample size increases for both Models 1 and 2. This is as expected. Sampling error tends to lead to overestimation of the variability of the estimated response propensities and this leads to underestimation of the R-indicator. The degree of underestimation is generally larger for population-based estimators than for the sample-based estimators, especially for higher response rates. The variation of response propensities is larger in this case than the variation under sample-based auxiliary variables. In addition, the RRMSE of the estimators decreases as sample size increases and is generally larger for population-based estimators. Thus, the population-based R-indicators are in general less accurate than their sample-based counterparts and allow for weaker conclusions regarding the nature of response.

Table 4.2
Properties of the estimated R-indicators for sample and population-based auxiliary variables for 500 samples in the evaluation study

Response Rate	Sample Rate	Estimator	Model 1				Model 2			
			Unadjusted		Adjusted		Unadjusted		Adjusted	
			%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
RR1	1%	Sample-based (log)	-1.73	2.39	0.32	2.01	-0.77	1.88	0.34	1.96
		Sample-based (lin)	-1.71	2.37	0.33	2.01	-0.75	1.87	0.35	1.95
		Type 1	-2.32	3.08	0.32	2.54	-1.08	2.32	0.30	2.39
		Type 1 - PC	0.04	2.28	0.22	2.42	0.59	2.44	0.38	2.41
		Type 2	-1.47	2.29	1.06	2.50	-0.20	1.74	1.01	2.27
		Type 2 - PC	0.71	2.11	0.94	2.34	1.19	2.32	1.05	2.28
	2%	Sample-based (log)	-0.90	1.53	0.14	1.36	-0.41	1.30	0.14	1.31
		Sample-based (lin)	-0.89	1.51	0.16	1.36	-0.40	1.29	0.15	1.31
		Type 1	-1.24	1.89	0.12	1.61	-0.51	1.57	0.17	1.59
		Type 1 - PC	0.04	1.56	0.10	1.59	0.38	1.68	0.21	1.62
		Type 2	-0.45	1.30	0.84	1.64	0.26	1.31	0.86	1.63
		Type 2 - PC	0.72	1.53	0.82	1.61	1.02	1.75	0.89	1.66
	4%	Sample-based (log)	-0.48	1.00	0.05	0.93	-0.27	0.90	0.00	0.88
		Sample-based (lin)	-0.46	0.99	0.06	0.92	-0.26	0.89	0.01	0.88
		Type 1	-0.63	1.23	0.05	1.12	-0.34	1.13	-0.01	1.11
		Type 1 - PC	0.15	1.14	0.07	1.12	0.18	1.18	0.01	1.13
		Type 2	0.12	0.92	0.78	1.25	0.40	1.01	0.69	1.19
		Type 2 - PC	0.83	1.29	0.79	1.26	0.83	1.30	0.70	1.20
RR2	1%	Sample-based (log)	-1.81	2.44	0.33	2.01	-0.76	1.83	0.34	1.94
		Sample-based (lin)	-1.79	2.42	0.34	2.01	-0.75	1.82	0.35	1.94
		Type 1	-5.17	5.95	-0.01	3.95	-2.45	3.77	0.25	3.43
		Type 1 - PC	-1.50	3.58	-0.47	3.69	0.69	3.37	0.49	3.46
		Type 2	-4.76	5.50	0.27	3.75	-1.95	3.29	0.58	3.23
		Type 2 - PC	-1.13	3.28	-0.12	3.51	0.74	3.13	0.71	3.25
	2%	Sample-based (log)	-1.00	1.59	0.08	1.37	-0.40	1.29	0.14	1.30
		Sample-based (lin)	-0.98	1.57	0.09	1.36	-0.40	1.28	0.14	1.30
		Type 1	-2.89	3.55	0.07	2.59	-1.19	2.58	0.37	2.72
		Type 1 - PC	-0.57	2.37	-0.12	2.49	0.53	2.67	0.41	2.69
		Type 2	-2.52	3.19	0.39	2.50	-0.79	2.28	0.69	2.63
		Type 2 - PC	-0.26	2.19	0.19	2.37	0.81	2.58	0.71	2.60
	4%	Sample-based (log)	-0.48	0.98	0.07	0.90	-0.16	0.81	0.12	0.83
		Sample-based (lin)	-0.46	0.97	0.08	0.90	-0.15	0.81	0.12	0.82
		Type 1	-1.42	2.12	0.13	1.81	-0.60	1.66	0.16	1.67
		Type 1 - PC	0.16	1.77	0.14	1.80	0.37	1.76	0.20	1.69
		Type 2	-1.07	1.82	0.46	1.78	-0.25	1.47	0.47	1.63
		Type 2 - PC	0.45	1.72	0.46	1.75	0.65	1.73	0.50	1.66
RR3	1%	Sample-based (log)	-1.07	1.59	0.10	1.30	-0.52	1.21	0.02	1.16
		Sample-based (lin)	-0.85	1.40	0.24	1.26	-0.41	1.13	0.10	1.13
		Type 1	-6.60	7.32	-0.76	4.24	-3.20	4.61	0.06	4.12
		Type 1 - PC	-2.22	4.15	-0.88	4.16	-0.28	3.70	0.09	3.92
		Type 2	-6.29	6.99	-0.53	4.08	-2.85	4.25	0.27	3.95
		Type 2 - PC	-2.12	3.97	-0.67	4.02	-0.04	3.52	0.33	3.78
	2%	Sample-based (log)	-0.73	1.13	-0.14	0.92	-0.30	0.88	-0.03	0.85
		Sample-based (lin)	-0.54	0.98	0.01	0.87	-0.20	0.82	0.06	0.82
		Type 1	-3.70	4.31	0.12	2.93	-1.74	2.98	0.20	2.86
		Type 1 - PC	-0.78	2.60	-0.15	2.78	0.42	2.81	0.36	2.94
		Type 2	-3.46	4.07	0.30	2.87	-1.46	2.73	0.41	2.77
		Type 2 - PC	-0.61	2.47	0.02	2.70	0.64	2.74	0.57	2.87
	4%	Sample-based (log)	-0.46	0.77	-0.16	0.66	-0.18	0.57	-0.05	0.55
		Sample-based (lin)	-0.29	0.66	-0.01	0.61	-0.09	0.53	0.04	0.53
		Type 1	-1.96	2.62	0.12	2.12	-0.89	1.81	0.13	1.76
		Type 1 - PC	-0.03	1.97	0.07	2.06	0.38	1.84	0.19	1.79
		Type 2	-1.74	2.42	0.31	2.07	-0.66	1.65	0.31	1.71
		Type 2 - PC	0.11	1.89	0.25	2.00	0.56	1.81	0.38	1.75

In general, the unadjusted population-based composite estimators have a better performance than the corresponding unadjusted population-based estimators, both in terms of %RB and RRMSE, especially for higher response rates. They still show some degree of overestimation under the correct Model 1 for low response rates and underestimation for high response rates. However, for Model 2 we see overestimation.

We now turn to the bias-adjusted estimated R-indicators in Table 4.2. For Type 1, the bias adjustment is able to remove the bias. The analytical bias adjustment for Type 1 population-based estimator works well and generally outperforms the analytical bias adjustment for Type 2 population-based estimates. It seems to pick up most of the bias and provides adjusted estimates that are closer to sample-based R-indicators. The RRMSE for the bias-adjusted estimator is generally similar to the corresponding RRMSE for the unadjusted estimator, meaning that the increase in variability is compensated by the bias reduction. For higher response rates, the adjusted population-based composite estimate reduces the bias and RRMSE of their corresponding population based R-indicators.

In unadjusted form, the Type 2 R-indicator behaves better than the Type 1 R-indicator. This is rather surprising as we seem to be able to have more accurate estimation of the true R-indicator when using less information. The reason for this is that for the Type 1 estimator we do not include any of the sampling variation when we "plug in" the population covariance matrix, whilst for the Type 2 estimator we use only the marginal information and "plug in" the response covariance matrix which accounts for more of the sampling variation. After the bias adjustment, the Type 2 estimators have higher %RB (especially for lower response rates) but similar RRMSE. Type 2 bias adjustment performs worse than the bias adjustment for Type 1 and overcompensates for the bias. This result was expected as the Type 2 bias adjustment is based on a linear approximation, while Type 1 bias adjustment is computed exactly.

Regarding increasing response rates, surprisingly, for the population-based unadjusted estimators, we observe a better performance for lower response rates, both in terms of percentage relative bias (%RB) and RRMSE. The RRMSE of RR3 are 2 to 3 times larger than for RR1. Analytical bias adjustments work very well under all response rates, although with higher RRMSEs for higher response rates. These RRMSEs are reduced by the use of the composite estimators.

Regarding the effect of the number of variables in the model, a lower %RB and RRMSE are observed under Model 2 for unadjusted population-based estimators compared to Model 1. The composite estimators show in general an opposite pattern. The bias-adjusted versions show similar performance under the two models.

Table 4.3 shows the mean of the estimated λ_{opt} for the composite population-based Type 1 and Type 2 estimators compared to the true value obtained from the population under the two extreme response rate scenarios, RR1 and RR3. It can be seen that the mean estimated λ_{opt} does not deviate greatly from their true values in the evaluation study.

Table 4.3
Mean λ_{opt} for population-based auxiliary variables for 500 samples in the evaluation study

Response Rate	Sample Rate	Model 1				Model 2			
		Type 1		Type 2		Type 1		Type 2	
		True	Pop-based	True	Pop-based	True	Pop-based	True	Pop-based
RR1	1%	0.40	0.33	0.36	0.33	0.31	0.29	0.26	0.28
	2%	0.25	0.21	0.22	0.21	0.19	0.22	0.15	0.19
	4%	0.14	0.13	0.13	0.13	0.10	0.10	0.08	0.09
RR3	1%	0.68	0.44	0.67	0.44	0.57	0.51	0.55	0.48
	2%	0.51	0.39	0.50	0.38	0.41	0.43	0.39	0.41
	4%	0.35	0.27	0.34	0.27	0.25	0.23	0.24	0.22

Table 4.4 analyses the performance of the bootstrap estimators for estimating the variance of population-based R-indicators under the two extreme response rate scenarios, RR1 and RR3. Analytical expressions for the variance of sample-based R-indicators have been developed and used in the evaluation study (see Shlomo et al., 2012). Simulation means of the variance estimators are compared in Table 4.4 with the simulation variances (calculated across the replicated samples), using percentage relative bias. The table also includes the Coverage Rate defined as the percentage of times that the true R_ρ is contained in the confidence interval $100\left\{\left[\sum_{j=1}^{500} I\left(R_\rho \in \hat{R}_{\hat{\rho}j} \pm 1.96\sqrt{\hat{V}_j\left(\hat{R}_{\hat{\rho}j}\right)}\right)\right] \Big/ 500\right\}$, where $\hat{V}_j\left(\hat{R}_{\hat{\rho}j}\right)$ is the estimated variance for the j^{th} sample (linearization variance estimator for sample-based estimator and bootstrap variance estimator for population-based estimators) and I is the indicator function. The bootstrap variance estimators for population-based estimators work well. The sample-based estimator show better coverage than the corresponding population-based versions. Type 1 and Type 2 population-based estimators have similar coverages. The coverage always improves as the sample size gets larger.

The behaviour under different response rates is mixed. There seems to be an interaction between sample size and response rate. The number of variables in the model does not have a large impact on coverage. However, we observe problems with coverage for the population-based estimators under the highest response rate (RR3), especially for the 1% sample rate.

Figures 4.1, 4.2 and 4.3 present box plots comparing the estimators and their bias adjusted versions when fitting Model 1, under different response rate scenarios RR1, RR2 and RR3 respectively. The gains from the bias adjustments are evident for Type 1 and Type 2 R-indicators. Standard errors for RR3 are much larger than for RR1 under the same sampling rates. The variability of the bias-adjusted estimator increases and it is larger for smaller sample sizes.

Table 4.4
Properties of variance estimators for R-indicators under sample and population-based auxiliary variables for 500 samples

Response rate	Sampling rate	Estimator	Model 1		Model 2	
			%RB	Coverage	%RB	Coverage
RR1	1%	Sample-based	1.84	0.95	-5.74	0.95
		Type 1	4.35	0.95	11.12	0.96
		Type 2	4.99	0.94	7.72	0.95
	2%	Sample-based	1.43	0.96	1.15	0.95
		Type 1	8.62	0.96	5.31	0.95
		Type 2	7.03	0.93	2.10	0.92
	4%	Sample-based	7.93	0.97	-4.58	0.95
		Type 1	13.23	0.96	3.42	0.95
		Type 2	13.38	0.89	2.53	0.90
RR3	1%	Sample-based	-1.05	0.95	-9.48	0.92
		Type 1	2.87	0.78	11.47	0.86
		Type 2	4.97	0.78	10.26	0.85
	2%	Sample-based	-4.34	0.94	-7.96	0.94
		Type 1	-7.61	0.92	2.37	0.91
		Type 2	-8.07	0.92	1.02	0.90
	4%	Sample-based	3.31	0.94	-3.54	0.95
		Type 1	-8.33	0.93	12.32	0.96
		Type 2	-8.13	0.93	10.89	0.96

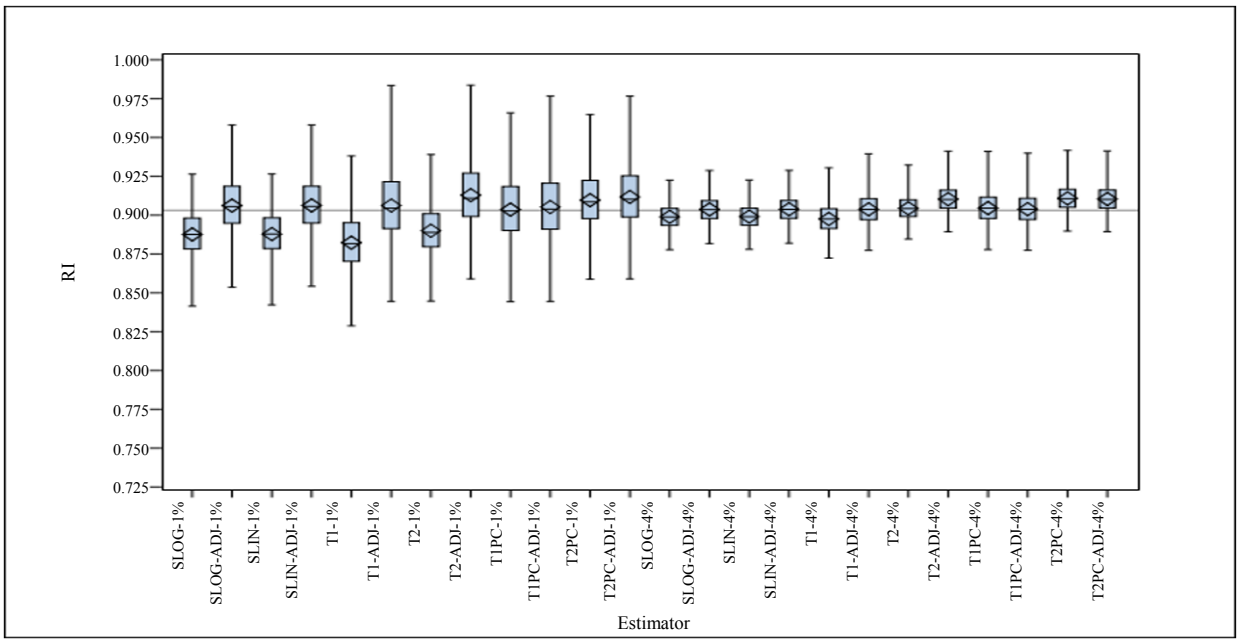


Figure 4.1 Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR1. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

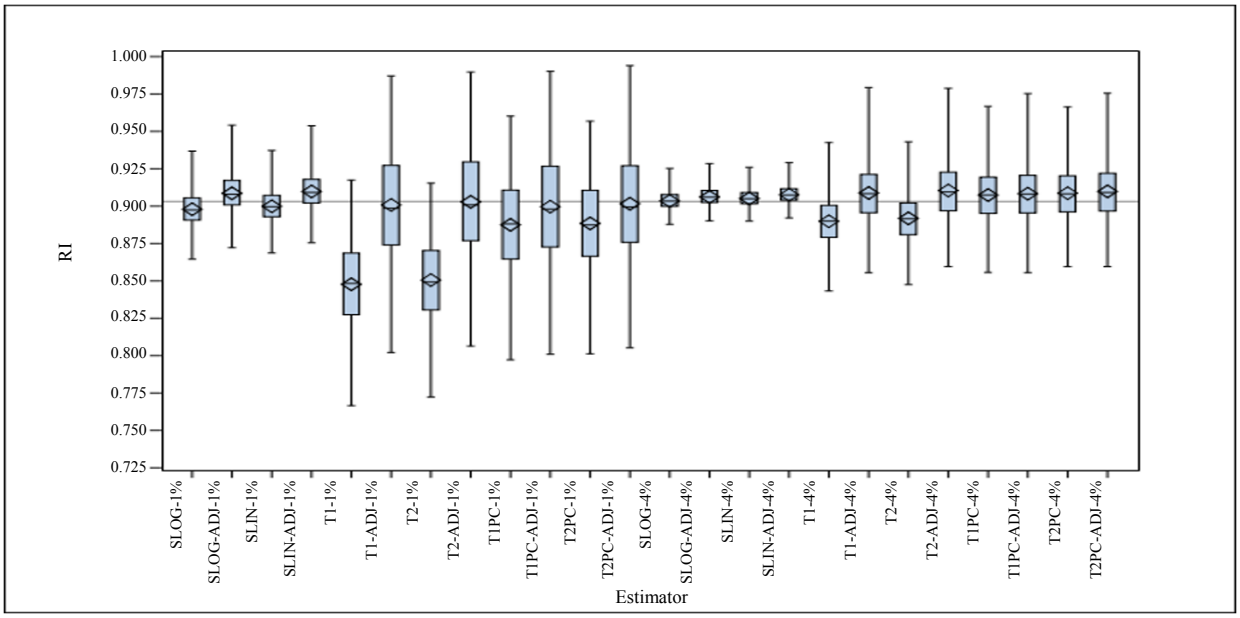


Figure 4.2 Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR2. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

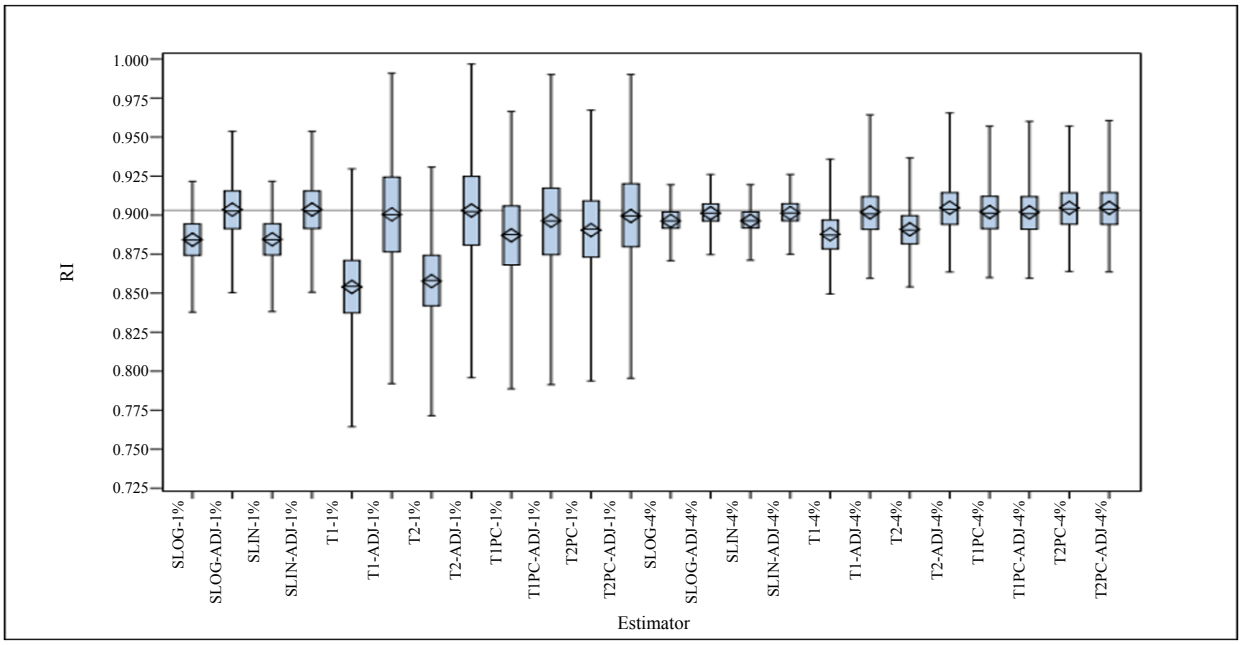


Figure 4.3 Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR3. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

5 Application to the Dutch Health Survey

In this section, we apply the population-based Type 1 and Type 2 estimators to the Dutch Health Survey conducted by Statistics Netherlands. We employ three auxiliary variables that are part of the gold standard for Dutch market research companies and compare population-based performance to sample-based performance.

The Dutch Health Survey (HS) was commissioned in 1998 as a repeated cross-sectional survey among the full population registered in the Dutch Population Register, but excluding the institutionalized population. It uses a two-stage, self-weighting sampling design in which the first stage is formed by municipalities and the second stage by persons living in the selected municipalities. Until 2012, the HS was a face-to-face survey. In 2012, it changed to a mixed-mode design involving online and face-to-face interviews. Over the years, the sample size was reduced considerably from around 35,000 to around 18,000. We use the 2002 HS data, one of the last years with the original sample size. The net sample size is 33,584 persons.

To calibrate national and regional samples, Dutch market research companies use the so-called Gold Standard population statistics produced by Statistics Netherlands (MOA, 2015). The Gold Standard is an explicitly defined set of auxiliary variables that affiliated companies include in their survey questionnaires. Three of these variables are age, gender and marital status. We focus on these three in the application.

Table 5.1 contains the HS sample and response distributions, and the Statistics Netherlands' population distributions for the three variables. Joint population distributions, needed to estimate the Type 1 population-based covariance matrices, are also available, but not given here. In practice, the sample distribution is, of course, unknown. The three variables show a different picture: for age and marital status, the response distribution is closer to the sample distribution than to the population distribution, and population-based response propensities give more variation. For gender, the population distribution is closer to the response distribution and less variation is found.

Table 5.1
Age, gender, and marital status distributions for the sample, respondents, and population

Variables	Categories	Respondents	Sample	Population
Age	20-24	7.5	7.9	8.1
	25-29	7.3	8.2	8.9
	30-34	9.9	10.2	10.9
	35-39	10.9	10.8	11
	40-44	10.3	10.3	10.4
	45-49	9.7	9.4	9.6
	50-54	9.4	9.6	9.5
	55-59	8.8	8.9	8
	60-64	7.1	6.7	6.3
	65-69	5.9	5.6	5.4
	70-74	5.4	4.7	4.6
	75+	7.7	7.8	7.2
Gender	Male	48.9	49.8	49.2
	Female	51.1	50.2	50.8
Marital status	Not married	23.7	26.8	26.9
	Married	63.3	59.3	58.8
	Widowed	6.5	6.7	6.7
	Divorced	6.4	7.2	7.6

We estimate Type 1 and Type 2 population-based R-indicators in Table 5.3. For the composite estimator, we used the estimated smoothing parameter $\tilde{\lambda}_{\text{opt}}$ based on the population-based response propensities. We also include an estimate for λ_{opt} calculated using sample-based response propensities. The latter cannot normally be computed and is included for comparison only. Table 5.2 contains the estimated smoothing parameter $\tilde{\lambda}_{\text{opt}}$ based on both the population-based response propensities and the sample-based response propensities. The sample-based $\tilde{\lambda}_{\text{opt}}$ are larger and tend to have a stronger smoothing effect. However, all $\tilde{\lambda}_{\text{opt}}$ are relatively small.

Table 5.3 contains the various population-based R-indicators. For comparison, the sample-based R-indicator is also provided where we used the logistic link function. The linear link function produced the same result. We can conclude that the population-based R-indicators, using only response and population distributions, are different from the sample-based R-indicators, using response and sample distributions. This difference increases, as expected, when Type 2 indicators are used. The composite estimators perform slightly better than the non-composite estimators, but there is still a considerable difference. This is not due

to a biased smoothing parameter, as the difference is only modestly smaller when sample-based propensities are used to estimate the smoothing parameter. Furthermore, after bias adjustment, the differences between the composite estimators for sample-based and population-based propensities vanish.

Table 5.2
Values for smoothing parameter λ_{opt} based on population-based response propensities and on sample-based response propensities for Type 1 and 2 composite estimators

	Smoothing parameter $\tilde{\lambda}_{opt}$	
	Type 1	**Type 2**
Population-based response propensities	0.043	0.038
Sample-based response propensities	0.076	0.095

Table 5.3
Unadjusted and bias-adjusted sample-based and Type 1 and Type 2 population-based R-indicators for the HS 2002 data. The population-based composite R-indicators are based on the smoothing parameter λ_{opt} using population-based and sample-based response propensities. 95% confidence intervals (CI) by normal approximation are provided

	Unadjusted			Bias-adjusted		
Estimator	**R-indicator**	**95% CI**		**R-indicator**	**95% CI**	
Sample-based	0.899	0.888	0.909	0.901	0.890	0.912
Type 1 – original	0.876	0.860	0.891	0.879	0.864	0.895
Type 1 – composite population-based	0.880	0.865	0.896	0.880	0.864	0.895
Type 1 – composite sample-based	0.883	0.868	0.898	0.880	0.865	0.895
Type 2 – original	0.873	0.858	0.889	0.877	0.861	0.894
Type 2 – composite population-based	0.878	0.863	0.894	0.878	0.862	0.893
Type 2 – composite sample-based	0.881	0.866	0.897	0.878	0.863	0.893

A conclusion from the application is that the lower population-based R-indicators result from the large differences between sample and population distributions of the auxiliary variables. For a sample size of 33,584 persons, a test of the differences between sample and population distributions is significant for all three variables at the 5% level. The available Dutch Health Survey net sample does not contain sampling units with frame and/or other administrative errors as well as out-of-scope populations such as institutionalized persons. This modification plus some additional, small tailoring to interviewer workloads, most likely caused sample distributions to differ from the original population counts. This points at the "Achilles heel" of population-based R-indicators: it is imperative that there is no disparity between definitions and populations.

6 Discussion

The extension of sample-based to population-based estimators of R-indicators is comprised of two steps: 1) the estimation of response propensities, and 2) the estimation of the R-indicators based on these

propensities. The population-based estimation of response propensities is straightforward when linear models are assumed for response propensities and response influences. The linear link function is reasonable when estimating response propensities under typical response rates seen for large-scale national social surveys as shown in the evaluation study in Section 4. The sample-based estimators contain sample covariance matrices and sample frequencies that can be replaced by population covariance matrices or population frequencies. We identified two types of settings: when population cross-products are available or when auxiliary information is restricted to marginal population counts only. We labelled the corresponding estimators as Type 1 and Type 2 estimators, respectively. The Type 2 setting is more restrictive than the Type 1 setting.

Following the estimation of population-based response propensities, we have constructed population-based estimators for the R-indicator and examined their properties both theoretically and empirically. The estimators are applied to samples drawn from real data from the 1995 Israel Census Data where "true" propensities were calculated according to realistic assumptions for national household social surveys. Thus, we have addressed the first two research questions at the beginning of the paper: How to extend sample-based response propensities and R-indicators to population-based response propensities and R-indicators? and What are the statistical properties of population-based R-indicators?

There are many options for the estimation of R-indicators depending on the response to the survey. We used propensity weighted response means as the propensities are available. However, any calibration method can be used such as linear weighting or adjustment classes. In fact, the set of auxiliary variables used for the estimation of the R-indicators may be a subset of the auxiliary variables used for the estimation of propensities and influences. Parsimonious models may prove to be more efficient as it is known that propensity-weighting may seriously affect the precision of the estimators. This is a topic for future research.

The two properties we examined are the bias and standard errors of the proposed population-based R-indicators. As expected the bias and standard errors are dependent on the size of the sample and the type of auxiliary information available where the smaller the sample, the larger the bias and the standard error. When samples are smaller, it becomes more difficult to distinguish sampling variation from response variation. Clearly, the confidence intervals become larger as there is less information in small samples.

The bias-adjusted Type 1 estimators (population cross-products) perform better than the bias-adjusted Type 2 estimators (population marginal counts). This is as expected given that they employ more information. However, the unadjusted Type 2 estimators have better RRMSE properties than the unadjusted Type 1 estimators. This is a surprising result and points to a suboptimal use of the population cross-products when they are used as "plug-ins" and do not account for any sampling variation. The standard errors of the population-based estimators are larger than their sample-based counterparts.

The evaluation study in scenario RR3 shows that, for very high response rates, the population-based R-indicators provide higher standard errors and larger bias, mainly due to propensities being estimated outside of the interval [0, 1]. For this reason, we proposed a composite estimator with varying smoothing parameters dependent on the response rate. Standard errors were reduced but at the cost of increased bias.

From the analyses it becomes apparent that the bias of the Type 1 and Type 2 estimators depends on the number of auxiliary variables, but this dependence was modest in our evaluations. The bias may increase when using detailed models with many variables for the estimation of response propensities. The rationale behind this is that detailed models allow for more sampling variation to be picked up as bias.

The population-based R-indicators have a number of caveats:

Firstly, the choice of auxiliary information that is available at a national level may be more limiting than sample-based auxiliary information depending on the availability of registers and administrative data. The selection of auxiliary variables should depend on whether they are correlated with the survey target variables. Also, it is strongly recommended that population statistics that are based on registers or administrative data are used rather than those based on weighted survey counts from other surveys since these statistics may not reflect the true population distribution accurately. One would draw erroneous conclusions about the representativeness of the response if the population estimates are biased.

Secondly, we make the assumption that the survey measures the same quantities as in the population information and we do not investigate the effect of possible departures from this assumption. However, we note that there is an imminent risk of measurement errors when comparing the representativeness of survey questions to population statistics. It must be ascertained that the survey questions that are employed have the same definitions and classifications as the population tables. Hence, it is best to avoid questions that are prone to measurement errors, such as questions that require a strong cognitive effort or that may lead to socially desirable answers.

Thirdly, in settings where only population information is available, options to improve representativeness during data collection are much more limited since there is no individual auxiliary information available for the nonrespondents. Nonetheless, in these settings, assessments of representativeness may still be useful in the design of advance and reminder letters, in interviewer training and in paradata collection.

Finally, we do not consider hybrid settings where the R-indicator is based on both linked data and population tables. In addition, we do not deal with the case where we could use weighted survey estimates if there is no aggregated population information available. This will impact on both the bias and variance estimates for the population based R-indicators. Such extensions are relatively straightforward but will be left to future papers.

The research into population-based R-indicators is still at the beginning stage and it is too early to provide a definitive answer to the last research question presented in the introduction regarding the feasibility and practicability of R-indicators based on aggregate population auxiliary information. As mentioned in the introduction, further usage of these R-indicators are being explored in the context of evaluating and monitoring streamed administrative data and assessing the representativeness of linked records. In addition, Schouten et al. (2011) introduced partial R-indicators under sample-based auxiliary information for evaluating the lack of representativeness due to a specific auxiliary variable or category. These were used for monitoring and evaluating data collection. Schouten and Shlomo (2017) demonstrate the use of partial

R-indicators for adaptive survey designs. It is straightforward, similarly, to define population-based partial R-indicators and this will be a subject of future work.

Regarding the evaluation study presented in Section 4 on survey representativeness, it is based on real data under realistic assumptions of response probabilities typically found in social surveys conducted at national statistical institutes. Future research needs to assess whether alternative estimators can be constructed that are more precise, and, consequently, allow for stronger conclusions regarding the nature of response. A natural avenue to explore is an iterative approach through a modification of the EM-algorithm, in which the score of the nonrespondents on the auxiliary variables is estimated and used to update response propensity estimates.

We did not consider population-based estimation for other types of models such as logistic or probit regression. As shown in the numerical evaluation in Section 4, differences in sample-based estimators between the linear and logistic link function are in general small, but when the response rates get very close to 1, they become more evident. For these cases, developing other link functions for population-based estimation is a subject of future research. This would be a useful and natural extension to the theory of R-indicators as these models are often used in practice and avoid propensities outside the [0, 1] interval.

Acknowledgements

Part of the research presented here was developed within project RISQ (Representativity Indicators for Survey Quality, www.risq-project.eu), funded by the European 7th Framework Programme. We thank the members of the RISQ project: Katja Rutar from Statistični Urad Republike Slovenije, Geert Loosveldt and Koen Beullens from Katholieke Universiteit, Leuven, Øyvin Kleven, Johan Fosen and Li-Chun Zhang from Statistisk Sentralbyrå, Norway, Ana Marujo from the University of Southampton, UK and Paul Knottnerus, Centraal Bureau voor de Statistiek, for their valuable input.

The first author was supported by a STSM Grant from the COST Action IS1004 and by the ex 60% University of Bergamo, Biffignandi grant.

Appendix A

Analytic approximation to the bias of Type 1 $\tilde{R}_{\tilde{\rho}_{T1}}$ estimators

First, we compute the bias of $\tilde{S}^2_{\tilde{\rho}_{T1}}$ under general sampling design. Letting $\hat{m}_1 = N^{-1}\sum_r d_i$ and $\hat{m}_2 = N^{-1}\sum_r d_i\tilde{\rho}_{i,T1}$, then we can write

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = E\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) - S^2_{\rho} = \frac{N}{N-1}\left\{E\left(\hat{m}_2\right) - V\left(\hat{m}_1\right) - \left[E\left(\hat{m}_1\right)\right]^2\right\} - \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in U}\rho_i^2 - \bar{\rho}_U^2\right\}. \quad \text{(A.1)}$$

Note that

$$E(\hat{m}_2) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i \tilde{\rho}_{i,T1}\right) = \frac{1}{N}\sum_{i\in U} \mathbf{x}_i^T \mathbf{T}_1^{-1} E_s\left\{E_m\left(d_i^2 s_i r_i \mathbf{x}_i + \sum_{\substack{k\in U\\ k\neq i}} d_i d_k s_i s_k r_i r_k \mathbf{x}_k \mid s\right)\right\}$$

$$= \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1}\sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik} \rho_k \mathbf{x}_k,$$

$$E(\hat{m}_1) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i\right) = E_s\left(\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right) = \bar{\rho}_U,$$

and

$$V(\hat{m}_1) = V_s\{E_m(\hat{m}_1 \mid s)\} + E_s\{V_m(\hat{m}_1 \mid s)\}$$

$$= V_s\left\{\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right\} + E_s\left\{\frac{1}{N^2}\sum_{i\in U} d_i^2 s_i \rho_i (1-\rho_i)\right\}$$

$$= \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik} \rho_i \rho_k + \frac{1}{N^2}\sum_{i\in U} d_i \rho_i (1-\rho_i),$$

where $\Delta_{ik} = \pi_{ik} - \pi_i\pi_k$ and π_{ik} are the second-order sample inclusion probabilities. Hence, the bias of $\tilde{S}^2_{\tilde{\rho}_{T1}}$ with respect to the joint distribution of sampling design and the response mechanism is given by

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1}\sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik} \rho_k \mathbf{x}_k \right.$$

$$\left. - \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik} \rho_i \rho_k - \frac{1}{N^2}\sum_{i\in U} d_i \rho_i (1-\rho_i) - \frac{1}{N}\sum_{i\in U} \rho_i^2\right]. \qquad \text{(A.2)}$$

Under simple random sampling without replacement, (A.2) can be simplified to

$$B^{\text{SRS}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{1}{n}\sum_{i\in U}\rho_i\left\{1 - \frac{n-1}{N-1}\rho_i\right\}\mathbf{x}_i^T \mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n(N-1)N}\sum_{i\in U}\rho_i^2 - \frac{\bar{\rho}_U}{n} - \left(1-\frac{n}{N}\right)\frac{S^2_\rho}{n}\right].$$

A response-set based estimator of $B^{\text{SRS}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$ is

$$\tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\mathbf{x}_i^T \mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n} - \frac{n_r}{n^2}\right].$$

More generally, the Horvitz-Thompson response-set estimator for (A.2) under complex sampling is given by

$$\tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in r} d_i\left(d_i - \tilde{\rho}_{i,T1}\right)\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i - \frac{1}{N^2}\sum_{i\in r} d_i^3\Delta_{ii}\tilde{\rho}_{i,T1}\right.$$
$$- \frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\ k\neq i}} d_i d_k \frac{\Delta_{ik}}{\pi_{ik}} - \frac{1}{N^2}\sum_{i\in r} d_i^2\left(1-\tilde{\rho}_{i,T1}\right)$$
$$\left. + \frac{1}{N}\sum_{i\in r} x_i^T T_1^{-1}\sum_{\substack{k\in r\\ k\neq i}} x_k\left(d_i d_k - \frac{1}{\pi_{ik}}\right)\right\}.$$

Appendix B

Analytic approximation to the bias of Type 2 $\tilde{R}_{\tilde{\rho}_{T2}}$ estimators

The strategy to compute an analytical bias adjustment for $\tilde{S}^2_{\tilde{\rho}_{T2}}$ is to first approximate $\tilde{\rho}_{i,T2}$ by a linear estimator using Taylor linearization techniques. Next, compute an approximate bias adjustment for $\tilde{S}^2_{\tilde{\rho}_{T2}}$, by inserting the linear approximation for $\tilde{\rho}_{i,T2}$ into $\hat{m}_2$.

In the following, define, for $j = 1, \ldots, p$ and $j' = 1, \ldots, p$, the estimated totals

$$\hat{t}_0 = \sum_s d_k r_k, \quad \hat{t}_{jj'} = \sum_s d_k r_k z_{jk} z_{j'k}, \quad \text{and} \quad \hat{t}_j = \sum_s d_k r_k x_{jk},$$

where $\mathbf{z}_k = (\mathbf{x}_k - \overline{\mathbf{x}}_U)$ and $z_{jk} = \left(x_{jk} - \overline{x}_{jU}\right)$. Let $\hat{t}$ be a p-vector with components $\hat{t}_j$, and $\hat{\mathbf{F}}$ be the symmetric $(p \times p)$-matrix with elements $\hat{t}_{jj'}$. We may write

$$\tilde{\rho}_{i,T2} = \mathbf{x}_i^T\left[N\hat{t}_0^{-1}\hat{\mathbf{F}} + N\overline{\mathbf{x}}_U\overline{\mathbf{x}}_U^T\right]^{-1}\hat{\mathbf{t}} = \mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}.$$

Define now the population totals

$$t_0 = \sum_U \rho_k, \quad \mathbf{F} = \sum_U \rho_k \mathbf{z}_k\mathbf{z}_k^T, \quad \text{and} \quad \mathbf{t} = \sum_U \rho_k \mathbf{x}_k.$$

Notice that $\hat{t}_0$ is unbiased for t_0, $\hat{\mathbf{F}}$ is unbiased for $\mathbf{F}$, and $\hat{\mathbf{t}}$ is unbiased for $\mathbf{t}$. Let $\mathbf{T}_2 = Nt_0^{-1}\mathbf{F} + N\overline{\mathbf{x}}_U\overline{\mathbf{x}}_U^T$.

Proposition 1. The estimator $\tilde{\rho}_{i,T2}$ defined in (2.7) may be approximated by

$$\tilde{\rho}_{i,T2} \cong \mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t}\left(\hat{t}_0 - t_0\right) - \mathbf{x}_i^T\mathbf{T}_2^{-1}Nt_0^{-1}\left(\hat{\mathbf{F}} - \mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t} + \mathbf{x}_i^T\mathbf{T}_2^{-1}\hat{\mathbf{t}}.$$

Proof. Following standard Taylor linearization (see Särndal, Swensson and Wretman, 1992 and Bethlehem, 1988), the estimator $\tilde{\rho}_{i,T2}$ may be approximated by

$$\tilde{\rho}_{i,T2} \cong \rho^*_{i,T2} + a_0\left(\hat{t}_0 - t_0\right) + \sum_{j=1}^{p}\sum_{j'\leq j} a_{jj'}\left(\hat{t}_{jj'} - t_{jj'}\right) + \sum_{j=1}^{p} a_j\left(\hat{t}_j - t_j\right), \tag{B.1}$$

where $\rho^*_{i,T2} = \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{t}$, and

$$a_0 = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_0}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\left[-\hat{\mathbf{T}}_2^{-1}\left(-N\hat{t}_0^{-2}\hat{\mathbf{F}}\right)\hat{\mathbf{T}}_2^{-1}\right]\hat{\mathbf{t}}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t},$$

$$a_{jj'} = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_{jj'}}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = -\mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-1}\mathbf{\Lambda}_{jj'}\right)\mathbf{T}_2^{-1}\mathbf{t},$$

$$a_j = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_j}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\boldsymbol{\lambda}_j,$$

where $\mathbf{\Lambda}_{jj'}$ is a $(p \times p)$-matrix with ones in positions $(j,\ j')$ and $(j',\ j)$ and zeros elsewhere and $\boldsymbol{\lambda}_j$ is a p-vector with the j^{th} component equal to one and zeros elsewhere. Inserting the partial derivatives into (B.1) gives the result.

Proposition 2. Under simple random sampling, an approximate bias for $\tilde{S}^2_{\tilde{\rho}_{T2}}$ with respect to the joint distribution of sampling design and the response mechanism is given by

$$\begin{aligned} B^{\text{SRS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\Bigg\{ & t_0^{-2}\frac{N}{n}\sum_U c_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} \\ & - t_0^{-1}\frac{N}{n}\sum_U \mathbf{b}_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\}\mathbf{z}_i\mathbf{z}_i^T\mathbf{T}_2^{-1}\mathbf{t} + \frac{1}{n}\sum_U \rho_i\mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{x}_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} \\ & + \frac{n-1}{n(N-1)}\sum_U \rho_i\rho^*_{i,T2} - \left(1-\frac{n}{N}\right)\frac{S^2_\rho}{n} - \frac{\overline{\rho}_U}{n} + \frac{1}{nN}\sum_U \rho_i^2 - \frac{1}{N}\sum_U \rho_i^2\Bigg\}, \end{aligned}$$

where $c_i = \mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{F}\mathbf{T}_2^{-1}\mathbf{t}$, $\mathbf{b}_i = \mathbf{x}_i^T\mathbf{T}_2^{-1}$ and $\rho^*_{i,T2} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{t}$.

A response-set based estimator of $B^{\text{SRS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ is

$$\begin{aligned} \tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\Bigg\{ & \frac{1}{n_r^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} \\ & - \frac{N}{nn_r}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{z}_i\mathbf{z}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} \\ & + \frac{N}{n^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{x}_i \\ & + \frac{n-1}{n^2(N-1)}\sum_r\tilde{\rho}_{i,T2} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T2}}}{n} - \frac{n_r}{n^2}\Bigg\}. \end{aligned}$$

Proof. Thanks to Proposition 1, $\hat{m}_2$ defined in Appendix A may be approximated as follows

$$\begin{aligned}\hat{m}_2 &= \frac{1}{N}\sum_U d_i s_i r_i \tilde{\rho}_{i,T2} \\ &\cong \frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t}\left(\hat{t}_0 - t_0\right) \\ &\quad -\frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1} Nt_0^{-1}\left(\hat{\mathbf{F}} - \mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t} + \frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\hat{\mathbf{t}} \\ &=: A + B + C.\end{aligned}$$

The expected values of the terms A, B, and C are

$$\begin{aligned}E(A) &= t_0^{-2}\sum_{i\in U} c_i d_i \rho_i + t_0^{-2}\sum_{i\in U} c_i d_i \sum_{k\neq i} d_k \rho_i \rho_k \pi_{ik} - t_0^{-1}\sum_{i\in U} c_i \rho_i, \\ E(B) &= -t_0^{-1}\sum_{i\in U} d_i b_i \rho_i \mathbf{z}_i \mathbf{z}_i^T \mathbf{T}_2^{-1}\mathbf{t} - t_0^{-1}\sum_{i\in U} d_i \mathbf{b}_i \sum_{k\neq i} d_k \rho_i \rho_k \pi_{ik} \mathbf{z}_k \mathbf{z}_k^T \mathbf{T}_2^{-1}\mathbf{t} + t_0^{-1}\sum_{i\in U} \rho_i \mathbf{b}_i \mathbf{F}\mathbf{T}_2^{-1}\mathbf{t},\end{aligned}$$

and

$$E(C) = \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\sum_{k\neq i} d_k \rho_k \pi_{ik}\mathbf{x}_k.$$

It follows that, under simple random sampling, $E(\hat{m}_2)$ becomes

$$\begin{aligned}E^{\text{SRS}}(\hat{m}_2) &= t_0^{-2}\frac{N}{n}\sum_U c_i \rho_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\} - t_0^{-1}\frac{N}{n}\sum_U \mathbf{b}_i \rho_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\}\mathbf{z}_i \mathbf{z}_i^T \mathbf{T}_2^{-1}\mathbf{t} \\ &\quad + \frac{1}{n}\sum_U \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{x}_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\} + \frac{n-1}{n(N-1)}\sum_U \rho_i \rho_{i,T2}^{*}.\end{aligned}$$

So the total bias under simple random sampling is obtained by inserting $E^{\text{SRS}}(\hat{m}_2)$ computed above into (A.1) and following the proof in Appendix A for the other terms.

The response-set based estimator $\tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ of $B^{\text{SRS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ is obtained by substituting t_0 with $\hat{t}_0 = Nn_r/n$, F with $\hat{\mathbf{F}} = Nn^{-1}\sum_r \mathbf{z}_k \mathbf{z}_k^T$, $\mathbf{T}_2$ with $\hat{\mathbf{T}}_2 = N\hat{t}_0^{-1}\hat{\mathbf{F}} + N\overline{\mathbf{x}}_U \overline{\mathbf{x}}_U^T$, and t with $\hat{\mathbf{t}} = Nn^{-1}\sum_r \mathbf{x}_k$.

Note that the bias adjustment $\tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ corresponds to "plugging-in" Type 2 quantities ($\tilde{\rho}_{i,T2}$ instead of $\tilde{\rho}_{i,T1}$, matrix $\hat{\mathbf{T}}_2$ instead of $\mathbf{T}_1$, and $\tilde{S}^2_{\tilde{\rho}_{T2}}$ instead of $\tilde{S}^2_{\tilde{\rho}_{T1}}$) into the analytical bias adjustment $\tilde{B}^{\text{SRS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$ developed for $\tilde{S}^2_{\tilde{\rho}_{T1}}$ with two additional terms due to the linearization of $\hat{\mathbf{T}}_2$.

More generally, the Horvitz-Thompson response-set estimator under complex sampling for the bias adjustment of Type 2 population-based R-indicator is given by

$$\begin{aligned}\tilde{B}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) &= \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in r} d_i\left(d_i - \tilde{\rho}_{i,T2}\right)\mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\mathbf{x}_i - \frac{1}{N^2}\sum_{i\in r} d_i^3 \Delta_{ii}\tilde{\rho}_{i,T2} - \frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\ k\neq i}} d_i d_k \frac{\Delta_{ik}}{\pi_{ik}}\right. \\ &\quad - \frac{1}{N^2}\sum_{i\in r} d_i^2\left(1 - \tilde{\rho}_{i,T2}\right) + \frac{1}{N}\sum_{i\in r} x_i^T \hat{T}_2^{-1}\sum_{\substack{k\in r\\ k\neq i}} x_k\left(d_i d_k - \frac{1}{\pi_{ik}}\right) \\ &\quad + \left(\sum_{k\in r} d_k\right)^{-2}\sum_{i\in r} d_i^2 \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} + \left(\sum_{k\in r} d_k\right)^{-2}\sum_{i\in r} d_i \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\sum_{k\neq i} d_k \\ &\quad \left. - \left(\sum_{k\in r} d_k\right)^{-1}\sum_{i\in r} d_i^2 \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\mathbf{z}_i \mathbf{z}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} - \left(\sum_{k\in r} d_k\right)^{-1}\sum_{i\in r} d_i \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\sum_{k\neq i} d_k \mathbf{z}_k \mathbf{z}_k^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\right\}.\end{aligned}$$

References

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

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

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 (428), 1282-1289.

Brick, J.M., and Jones, M.E. (2008). Propensity to respond and nonresponse bias. *METRON – International Journal of Statistics*, LXVI (1), 51-73.

Copas, J.B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B*, 45, 311-354.

Copas, J.B. (1993). The shrinkage of point scoring methods. *Journal of the Royal Statistical Society, Series C*, 42, 315-331.

De Heij, V., Schouten, B. and Shlomo, N. (2015). RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators and partial R-indicators, available at www.risq-project.eu.

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

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Kreuter, F. (2013). *Improving Surveys with Process and Paradata*, Edited monograph, New Jersey: John Wiley & Sons, Inc.

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

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29 (4), 557-582.

MOA (2015). User Instruction Gold Standard, Dutch Market Research Association, available at www.moaweb.nl/sevrices/services/gouden-standaard.html.

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

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

Särndal, C.-E., and Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2 (4), 361-387.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, New York: John Wiley & Sons, Inc.

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

Schouten, B., and Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *International Statistical Review*, 85 (1), 143-163.

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-eng.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016). Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society, Series A*, 179 (3), 727-748.

Schouten, B., Bethlehem, J., Beulens, K., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N. and Skinner, C. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80 (3), 382-399.

Shlomo, N., Skinner, C. and Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference,* 142, 201-211.

Van der Laan, D., and Bakker, B. (2015). Indicators for the representativeness of linked sources, NTTS 2015 Proceedings, available at https://ec.europa.eu/eurostat/cros/system/files/NTTS2015%20proceedings.pdf.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76 (3), 555-575.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7 (1), 45-55.

Wagner, J., and Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology,* 2, 323-342.

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

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 249-270
Statistics Canada, Catalogue No. 12-001-X

Semiparametric quantile regression imputation for a complex survey with application to the Conservation Effects Assessment Project

Emily Berg and Cindy Yu[1]

Abstract

Development of imputation procedures appropriate for data with extreme values or nonlinear relationships to covariates is a significant challenge in large scale surveys. We develop an imputation procedure for complex surveys based on semiparametric quantile regression. We apply the method to the Conservation Effects Assessment Project (CEAP), a large-scale survey that collects data used in quantifying soil loss from crop fields. In the imputation procedure, we first generate imputed values from a semiparametric model for the quantiles of the conditional distribution of the response given a covariate. Then, we estimate the parameters of interest using the generalized method of moments (GMM). We derive the asymptotic distribution of the GMM estimators for a general class of complex survey designs. In simulations meant to represent the CEAP data, we evaluate variance estimators based on the asymptotic distribution and compare the semiparametric quantile regression imputation (QRI) method to fully parametric and nonparametric alternatives. The QRI procedure is more efficient than nonparametric and fully parametric alternatives, and empirical coverages of confidence intervals are within 1% of the nominal 95% level. An application to estimation of mean erosion indicates that QRI may be a viable option for CEAP.

Key Words: Informative sample design; B-spline; Erosion.

1 Introduction

Missing data have important implications for analyses of survey data. Missing data can arise because sampled units refuse to participate in the survey, are difficult to locate, do not respond to sensitive questions, or drop out of longitudinal studies. If the missing values are related to the variable of interest, an analysis of the complete data with no modification for missing values, is biased. Weighting and imputation are two broad classes of missing data adjustments.

Two types of weighting adjustments are calibration (D'arrigo and Skinner, 2010 and Kott, 2006) and propensity score estimation (Kim and Riddles, 2012). In calibration, the weights for the respondents are adjusted so that the weighted sum of an auxiliary variable for the respondents is equal to the corresponding mean for the full sample or a population mean. In propensity score estimation, the sampling weight is multiplied by the inverse of an estimated response probability.

Imputation completes the data set, replacing missing response variables with imputed values. Imputation can simplify analyses in the presence of item nonresponse and improve consistency in results across users. We consider imputation of a response y, which may be missing, using an auxiliary variable x that is observed for the full sample. To allow flexibility in the model assumptions, we use a semiparametric quantile regression model to describe the relationship between x and y.

1. Emily Berg, Department of Statistics, Iowa State University. E-mail: emilyb@iastate.edu; Cindy Yu, Department of Statistics, Iowa State University.

A diverse range of imputation procedures exists (Kim and Shao, 2013). Parametric fractional imputation (Kim, 2011) and parametric multiple imputation (Rubin, 2004) generate imputed values from an estimate of a fully parametric model for the conditional distribution of the response given covariates. Hot deck imputation (i.e., Andridge and Little, 2010), in contrast, includes, a class of nonparametric procedures in which imputed values are selected from respondents. In some hot deck procedures, weights are assigned according to a proximity measure, defined by imputation classes (Brick and Kalton, 1996) or a metric (Rubin, 2004; Little, 1988) such as a kernel distance (Wang and Chen, 2009). Nonparametric imputation is more robust to model misspecification than fully parametric methods, but estimators based on nonparametric procedures can have poor efficiency in small samples. Semiparametric quantile regression imputation (QRI) is a compromise between nonparametric and fully parametric imputation procedures. In QRI, the imputed values for a single missing value are the estimated quantiles of the distribution of the missing observation conditional on a function of auxiliary variables. Because a semiparametric model for the quantile function is used, QRI is robust to model misspecification, and because values are imputed from estimated quantiles, QRI is resistant to extreme values. Chen and Yu (2016) develop QRI for simple random sampling from an infinite population. We extend Chen and Yu (2016) to allow unequal selection probabilities.

Many imputation procedures rely on a missing at random (MAR) assumption (Rubin, 1976). A common assumption is that the response variable (y, which may be missing) is conditionally independent of the missing indicator (1 if a response is provided and 0 otherwise) given the observed data. A direct application of this MAR definition to a complex survey specifies independence of the response variable and missing indicator variable conditional on the auxiliary variable and the sample inclusion indicators (Little, 1982; Pfeffermann, 2011). Berg, Kim and Skinner (2016) call the missing at random assumption that is defined conditional on the sample inclusion indicators sample missing at random. An alternative assumption, called population missing at random (Berg et al., 2016), is that the response variable is conditionally independent of the missing indicator given the auxiliary variable in the superpopulation, unconditional on the sample inclusion indicators. Berg et al. (2016) show that these two assumptions are not equivalent. We discuss these MAR concepts precisely in Section 2 and develop our procedure to be sufficiently flexible to accommodate either condition.

Our interest in semiparametric quantile regression for a complex survey is motivated in part by the Conservation Effects Assessment Project (CEAP), a complex survey intended to quantify soil and nutrient loss from crop fields. Because distributions of the response variables are highly skewed and contain extreme values, specification of an adequate fully parametric imputation model is difficult, and hot deck imputation procedures may have large variances. We investigate the use of QRI to address these issues in imputation for CEAP.

We demonstrate the theoretical validity and applicability of semiparametric quantile regression imputation in the context of a complex survey. Section 2 and Section 3, respectively, present the imputation algorithm and asymptotic properties. Section 4 and Section 5 demonstrate the properties of QRI through the

CEAP application and simulations, respectively. Section 6 concludes with a summary and a discussion of areas for future research.

2 Quantile regression imputation for complex survey data

Consider a conceptual framework in which samples are drawn from a finite population generated from a superpopulation model (Fuller, 2009b, Chapter 6). Let x_i and y_i have joint distribution $f(x_i, y_i)$ in the superpopulation. We define the conditional distribution of y_i given x_i through the conditional quantile function. Let $q_\tau(x_i)$ denote the τ^{th} quantile of the conditional distribution of y_i given x_i in the superpopulation, where $q_\tau(x_i)$ is defined by

$$P(y_i \leq q_\tau(x_i) \,|\, x_i) = \tau. \tag{2.1}$$

We specify a model for the quantiles because quantile regression models can describe a wide variety of distributions, as illustrated in Figure 2.1. The left panel of Figure 2.1 depicts a linear quantile regression model in which each conditional quantile function is represented with a different intercept and a different slope. The use of a different slope allows describing data with nonconstant variances. The right panel of Figure 2.1 illustrates a generalization to semiparamtric quantile regression, where the τ^{th} quantile of the conditional distribution of y_i is represented as a continuous function of x_i. In the imputation procedure, we assume $q_\tau(\cdot)$ is a function with $p+1$ continuous derivatives. We approximate $q_\tau(x_i)$ with a B-spline (de Boor, 2001; Chen and Yu, 2016; Yoshida, 2013; Hastie, Tibshirani and Friedman, 2009), as we explain in more detail in Section 2.2. To enable the use of the B-spline, we assume x_i has compact support but do not require further distributional assumptions for x_i.

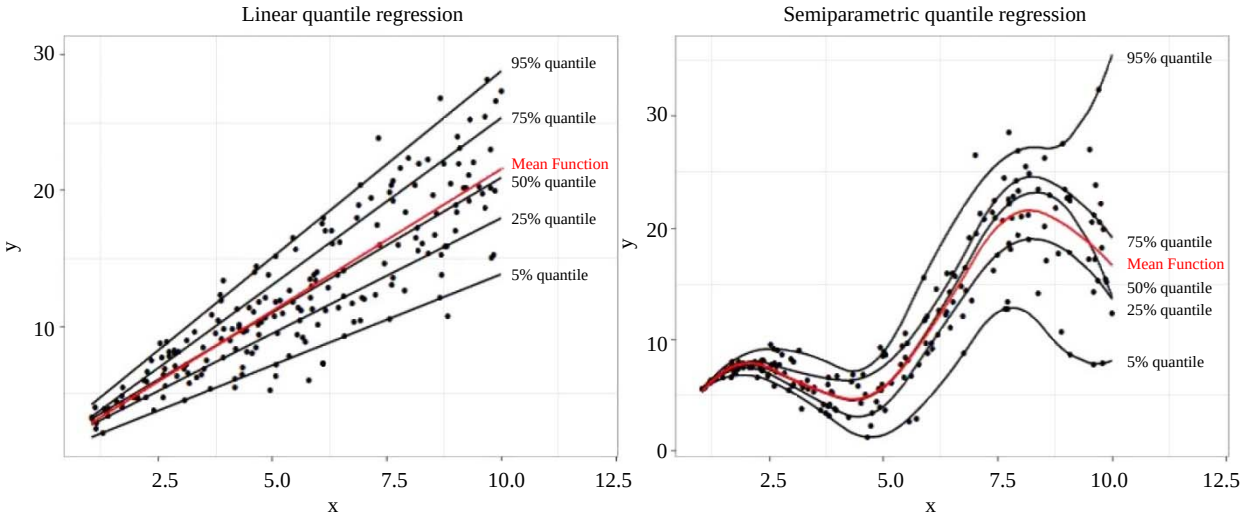


Figure 2.1 Illustration of linear quantile regression (left) and semiparametric quantile regression (right).

We consider estimation of parameters that are defined in terms of the superpopulation model relating y_i to x_i, rather than finite population parameters. The true parameter of interest, $\boldsymbol{\theta}_o$, is a $d-$dimensional vector satisfying,

$$E\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] = \mathbf{0}, \tag{2.2}$$

where $\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)$ is an $r-$dimensional function with two continuous derivatives, and $r \geq d$. The expectation operator $E\left[\cdot\right]$ denotes expectation with respect to the superpopulation model. Note that $E\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] = E\left[E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right]\right]$, where

$$\begin{aligned} E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] &= \int_{-\infty}^{\infty} \mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right) f_{y|x}\left(y_i \mid x_i\right) dy_i \\ &= \int_0^1 \mathbf{g}\left(F_{y|x}^{-1}\left(\tau\right),\, x_i;\, \boldsymbol{\theta}_o\right) \frac{f_{y|x}\left(F_{y|x}^{-1}\left(\tau\right) \mid x_i\right)}{f_{y|x}\left(F_{y|x}^{-1}\left(\tau\right) \mid x_i\right)} d\tau = \int_0^1 \mathbf{g}\left(q_\tau\left(x_i\right),\, x_i;\, \boldsymbol{\theta}_o\right) d\tau, \end{aligned} \tag{2.3}$$

and $F_{y|x}\left(y_i \mid x_i\right)$ and $f_{y|x}\left(y_i \mid x_i\right)$, respectively, denote the cumulative distribution function (cdf) and probability density function (pdf) of the conditional distribution of y_i given x_i. The second equality in (2.3) follows from the probability integral transform and a change of variables from y_i to the uniformly distributed τ with pdf $f\left(\tau\right) = I\left[\tau \in (0,1)\right]$, where $I\left[\cdot\right]$ is the indicator variable that takes the value 1 if the argument is true and 0 otherwise. The relationship defined by the third equality in (2.3) plays an important role in the imputation procedure. For each missing y_i, we construct J imputed values defined $\left\{\hat{q}_{\tau_1}\left(x_i\right), \ldots, \hat{q}_{\tau_J}\left(x_i\right)\right\}$, where $\hat{q}_{\tau_j}\left(x_i\right)$ estimates $q_{\tau_j}\left(x_i\right)$, and $\tau_1, \ldots, \tau_J$ form a fine grid on the interval $[0, 1]$. We then estimate $E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right]$ by approximating the integral in the last expression of (2.3) with an average of the J imputed values.

The imputation procedure consists of two main steps. We first construct the imputed values, estimating $q_\tau\left(x_i\right)$ using a linear combination of B-spline basis functions. We then estimate $\boldsymbol{\theta}_o$ using the generalized method of moments (GMM), replacing missing y_i with the estimate of $E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right]$ based on the imputed values and the relationship (2.3). To formalize the procedure, we require specific assumptions about the design and the response mechanism, which we specify in Section 2.1. Section 2.2 explains estimation of the quantile function, and Section 2.3 describes the generalized method of moments. Software for implementing the procedures is available from the authors.

2.1 Assumptions on design and response mechanism

Let I_i be the sample membership indicator, defined by $I_i = 1$ if unit i is selected. Let π_i and π_{ij} be the first and second order inclusion probabilities, respectively, defined by

$$\left[\pi_i,\, \pi_{ij}\right] = \left[P\left(I_i = 1 \middle| y_i,\, x_i\right),\, P\left(I_i = 1,\, I_j = 1 \middle| y_i,\, x_i,\, y_j,\, x_j\right)\right]. \tag{2.4}$$

Dependence of π_i on y_i in (2.4) represents a possible correlation between y_i and π_i that can cause the sample design to be informative for the quantile regression model (2.1). We denote the selected sample by A, where $A = \left\{i : I_i = 1\right\}$.

We assume x_i is observed for all i in A, while y_i may be missing. Let δ_i be the response indicator, defined by $\delta_i = 1$ if y_i is observed, and $\delta_i = 0$ if y_i is missing. Assume $\delta_i \sim \text{Bernoulli}(p_i)$, where the response probability p_i is defined as

$$p_i = P(\delta_i = 1 | y_i, x_i, I_i). \tag{2.5}$$

To define an approximately unbiased imputation procedure, we require an assumption about the relationship between δ_i and y_i. A common approach in missing data analysis is to assume that the response variable, y_i, is independent of the missing indicator, δ_i, conditional on the observed values (Little, 1982 and Pfeffermann, 2011). This assumption is a widely used interpretation of the missing at random (MAR) definition given in Rubin (1976) and clarified in Mealli and Rubin (2015). For a complex survey, the relationship between the inclusion probabilities, the response probabilities, and y can be complex if the response indicators and the sample inclusion indicators depend on a variable that is not included in the imputation model.

We follow the approach of Berg et al. (2016) and consider two assumptions about the relationship between δ_i and y_i. We define sample missing at random (SMAR) to mean

$$P(\delta_i = 1 | x_i, y_i, I_i) = P(\delta_i = 1 | x_i, I_i). \tag{2.6}$$

In contrast, we define population missing at random (PMAR) to mean

$$P(\delta_i = 1 | x_i, y_i) = P(\delta_i = 1 | x_i). \tag{2.7}$$

Berg et al. (2016) discuss situations in which the PMAR assumption may be viewed as reasonable and provide examples where PMAR holds while SMAR fails. If the response probabilities and the sample inclusion probabilities depend on a variable that is not included in the imputation model, then the PMAR may hold while SMAR does not. One example of a variable that may be excluded from the imputation model is a design variable. The analyst may omit a design variable from the imputation model if the design variable is unavailable at the imputation stage or because the imputation model is a subject-matter model relating y_i to x_i. We develop the QRI procedure to be flexible enough to accommodate either PMAR or SMAR. In practice, the analyst can decide whether PMAR or SMAR is more realistic for a particular application. In Section 2.2, we explain precisely how the nature of the missing at random assumption can impact the use of sampling weights in the estimation procedure. In the theory of Section 3, we focus on the situation in which assumption (2.7) holds.

2.2 Quantile regression with penalized B-Splines

We approximate the quantile function defining the relationship between y_i and x_i in the superpopulation with a linear combination of B-spline basis functions. A B-spline basis of order p spans the linear space of piecewise polynomials of degree $p-1$ with continuous derivatives up to order $p-2$.

B-splines allow improvements in computational efficiency over direct use of polynomial splines (Hastie, Tibshirani and Friedman, 2009).

To define the B-spline, we borrow terminology from Hastie, Tibshirani, and Friedman (2009) and Chen and Yu (2016). Assume x_i has compact support on the interval $[M_1, M_2]$. Define $K_n - 1$ interior knots, spaced at equidistant locations in the interval $[M_1, M_2]$ by, $\kappa_i = M_1 + [M_2 - M_1][K_n]^{-1} i$, for $i = 1, \ldots, K_n - 1$. Define p boundary knots at M_1 by κ_k for $k = -p+1, \ldots, 0$, and denote the p boundary knots at M_2 by κ_k for $k = K_n, \ldots, K_n + p - 1$. The $p^{\text{th}}-$degree B-spline basis functions for the knot sequence $\kappa_{-p+1}, \ldots, \kappa_{K_n+p-1}$ are the elements of the $K_n + p-$dimensional vector,

$$\mathbf{B}(x) = \left(B_{-p+1}^{[p]}(x), \ldots, B_{K_n}^{[p]}(x)\right)', \tag{2.8}$$

where $B_i^{[s]}(x)$ $(s = 1, \ldots, p)$ is defined recursively through divided differences. Specifically,

$$B_i^{[1]}(x) = I[\kappa_i \le x \le \kappa_{i+1}], \quad \text{for} \quad i = -p+1, \ldots, K_n + p - 2, \tag{2.9}$$

and

$$B_i^{[s]}(x) = \frac{x - \kappa_i}{\kappa_{i+s-1} - \kappa_i} B_i^{[s-1]}(x) + \frac{\kappa_{i+s} - x}{\kappa_{i+s} - \kappa_{i+1}} B_{i+1}^{[s-1]}(x), \tag{2.10}$$

for $i = -p+1, \ldots, K_n + p - 1 - s$ and $s = 2, \ldots, p$.

The estimator of the quantile regression function is defined by

$$\hat{q}_\tau(x) = \mathbf{B}(x)' \hat{\boldsymbol{\beta}}_\tau, \tag{2.11}$$

where the estimator $\hat{\boldsymbol{\beta}}_\tau$ is obtained by minimizing the quadratic form,

$$Q_\tau(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i w_i b_i \rho_\tau\left(y_i - \mathbf{B}(x_i)' \boldsymbol{\beta}\right) + \frac{\lambda_n}{2} \boldsymbol{\beta}' \mathbf{D}_m' \mathbf{D}_m \boldsymbol{\beta}, \tag{2.12}$$

where $w_i = \pi_i^{-1}\left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1}$, λ_n is a specified smoothing parameter, and $\rho_\tau(\cdot)$, b_i, and $\mathbf{D}_m$ are defined as follows. The function $\rho_\tau(u)$ in the first term of (2.12), is the check function of Koenker and Bassett (1978) defined by

$$\rho_\tau(u) = u(\tau - I[u < 0]). \tag{2.13}$$

Koenker's check function (2.13) is a standard optimization criterion for quantile regression because $q_\tau(x)$ minimizes the function $R(a) = E[\rho_\tau(y - a) | x]$ across a. The second term of (2.12) imposes a roughness penalty on the estimated quantile regression function. The matrix $\mathbf{D}_m$ is the m^{th} difference matrix with (i, j) element, $d_{ij} = (-1)^{j-i} C(m, j-i) I[0 \le j - i \le m] + (1 - I[0 \le j - i \le m])$, where $C(a, b)$ is the choose function. When $m = 2$, $\mathbf{D}_m$ has an interpretation related to the integral of the square of the second derivative of the function defined by the B-spline. Because the second derivative of a straight line is zero, the use of $\mathbf{D}_m$ for $m = 2$ shrinks the estimated quantile regression function toward a straight line. The appropriate choice of b_i in the first term of (2.12) depends on the assumptions about the nonresponse

mechanism. If (2.6) holds, then one may set $b_i = w_i^{-1}$, which leads to the unweighted estimating equation of Chen and Yu (2016). If (2.6) is not satisfied, the unweighted estimator may lead to bias, and setting $b_i = 1$ is one way to attain an approximately unbiased estimator (Berg et al., 2016). We focus on the conservative choice of $b_i = 1$, which leads to consistent estimators under (2.7) without requiring (2.6).

Remark 1. For simplicity, we consider a univariate x_i with support on a closed interval. Chen and Yu (2016) show that the procedure extends directly to a $h-$dimensional vector $\mathbf{x}_i$, each element of which has support on a closed interval. To extend the procedure to a vector $\mathbf{x}_i$, Chen and Yu (2016) define $\mathbf{B}(\mathbf{x}_i) = \left(\mathbf{B}(x_{1i})', \mathbf{B}(x_{2i})', \ldots, \mathbf{B}(x_{hi})'\right)$, where $x_{\tilde{h}i}$ is the $\tilde{h}^{\text{th}}$ element of $\mathbf{x}_i$, for $\tilde{h} = 1, \ldots, h$.

2.3 GMM estimation based on quantile regression imputation

Recall that the population parameter of interest is defined by the estimating equation in (2.2). We define a full sample estimator of $\boldsymbol{\theta}_o$ by

$$\hat{\boldsymbol{\theta}}_A = \text{argmin}_{\theta} \mathbf{G}_{n,A}(\boldsymbol{\theta})' \mathbf{G}_{n,A}(\boldsymbol{\theta}), \tag{2.14}$$

where

$$\mathbf{G}_{n,A}(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}), \tag{2.15}$$

w_i is defined following (2.12), and $i = 1, \ldots, n$ index the elements in A. The estimator defined by (2.15) is a a generalized method of moments estimator, where each element of $\mathbf{G}_{n,A}$ defines a deviation between a sample moment and the corresponding population parameter. For instance, if $\boldsymbol{\theta}_o = E[y_i]$, then $\mathbf{g}_i(y_i; \boldsymbol{\theta}_o) = (y_i - \boldsymbol{\theta}_o)$. Additional examples are provided in the simulation study of Section 5. Because y_i is unobserved for nonrespondents, $\hat{\boldsymbol{\theta}}_A$ is unattainable.

An imputed version of (2.15) is defined by replacing $\mathbf{g}(y_i, x_i, \boldsymbol{\theta})$ for an unobserved unit i by an estimator of the expected value. From (2.3), an estimator of $E_{y|x}[\mathbf{g}(y_i, x_i, \boldsymbol{\theta})]$ is $\int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau$, where $\hat{q}_{\tau i} = \hat{q}_\tau(x_i) = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_\tau$. We then define the estimator $\hat{\boldsymbol{\theta}}$ by,

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \left\{\mathbf{G}_n(\boldsymbol{\theta})' \mathbf{G}_n(\boldsymbol{\theta})\right\}, \tag{2.16}$$

where

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \left\{\delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1 - \delta_i) \int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau\right\}. \tag{2.17}$$

For specific $\mathbf{g}_i$ the minimizer of (2.16) has a closed form expression. For the case in which $\boldsymbol{\theta}_o = E[y_i]$, and $\hat{\boldsymbol{\theta}}$ is the Hájek estimator defined by

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n} w_i \left\{\delta_i y_i + (1 - \delta_i) \int_0^1 \hat{q}_\tau(x_i)\, d\tau\right\}.$$

In other situations, a closed form expression may not exist and standard numerical procedures, such as Newton-Raphson, can be used to minimize (2.17). In deriving the asymptotic results of Section 3, we

assume that $\boldsymbol{\theta}_o$ is the unique value such that $E[\mathbf{g}_i(y_i, \boldsymbol{\theta}_o)] = \mathbf{0}$, which relates to the existence of a unique minimum of (2.16). See Fuller (1996, page 252) for a similar condition and a discussion of the theory of estimators that minimize a quadratic form.

In practice, an approximation for the integral is required. We use a midpoint approximation (i.e., Nusser, Carriquiry, Dodd and Fuller, 1996). Let the fixed sequence $0 < \tau_1 \leq \tau_2 \cdots \leq \tau_J < 1$ be the mid-points of J evenly-spaced sub-intervals of [0, 1]. For non-respondent i, construct J imputed values,

$$y_{ij}^* = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_{\tau_j}, \quad j = 1, \ldots, J, \tag{2.18}$$

where $\hat{\boldsymbol{\beta}}_{\tau_j}$ is obtained by minimizing $Q_{\tau_j}(\boldsymbol{\beta})$ in (2.12). We define the estimator $\hat{\boldsymbol{\theta}}_J$ to satisfy

$$\hat{\boldsymbol{\theta}}_J = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \mathbf{G}_{n,J}(\boldsymbol{\theta})' \mathbf{G}_{n,J}(\boldsymbol{\theta}) \right\}, \tag{2.19}$$

where

$$\mathbf{G}_{n,J}(\boldsymbol{\theta}) := \mathbf{G}_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} w_i \left\{ \delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1 - \delta_i) J^{-1} \sum_{j=1}^{J} \mathbf{g}\left(y_{ij}^*, x_i, \boldsymbol{\theta}\right) \right\}, \tag{2.20}$$

w_i is defined following (2.12), and $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_{\tau_1}, \ldots, \hat{\boldsymbol{\beta}}_{\tau_J}\right)'$. The imputation procedure above differs from Chen and Yu (2016) in that the midpoint approximation for the integral is used instead of Monte Carlo integration. Both the midpoint approximation and Monte Carlo integration are justified by the probability integral transform, which relates the expectation to the conditional quantile function, as explained in (2.3). For functions with bounded second derivatives, the error in the midpoint approximation is $O(J^{-2})$. We also prefer the midpoint approximation because in simulations, it reduces the variance of the estimator and reduces instability in the variance estimator due to extreme quantiles relative to Monte Carlo simulation. Jang and Wang (2015) discuss the potential problem of unstable estimators for extreme quantiles from unstructured quantile regression models.

3 Asymptotic distributions and variance estimation

We derive an asymptotic normal distribution for the QRI estimator $\hat{\boldsymbol{\theta}}$ defined in (2.16), although the estimator $\hat{\boldsymbol{\theta}}_J$, defined in (2.19), with a finite number of (J) imputations is necessary in practice. This approach of developing theory under an assumption of an infinite number of imputed values has been used previously. See, for example, Clayton, Spiegelhalter, Dunn and Pickles (1998) and Robins and Wang (2000). The simulations in Section 5 demonstrate that the asymptotic normal distribution derived for $J = \infty$ is a reasonable approximation for the distribution of the estimator constructed with finite J. We outline the main concepts underlying the proofs of lemma 1, lemma 2, and Theorem 1, deferring details to Section B of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

The derivation of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ proceeds in three main steps. Lemma 1 gives the asymptotic distribution of the estimators of the quantile regression coefficients. Lemma 2 presents the asymptotic distribution of the estimating equation (2.17). These two lemmas are analogous to lemma 1 and lemma 2 of Chen and Yu (2016). Theorem 1 then provides the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

3.1 Asymptotic normality of $\hat{\boldsymbol{\theta}}$

We consider a sequence of samples and finite populations indexed by N, where the sample size $n \to \infty$ as $N \to \infty$. To define the regularity conditions, we introduce the notation $\mathcal{F}_N$ to represent an element of the sequence of finite populations with size N and use the notation "$|\mathcal{F}_N$" to indicate that the reference distribution is the distribution based on repeated sampling conditional on the finite population of size N. For example, $E\left[\hat{Y} \mid \mathcal{F}_N\right]$ and $V\left\{\hat{Y} \mid \mathcal{F}_N\right\}$, respectively, denote the conditional expectation and variance of the outcome $\hat{Y}$ with respect to the randomization distribution generated from repeated sampling from $\mathcal{F}_N$. Similarly, $\hat{Y} \xrightarrow{d} Y \mid \mathcal{F}_N$ a.s., means that $\hat{Y}$ converges in distribution to Y almost surely with respect to the process of repeated sampling from the sequence of finite populations as $N \to \infty$. The convergence is with probability 1 because $\mathcal{F}_N$ is a random realization from the superpopulation model (2.1).

The regularity conditions on the sample design and tuning parameters for the estimator of the B-spline model are as follows:

1. Any variable v_i such that $E\left[\,|v_i|^{2+\delta}\right] < \infty$, where $\delta > 0$, satisfies,

$$\sqrt{n}\left(\bar{v}_{\mathrm{HT}} - \bar{v}_N\right) \mid \mathcal{F}_N \xrightarrow{d} N\left(0, V_\infty\right) \quad \text{a.s.}, \tag{3.1}$$

   where $\left(\bar{v}_{\mathrm{HT}}, \bar{v}_N\right) = N^{-1}\sum_{i=1}^{N}\left(\pi_i^{-1} v_i I_i, v_i\right)$, $V_\infty = \lim_{N\to\infty} V_N$, and $V_N = nV\left\{\bar{v}_{\mathrm{HT}} \mid \mathcal{F}_N\right\}$ is the conditional variance of the Horvitz-Thompson mean, $\bar{v}_{\mathrm{HT}}$, given $\mathcal{F}_N$.

2. $nn_B^{-1} \to 1$ and $n_B N^{-1} \to f_\infty \in [0, 1]$, where n_B is the expected sample size.
3. There exist constants C_1, C_2, and C_3 such that $0 < C_1 \le n_B N^{-1}\pi_i^{-1} \le C_2 < \infty$, and

$$\left| n_B\left(\pi_{ij} - \pi_i\pi_j\right)\pi_i^{-1}\pi_j^{-1}\right| \le C_3 < \infty \quad \text{a.s.} \tag{3.2}$$

4. The value determining the number of interior knots $K_n = O\left(n_B^{\frac{1}{2p+3}}\right)$.
5. $\lambda_n = O\left(n_B^{\nu}\right)$ for $\nu \le (2p+3)^{-1}(p+m+1)$.

Condition 3 is also used in Fuller (2009a). Condition 3 holds for simple random sampling, where $(\pi_{ij} - \pi_i\pi_j)\,\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)\,(N-1)^{-1}N - 1$, and for Poisson sampling, where $(\pi_{ij} - \pi_i\pi_j)\,\pi_i^{-1}\pi_j^{-1} = 0$. Fuller (2009a) explains that condition 3 holds for many stratified designs and that the designer has the control to ensure condition 3.

Under assumptions 4-5, Barrow and Smith (1978) show that a $\boldsymbol{\beta}_\tau^*$ exists that satisfies,

$$\sup_{x\in[M_1, M_2]}\left| q_\tau(x) - b_\tau^q(x) - \mathbf{B}(x)'\,\boldsymbol{\beta}_\tau^*\right| = o\left(K_n^{-(p+1)}\right), \tag{3.3}$$

where $\mathbf{B}(x)' \boldsymbol{\beta}_\tau^*$ is the best L_∞ approximation for $q_\tau(x)$, and $b_\tau^{(a)}(x)$ is a bias of the B-spline approximation for the true quantile function, satisfying, $b_\tau^a(x) = O\left(K_n^{-(p+1)}\right)$. For details of the form of the bias term, see Chen and Yu (2016) and Yoshida (2013). The property (3.3) is used extensively in the derivation of lemma 1.

The proofs of both lemma 1 and lemma 2 use a result given in Theorem 1.3.6 of Fuller (2009b). Because of the importance of this theorem to the results of this section, we state this theorem as Fact 1:

Fact 1. (Theorem 1.3.6 of Fuller (2009b)): Suppose

$$\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N(0, V_{11}) \quad \text{a.s.,} \quad \text{and} \quad \theta_N - \theta_o \xrightarrow{d} N(0, V_{22}). \tag{3.4}$$

Then, $\left(\hat{\theta} - \theta_o\right) \xrightarrow{d} N(0, V_{11} + V_{22})$.

Note that V_{11} in Fact 1 is a fixed limit and not a design variance because the design variance is a random function of the finite population in this framework. The condition $\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N(0, V_{11})$ a.s., holds for a broad class of designs, such as those discussed in Isaki and Fuller (1982).

Lemma 1. Under assumptions 1-5 and for fixed $x_i \in [M_1, M_2]$ and $\tau \in (0, 1)$,

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau(x_i) - \mathbf{B}(x_i)' \boldsymbol{\beta}_\tau^* + b_\tau^\lambda(x_i)\right) \xrightarrow{d} N\left(0, \mathbf{B}(x_i)' \boldsymbol{\Sigma}_\infty(\tau) \mathbf{B}(x_i)\right), \tag{3.5}$$

and

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau(x_i) - q_\tau(x_i) + b_\tau^a(x_i) + b_\tau^\lambda(x_i)\right) \xrightarrow{d} N\left(0, \mathbf{B}(x_i)' \boldsymbol{\Sigma}_\infty(\tau) \mathbf{B}(x_i)\right), \tag{3.6}$$

where

$$\begin{aligned}
b_\tau^\lambda(x_i) &= \lim_{N\to\infty} \frac{\tilde{\lambda}_n}{n} \mathbf{B}(x_i)' \boldsymbol{\Omega}_n(\tau)^{-1} \mathbf{D}_m' \mathbf{D}_m \boldsymbol{\beta}_\tau^*, \qquad (3.7)\\
\boldsymbol{\Omega}_n(\tau) &= \mathbf{H}(\tau) + \frac{\tilde{\lambda}_n}{n} \mathbf{D}_m' \mathbf{D}_m,\\
\boldsymbol{\Sigma}_\infty(\tau) &= \lim_{N\to\infty} \frac{1}{K_n} \boldsymbol{\Omega}_n(\tau)^{-1} \left(\mathbf{V}_{1,\infty}(\tau) + f_\infty \tau(1-\tau) \boldsymbol{\Phi}\right) \boldsymbol{\Omega}_n(\tau)^{-1},\\
\mathbf{H}(\tau) &= E\left[p_i \mathbf{B}(x_i) f_{y|x,i}(q_{\tau i}) \mathbf{B}(x_i)'\right],\\
\boldsymbol{\Phi} &= E\left[p_i \mathbf{B}(x_i) \mathbf{B}(x_i)'\right],\\
\mathbf{V}_{1,\infty}(\tau) &= \lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \delta_i \delta_j \mathbf{B}(x_i) \psi_\tau(u_{i\tau}) \mathbf{B}\left(x_j\right)' \psi_\tau\left(u_{j\tau}\right),
\end{aligned}$$

$u_{i\tau} = y_i - \mathbf{B}(x_i)' \boldsymbol{\beta}_\tau^*$, $\psi_\tau(u) = \tau - I[u < 0]$, $\tilde{\lambda}_n = n\hat{N}N^{-1}\lambda_n$, $\hat{N} = \sum_{i=1}^n \pi_i^{-1}$, and $f_{y|x,i}(q)$ is the pdf of y_i given x_i evaluated at q.

The main idea of the proof of lemma 1 is to show that the estimator of the quantile regression coefficient has a Bahadur representation given in corollary 1 below:

Corollary 1: By the proof of lemma 1, the estimator of the quantile regression coefficient has the following Bahadur representation:

$$\sqrt{\frac{n}{K_n}}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^* + \frac{\tilde{\lambda}_n}{n}\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{D}_m'\mathbf{D}_m\boldsymbol{\beta}_\tau^*\right) = \sqrt{\frac{n}{K_n}}\boldsymbol{\Omega}_n(\tau)^{-1}\frac{1}{N}\sum_{i=1}^n \pi_i^{-1}\delta_i\mathbf{B}(x_i)\psi_\tau(u_{i\tau}) + o_p(1). \quad (3.8)$$

The derivation of the Bahadur representation follows the basic approach of Koenker (2005) and Yoshida (2013). To account for the complex sample design, condition (3.2) is used to bound sums of covariances induced by nontrivial second order inclusion probabilities. For independent random variables from an infinite population (as in Chen and Yu (2016), Yoshida (2013) and Koenker (2005)), the corresponding covariances are zero. Given the Bahadur representation (3.8), lemma 1 follows from an application of the regularity condition in (3.1) and Fact 1 to the elements of the Horvitz-Thompson mean in (3.8). The $V_{1,\infty}$ in $\boldsymbol{\Sigma}_\infty(\tau)$ essentially plays the role of V_{11} in Fact 1 and is the limit of the design variance of the Horvitz-Thompson mean. The second term in $\boldsymbol{\Sigma}_\infty(\tau)$ is the asymptotic variance of the design-expectation of the Horvitz-Thompson mean and plays the role of V_{22} in Fact 1.

Lemma 2 and Theorem 1 require additional regularity conditions about the estimating equation. The regularity conditions on the estimation are similar to those in Chen and Yu (2016) and are therefore deferred to Section A of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

Lemma 2. Under the assumptions of lemma 1 and the regularity conditions provided in Section A of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016),

$$\sqrt{n}\,\mathbf{G}_n(\boldsymbol{\theta}_o) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_G(\boldsymbol{\theta}_o)), \quad (3.9)$$

where

$$\mathbf{V}_G(\boldsymbol{\theta}_o) = f_\infty V\{\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\} + \lim_{N\to\infty}\mathbf{V}_{\xi,N}(\boldsymbol{\theta}_o), \quad (3.10)$$

$$\mathbf{V}_{\xi,N} = nN^{-2}\sum_{i=1}^N\sum_{j=1}^N \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\boldsymbol{\xi}_j(\boldsymbol{\theta}_o)',$$

$$\boldsymbol{\xi}_i(\boldsymbol{\theta}_o) = \delta_i\mathbf{g}_i(y_i;\boldsymbol{\theta}_o) + (1-\delta_i)\int_0^1 \mathbf{g}_i(q_\tau(x_i);\boldsymbol{\theta}_o)\,d\tau + \delta_i\mathbf{h}_{ni}(\boldsymbol{\theta}_o),$$

$$\mathbf{h}_{n_i}(\boldsymbol{\theta}_o) = \int_0^1 E\left[(1-p_j)\dot{\mathbf{g}}_{j,y}(q_\tau(x_j);\boldsymbol{\theta}_o)\mathbf{B}(x_j)'\right]\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{B}(x_i)\psi_\tau(u_{i\tau})\,d\tau,$$

and $\dot{\mathbf{g}}_{i,y}(y_i; \boldsymbol{\theta}_o)$ is the partial derivative of $\mathbf{g}_i(a; \boldsymbol{\theta})$ with respect to a evaluated at y_i.

The proof of lemma 2 centers on the Taylor expansion given by

$$\begin{aligned}\mathbf{g}_i(\hat{q}_\tau(x_i); \boldsymbol{\theta}_o) &= \mathbf{g}_i(q_\tau(x_i); \boldsymbol{\theta}_o) + \dot{\mathbf{g}}_{i,y}(q_\tau(x_i); \boldsymbol{\theta}_o)(\hat{q}_\tau(x_i) - q_\tau(x_i)) \\ &\quad + \ddot{\mathbf{g}}_{i,y}(q_\tau(x_i); \boldsymbol{\theta}_o)(\tilde{q}_\tau(x_i) - q_\tau(x_i))^2, \end{aligned} \tag{3.11}$$

where $\tilde{q}_\tau(x_i)$ is between $\hat{q}_\tau(x_i)$ and $q_\tau(x_i)$, and $\ddot{\mathbf{g}}_{i,y}(q_\tau(x_i); \boldsymbol{\theta}_o)$ denotes the vector of partial derivatives of the elements of $\dot{\mathbf{g}}_{i,y}(a, \boldsymbol{\theta}_o)$ with respect to a evaluated at $q_\tau(x_i)$. By arguments similar to those of Chen and Yu (2016), $n \left\| \ddot{g}_{i,y}(q_\tau(x_i); \boldsymbol{\theta}_o)(\tilde{q}_\tau(x_i) - q_\tau(x_i))^2 \right\| = O(1)$. Lemma 2 then follows from the linear approximation for $\hat{q}_\tau(x_i) - q_\tau(x_i)$ in lemma 1.

Theorem 1. Under the assumptions of lemmas 1 and 2, the QRI estimator $\hat{\boldsymbol{\theta}}$ defined in (2.16), constructed with $J = \infty$, satisfies, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$, where

$$\boldsymbol{\Sigma}_\theta = \left[\boldsymbol{\Gamma}(\boldsymbol{\theta}_o)' \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)\right]^{-1} \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)' \mathbf{V}_G(\boldsymbol{\theta}_o) \boldsymbol{\Gamma}(\boldsymbol{\theta}_o) \left[\boldsymbol{\Gamma}(\boldsymbol{\theta}_o)' \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)\right]^{-1}, \tag{3.12}$$

$\mathbf{G}(\boldsymbol{\theta}) = E[\mathbf{G}_N(\boldsymbol{\theta}, \mathbf{y})]$, $\mathbf{G}_N(\boldsymbol{\theta}, \mathbf{y}) = N^{-1} \sum_{i=1}^N \delta_i g(y_i, x_i)$, and $\boldsymbol{\Gamma}(\boldsymbol{\theta}_o) = E[\partial / \partial \boldsymbol{\theta} \mathbf{G}_N(\boldsymbol{\theta})]$.

By Pakes and Pollard (1989), Theorem 1 is satisfied if the following hold:

1. $\sup_\theta |\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})| = o_p(1)$,
2. For $\zeta_n \to 0$, $\sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_o| < \zeta_n} |\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) - \mathbf{G}_n(\boldsymbol{\theta}_o)| = o_p(n_B^{-0.5})$,

where ζ_n is arbitrarily small. Because of the complex sample design, the proof that these conditions hold proceeds in two steps, considering first the deviation $|\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}_N(\boldsymbol{\theta})|$ and then the deviation $|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})|$. The result then follows from the triangle inequality.

3.2 Variance estimation

We estimate the variance of $\hat{\boldsymbol{\theta}}_J$ using the linearization method (Fuller, 2009b, page 64). We use the asymptotic covariance matrix in (3.12) to estimate the variance of $\hat{\boldsymbol{\theta}}_J$, the estimator of $\boldsymbol{\theta}_o$ defined in (2.19), constructed with a finite number of imputed values. To estimate $\mathbf{V}_G(\boldsymbol{\theta}_o)$, a design-consistent variance estimator is applied to an estimator of the mean of an estimator of $\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)$ defined in (3.10). The estimator of $\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)$ is obtained by replacing $\boldsymbol{\theta}_o$ and $\boldsymbol{\beta}_\tau^*$ with estimators $\hat{\boldsymbol{\theta}}_J$ and $\hat{\boldsymbol{\beta}}_\tau$, respectively.

The estimator of variance is defined,

$$\hat{\boldsymbol{\Sigma}}_\theta = \left[\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J)' \hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J)\right]^{-1} \hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J)' \left[\hat{\mathbf{V}}_{G,\infty}(\hat{\boldsymbol{\theta}}_J)\right] \hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J) \left[\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J)' \hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}}_J)\right]^{-1}, \tag{3.13}$$

where $\hat{\mathbf{V}}_{G,\infty}(\hat{\boldsymbol{\theta}}_J) = \hat{f}_\infty \hat{V}\{\hat{\boldsymbol{\xi}}_i(\hat{\boldsymbol{\theta}}_J)\} + \hat{\mathbf{V}}_{\xi,N}(\hat{\boldsymbol{\theta}}_J)$,

$$\hat{V}\left\{\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right\} = \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)' - \frac{1}{\hat{N}\left(\hat{N}-1\right)}\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)', \tag{3.14}$$

$$\hat{\mathbf{V}}_{\xi,N}\left(\hat{\boldsymbol{\theta}}_J\right) = \frac{n}{\hat{N}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\boldsymbol{\xi}}_j\left(\hat{\boldsymbol{\theta}}_J\right)',$$

$$\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right) = \delta_i\mathbf{g}_i\left(y_i;\hat{\boldsymbol{\theta}}_J\right) + (1-\delta_i)J^{-1}\sum_{j=1}^{J}\mathbf{g}_i\left(\mathbf{B}(x_i)'\hat{\boldsymbol{\beta}}_{\tau_j};\hat{\boldsymbol{\theta}}_J\right) + \delta_i\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right),$$

$$\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right) = J^{-1}\sum_{j=1}^{J}N^{-1}\sum_{k=1}^{n}\pi_k^{-1}(1-\delta_k)\dot{\mathbf{g}}_{k,y}\left(\mathbf{B}(x_k)'\hat{\boldsymbol{\beta}}_{\tau_j};\hat{\boldsymbol{\theta}}_J\right)\mathbf{B}(x_k)'\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right)^{-1}\mathbf{B}(x_i)\psi_\tau\left(\hat{u}_{i\tau_j}\right),$$

$$\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right) = \hat{\mathbf{H}}\left(\tau_j\right) + \frac{\hat{f}_\infty\tilde{\lambda}_n}{n}\mathbf{D}_m'\mathbf{D}_m,$$

$$\hat{\mathbf{H}}(\tau) = \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\delta_i\mathbf{B}(x_i)\,\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)\mathbf{B}(x_i)',$$

$\hat{f}_\infty = n\hat{N}^{-1}$, $\hat{N} = \sum_{i=1}^{n}\pi_i^{-1}$, and $\hat{u}_{i\tau_j} = y_i - \mathbf{B}(x_i)'\hat{\boldsymbol{\beta}}_{\tau_j}$. An estimator of $\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)$ is the inverse of an estimator of the derivative of the quantile function and is defined by

$$\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right) = \max\left\{\frac{2a_{n,\tau}}{\mathbf{B}(x_i)'\left(\hat{\boldsymbol{\beta}}_{\tau+a_{n,\tau}} - \hat{\boldsymbol{\beta}}_{\tau-a_{n,\tau}}\right)}, 0\right\}, \tag{3.15}$$

where the bandwidth $a_{n,\tau}$ is given by

$$a_{n,\tau} = n^{-0.2}\left[\frac{4.5\phi\left(\Phi^{-1}(\tau)\right)^4}{\left(2\Phi^{-1}(\tau)^2+1\right)^2}\right], \tag{3.16}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$, respectively, the pdf and cdf of a standard normal distribution. See Wei, Ma and Carroll (2012) and Koenker (2005) for discussions of (3.15) and (3.16), respectively.

4 Application to Conservation Effects Assessment Project

The cropland component of the Conservation Effects Assessment Project (CEAP) consists of a series of surveys meant to measure soil and nutrient loss from crop fields. The first cropland assessment was a national survey conducted over the period 2003-2006. Data collection for a second national survey, planned for 2015-2016, was on-going at the time of writing this paper. Each of the time periods 2003-2006 and 2015-2016 is considered one time point for estimation. Data are collected over multiple years (i.e., 2003-2006 or 2015-2016) for operational reasons, and no unit is in the sample for two years in the same time period. Temporal changes of interest are changes between the two time periods, rather than changes between two years in the same time period. The temporal structure leads to unbalanced data because some units respond in both time periods, some units never respond, and some units respond in only one of the two time periods. Providing the data user with a complete, imputed data set with a single set of weights simplifies analyses involving more than one time point.

We investigate the feasibility of imputation for CEAP using a subset of the data collected during 2003-2005. We omit the data collected in 2006 because the sample design changed, and we do not have the information required to compute sampling weights for the 2006 survey. The data from the 2015-2016 survey are not yet collected. This analysis is considered an investigation of the feasibility of using QRI to impute missing data in CEAP in the direction of addressing the broader problem of estimation of change over time.

An understanding of the CEAP sample design requires an understanding of the design of the National Resources Inventory (NRI). The NRI monitors status and trends in land use, land cover, and erosion, with emphasis on characteristics related to natural resources and agriculture. Primary sampling units in the NRI are land areas called segments, which are approximately 160 acres. Each segment contains approximately three secondary sampling units, which are randomly selected locations called points. From 1982-1997, the same sample of approximately 300,000 segments, referred to as the foundation sample, was revisited every five years. The foundation sample is a stratified sample of segments, with a typical sampling rate of approximately 4%. See Nusser and Goebel (1997) for details of the design of the NRI foundation sample. In 2000, the NRI transitioned to annual sample design. Because revisiting every sampled segment in the foundation sample on an annual basis is infeasible, a rotating panel design is used. A subsample of the foundation segments, called the core panel, is revisited annually. The core panel is supplemented with a rotation panel, which changes each year. In essence, the core and rotation panels are stratified samples of the foundation sample. The strata, called sample classes, depend on the characteristics of the NRI segment observed in 1982-1997, such as presence of wetlands, cropland, and forest. See Nusser (2006) and Breidt and Fuller (1999) for further detail on the NRI annual samples.

For the Conservation Effects Assessment Project (CEAP), data collectors visit a subset of the NRI points that are located in sampled crop fields and collect more detailed information on crop choices and conservation practices. The sample for the 2003-2005 CEAP survey essentially consists of segments in the NRI core panel, 2002 rotation panel, and 2003 rotation panel that contain at least one cropland point. For segments containing more than one cropland point, one cropland point was selected randomly. The selection of one point per segment is an effort to improve geographic spread and reduce the number of instances in which a farm operator associated with multiple sampled points is selected into the sample, thereby reducing the respondent burden.

Because the first phase sampling rate for the NRI is small $(\approx 4\%)$, we approximate the CEAP sample as a probability proportional to size with replacement sample. The selection probabilities for CEAP largely reflect the sample design for the NRI. Details of construction of first and second order selection probabilities for CEAP are provided in Section C of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

Data collection for crop fields sampled for the CEAP survey consists of multiple components. An important component is a farmer interview survey that collects detailed information on farming managements and conservation practices. Nonresponse can occur in CEAP if a farmer refuses to participate in the interview.

Response variables in CEAP are measurements of different types of soil and nutrient loss, obtained from a physical process model called the Agricultural Policy Environmental Extender (APEX). The APEX model converts data from the farmer surveys as well as information from administrative sources and the NRI to numerical measures of erosion. For this study, we consider a measure of soil loss due to sheet and rill erosion called RUSLE2, discussed further in Section 4.1.

The NRI survey provides a convenient source of auxiliary information for imputing CEAP response variables. Because the NRI survey data are collected through aerial photographs of sampled segments, nonresponse due to refusals does not occur in the NRI. As a consequence, NRI data are available for all sampled points in CEAP. Furthermore, the NRI collects data related to land use, conservation practices and erosion – characteristics that are expected to be correlated with outputs of the APEX model. As an auxiliary variable, we use USLE, a measure of sheet and rill erosion collected in the NRI.

Domains of interest in CEAP are ten "CEAP production regions". We focus on estimation of mean RUSLE2 for seven states (Iowa, Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin) that comprise the majority of the CEAP production region called the Corn Belt. We use semiparametric quantile regression to impute missing values for RUSLE2 using USLE as an auxiliary variable for each of these seven states in the Corn Belt region.

4.1 Imputation model and procedures

The variable of interest, RUSLE2, is a measure of sheet and rill erosion obtained from the APEX model. Because interest is in mean erosion on a per acre basis, the parameter of interest θ, the mean RUSLE2 erosion in the state, is defined as a ratio by,

$$\theta = \frac{E\left[\sum_{i=1}^{m_{ek}} R_{ik} D_k m_k^{-1}\right]}{E\left[m_{ek} D_k m_k^{-1}\right]}, \tag{4.1}$$

where R_{ik} is the RUSLE2 erosion for point i in segment k sampled in the period 2003-2005, D_k is the area of segment k, m_k is the total number of points in segment k, and m_{ek} is the number of points in segment k that are eligible for the CEAP survey. As discussed above, the period 2003-2005 is considered one time point, and no point is sampled more than once in this collection of years. Therefore, each sampled unit has one value R_{ik} for this set of years, and R_{ik} does not need a subscript of t for year.

The RUSLE2 erosion is an advancement of a simpler measure of sheet and rill erosion called USLE. The USLE is a product of five numerical indexes associated with slope steepness and length, rainfall, soil erodibility, conservation practices, and crop managements. While RUSLE2 is only observed for respondents to the CEAP survey, USLE is available from the main NRI sample for all points in the CEAP sample. We use the average USLE across years 2003-2005 as the covariate in the imputation model. Specifically, for point i in segment k, we define, $U_{ik} = 3^{-1}\sum_{t=2003}^{2005} U_{tik}$, where U_{tik} is the USLE soil loss in the NRI for point i in segment k for year t.

Because the RUSLE2 and USLE are highly skewed, the quantile regression model is applied after transforming both R_{ik} and U_{ik} by a power of 0.2. The quantile regression model postulated for the

superpopulation can be expressed as, $P(y_{ik} \le \tau \mid x_{ik}) = q_\tau(x_{ik})$, where $y_{ik} = R_{ik}^{0.2}$, and $x_{ik} = U_{ik}^{0.2}$. The unknown function $q_\tau(x_{ik})$ is approximated by a linear combination of B-spline basis functions generated from x_{ik}. To define the penalized B-spline, we set $p = 3$, $m = 2$, $K_n = 16$, and $\lambda = 0.004$.

Because the quantity of interest is erosion on a per acre basis, the estimator $\hat{\theta}$ of θ defined in (4.1) is a ratio of two estimators. That is, $\hat{\theta} = \hat{\theta}_2^{-1}\hat{\theta}_1$, where $\hat{\theta}_1$ is an estimator of $\theta_1 = E[D_k U_{ik}]$, and $\theta_2 = E[D_k]$. The estimator of θ_2 is the Hájek estimator, $\hat{\theta}_2 = \left(\sum_{k=1}^{n} \pi_{ik}^{-1} D_k\right)\left(\sum_{k=1}^{n} \pi_{ik}^{-1}\right)^{-1}$, where π_{ik} is the probability of selecting point i in segment k into the CEAP sample. The estimator $\hat{\theta}_1$ of θ_1 is obtained from GMM with $g(y, \theta_1) = (D_k y^5 - \theta_1)$.

4.2 Estimates and variance estimates

Table 4.1 contains estimates of average RUSLE2 soil loss based on QRI, along with estimated standard errors for seven states in the Corn Belt CEAP region. For comparison, the complete case estimator $(\bar{R}_{cc})$ and corresponding estimated standard error is also provided in Table 4.1. The complete case estimator is the ratio of Hájek estimators constructed using only the units that provide a usable response for RUSLE2.

For each of the seven states, the complete case estimator is larger than the estimator based on the imputed data. The imputation procedure reduces the estimator of θ, relative to the complete case estimator, because the weighted mean of U_{ik} among sampled units is smaller than the mean of U_{ik} among respondents, as shown in the last two rows of Table 4.1.

As expected, the estimated standard error for $\hat{\theta}$ is smaller than the estimated standard error for the complete case estimator. The ratios of the estimated variances for the complete case estimator to the estimated variances of $\hat{\theta}$ range from 1.103 for MN to 1.252 for IN. This comparison demonstrates the potential for efficiency gain due to the use of imputation. The reduction in estimated standard deviation occurs because the imputation procedure uses U_{ik} for the full sample, while the complete case estimator is based only on R_{ik} for the subset of respondents.

Table 4.1
Complete-case estimator $(\bar{R}_{cc})$ and QRI-GMM estimator $(\hat{\theta})$ of mean RUSLE2 soil loss (θ), corresponding standard errors, sample sizes (n), number of respondents (n_r), and weighted covariate means for sampled units $(\hat{U}_s)$ and weighted covariate means among respondents $(\hat{U}_r)$ for seven states in the Corn Belt

	IL	IN	IA	MI	MN	OH	WI
$\bar{R}_{cc}$	0.3301	0.2994	0.3464	0.3214	0.1741	0.3700	0.5226
$\mathrm{SE}(\bar{R}_{cc})$	0.0112	0.0179	0.0144	0.0209	0.0068	0.0213	0.0354
$\hat{\theta}$	0.3281	0.2901	0.3408	0.3145	0.1646	0.3636	0.4977
$\mathrm{SE}(\hat{\theta})$	0.0106	0.0160	0.0134	0.0189	0.0063	0.0201	0.0337
n	1,823	1,151	1,492	935	1,649	1,053	662
n_r	1,275	751	1,011	585	1,008	698	414
$\hat{U}_r$	4.0775	3.7781	5.2046	1.6029	2.1063	2.1071	4.7586
$\hat{U}_s$	4.0909	3.6107	5.0385	1.5776	1.8973	2.0761	4.2232

5 Simulations

We construct a simulation study to represent properties of the CEAP data and design. An extended set of simulations using the simulation models of Chen and Yu (2016) yields similar results and is not presented here for brevity. The objectives of the simulations are to evaluate the variance estimator and to compare QRI to nonparametric and fully parametric alternatives.

The fully parametric imputation procedure is parametric fractional imputation (Kim, 2011). The imputation model specified for parametric fractional imputation (PFI) is $y_i = \gamma_0 + \gamma_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The imputed values for PFI are generated as, $y_{ij}^* \sim N(\hat{\gamma}_0 + \hat{\gamma}_1 x_i, \hat{\sigma}_\epsilon^2)$, where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}_\epsilon^2)'$ satisfies $\mathbf{S}_w(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$,

$$\mathbf{S}_w(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \pi_i^{-1} \delta_i \mathbf{d}_i, \tag{5.1}$$

and $\mathbf{d}_i = \left(y_i - \gamma_0 - \gamma_1 x_i, (y_i - \gamma_0 - \gamma_1 x_i) x_i, (y_i - \gamma_0 - \gamma_1 x_i)^2 / \sigma_\epsilon^2 - 1\right)'$. By incorporating π_i^{-1} in the score function (5.1), the estimator is consistent if the population model is a linear model with iid normally distributed errors and either the MAR assumption in (2.7) or (2.6) holds.

The non-parametric imputation (NPI) procedure is based on Wang and Chen (2009). For NPI, the j^{th} imputed value for nonrespondent i, y_{ij}^*, is generated from a multinomial distribution with sample space $\{y_s: I_s = \delta_s = 1\}$. Specifically,

$$P\left(y_{ij}^* = y_s\right) = \frac{\pi_i^{-1} K\{(x_i - x_s)/h\}}{\sum_{j=1}^{N} I_j \delta_j \pi_j^{-1} K\{(x_i - x_j)/h\}}, \tag{5.2}$$

where $K(\cdot)$ is a normal kernel with bandwidth h selected by applying the method of Sheather and Jones (1991), as implemented in the R function $dpik$, to $\{x_i: I_s = \delta_s = 1\}$.

The QRI procedure is implemented as described in Sections 2-3. To define the penalized B-spline, we set $p = 3$, $m = 2$, $K_n = 16$, and $\lambda = 0.004$. The value of $\lambda = 0.004$ is the median of the values selected using the R function "cobbs" across 1,000 samples of a preliminary simulation. To select λ using "cobbs", we first use the R function "cobbs" to obtain λ_{τ_j} for $\tau_1, \ldots, \tau_J$. The selected λ is the minimum of the $\left\{\lambda_{\tau_j}: j = 1, \ldots, J\right\}$, which introduces the least amount of smoothing from among the selected λ_{τ_j}.

In simulations not presented here, we also consider multiple imputation. Modifications to standard multiple imputation procedures are needed to produce unbiased estimators for a situation in the sample missing at random assumption (2.6) does not hold (Berg et al., 2016; Reiter, Raghunathan and Kinney, 2006). Because an exploration of the modifications to multiple imputation needed to ensure consistent estimation is beyond the scope of this study, we restrict attention to PFI, NPI, and QRI.

For all three imputation procedures, GMM based on the imputed values is used to estimate the parameters. Note that this differs from Wang and Chen (2009), which uses empirical likelihood instead of

GMM. The number of imputations for the simulation is $J = 50$. The Monte Carlo (MC) sample size is 1,000.

We consider estimation of several parameters: $\theta_1 = E[y_i]$, $\theta_2 = V\{y_i\}$, $\theta_3 = \text{Cor}\{y_i, x_i\}$, $\theta_4 = E[E[y_i \mid x_i \leq 0.65]]$, and $\theta_5 = P(y_i \leq 8)$. With the exception of θ_5, GMM estimators of these parameters satisfy the assumptions required for the theory of Section 3. In particular, the function $\mathbf{g}_i(\cdot; \boldsymbol{\theta})$ defining the estimator of $(\theta_1, \theta_2, \theta_3, \theta_4)$ has two continuous derivatives. The estimator of θ_5 does not fall in the framework of Section 3 because $I[a \leq 8]$ is a non-smooth function of a; however, we evaluate the empirical properties of $\hat{\theta}_5$ defined as

$$\hat{\theta}_5 = \left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} \left\{\delta_i I[y_i \leq 8] + (1-\delta_i) J^{-1} \sum_{j=1}^{J} I\left[y_{ij}^* \leq 8\right]\right\}. \tag{5.3}$$

For details on the function $\mathbf{g}_i(\cdot; \boldsymbol{\theta})$ defining the estimators for the simulation, see Section D of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

5.1 Superpopulation model and design for simulations

The superpopulation model represents four aspects of the CEAP data and survey: (1) the shape of the expectation function, (2) the inclusion of a mean-variance relationship, (3) the use of probability proportional to size (PPS) with-replacement sampling, and (4) the sample sizes and response rates. The specific model for the simulation is $y_i = m(x_i) + e_i$, where $e_i \sim N\left(0, \sigma_e^2 m(x_i)^2\right)$, $m(x_i) = 2 + 10(1 + 8\exp(-5x_i))^{-\frac{5}{4}}$, and $x_i \sim$ Trunc. Norm.$(0.5, 0.3)$. The sample design is PPS with replacement, where the probability of selecting unit i on a single draw is $\left(\sum_{i=1}^{N} \tilde{\psi}_i\right)^{-1} \tilde{\psi}_i$, $\text{logit}(\tilde{\psi}_i) = -3 - 0.33z_i + 0.1y_i$, $z_i \sim$ Trunc. Norm.$(0.5, 0.3)$, and $N = 50{,}000$. The number of draws is $n = 1{,}500$, leading to a median sample size of 1,477, where the sample size is the number of unique units in the sample. The first and second order selection probabilities corresponding to $\tilde{\psi}_i$ are, $\pi_i = 1 - (1 - \tilde{\psi}_i)^n$, and $\pi_{ij} = 1 - (1 - \tilde{\psi}_i)^n - (1 - \tilde{\psi}_j)^n + (1 - \tilde{\psi}_j - \tilde{\psi}_i)^n$. The response indicator $\delta_i \sim$ Bernoulli(p_i), where $\text{logit}(p_i) = 0.5x_i + 1.5z_i$, which yields a median response rate of 0.631.

By the model for y_i given x_i, the assumption of population missing at random (2.7) holds for this simulation. Incorporating z_i in the models for p_i and π_i is the approach used in Berg et al. (2016) that causes the sample missing at random assumption (2.6) to fail. The variable z_i can be interpreted a design variable that is omitted from the imputation model.

5.2 Results

Table 5.1 contains three measures for comparing the QRI estimator to the PFI and NPI estimators. The percent relative MC MSE for estimator k $(k = \text{PFI, NPI})$ is defined,

$$\text{Pct. Rel. MSE}(k) = 100 \frac{\text{MSE}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{MSE}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}{\text{MSE}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}, \tag{5.4}$$

where $\hat{\theta}(k)$ is the estimator based on imputation procedure k. The percent relative variance for estimator k is defined,

$$\text{Pct. Rel. Var}(k) = 100\frac{\text{Var}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{Var}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}{\text{Var}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}, \tag{5.5}$$

for k = NPI, PFI. The percent of mean squared error due to squared bias is defined by

$$\text{Pct. Bias}(k) = 100\frac{\left(E_{\text{MC}}\left(\hat{\theta}(k)\right) - \theta\right)^2}{\text{MSE}_{\text{MC}}\left(\hat{\theta}(k)\right)}, \tag{5.6}$$

where k = NPI, PFI, QRI. The MSE of the QRI estimator is smaller than the MSE of the NPI and PFI estimators for all parameters. The PFI estimator is biased because the model underlying the PFI procedure does not account for the nonlinearity in the quantile curves or the nonconstant variances. The NPI procedure has a relatively large variance for sample sizes such as those obtained in the CEAP survey. The squared MC bias of the QRI procedure is less than 0.5% of MC MSE for all parameters.

The last two columns of Table 5.1 contain the relative bias of the variance estimator and the empirical coverage of normal theory 95% confidence intervals. The relative bias of the variance estimator defined as

$$\text{Rel. Bias} = \frac{E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right] - V_{\text{MC}}\left(\hat{\theta}\right)}{V_{\text{MC}}, \left(\hat{\theta}\right)}, \tag{5.7}$$

where $E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right]$ is the MC mean of the variance estimators and $V_{\text{MC}}\left(\hat{\theta}\right)$ is the MC variance of the QRI estimator. The MC relative bias of the variance estimator for the QRI estimator is between -6% and -1%. Empirical coverages of normal theory confidence intervals are within 1% of the nominal 95% level.

Table 5.1
MC properties of estimators and variance estimators for simulation with PPS with replacement sample design. Pct. Rel. MSE (5.4): Difference between the MC variance of the PFI or NPI estimator and the MC MSE of the QRI estimator, relative to the MC MSE of the QRI estimator. Pct. Rel. Var. (5.5): Difference between the MC variance of the PFI or NPI estimator and the MC MSE of the QRI estimator, relative to the MC MSE of the QRI estimator. Pct. Bias (5.6): percent of MC MSE of PFI, NPI, and QRI estimators due to squared MC bias. Rel. Bias = MC relative bias of variance estimator defined in (5.7). Coverage = MC coverage of 95% confidence intervals

	Pct. Rel. MSE		Pct. Rel. Var.		Pct. Bias			Rel. Bias	Coverage
	NPI	PFI	NPI	PFI	NPI	PFI	QRI	QRI	QRI
θ_1	0.509	1.624	0.211	1.589	0.304	0.041	0.006	-2.386	0.945
θ_2	3.308	1.882	1.011	-0.151	2.225	1.998	0.002	-1.113	0.951
θ_3	1.518	5.449	0.979	2.605	0.840	2.999	0.311	-5.772	0.943
θ_4	515.980	26.752	10.501	12.415	82.101	11.508	0.222	-3.182	0.952
θ_5	5.879	61.416	5.659	-2.345	0.223	39.510	0.015	–	–

6 Discussion

QRI is developed for a complex survey setting. Alternative choices of weights are discussed, and a closed form variance estimator is provided based on a linear approximation. Consistency and asymptotic normality of the estimators are demonstrated under the framework of an infinite number of imputed values. In simulations designed to represent the CEAP data, the variance estimator based on the asymptotic distribution has a relative bias less than 6% in absolute value and leads to confidence intervals with coverage close to the nominal level for finite J. Further, the estimator based on QRI is more efficient than an estimator based on PFI or NPI because QRI provides a reasonable compromise between bias and variance.

The quantile regression imputation procedure is applied to estimate mean erosion in seven states in the midwestern United States using data from the Conservation Effects Assessment Project. The analysis demonstrates that QRI presents a viable alternative to weighting adjustments currently used to account for nonresponse in CEAP.

Areas for improvement to QRI include the choice of τ_j, the choice of b_i, refinements to estimation of the quantile curves, and variance estimation for non-differentiable $\mathbf{g}(\cdot)$ functions. Development of automated methods to select the nuisance parameters, appropriate for selection of multiple quantiles in a complex survey setting, is an area for future research. Estimation of the quantile curves subject to a restriction that the estimated curves are non-overlapping, has potential to improve estimation of the derivatives needed for the variance estimator. Section E of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016) provides further discussion of areas for improvement.

Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78, 40-64.

Barrow, D.L., and Smith, P.W. (1978). Asymptotic properties of the best L_2 [0, 1] approximation by Splines with variable knots. *Quaterly of Applied Mathematics*, 33, 293-304.

Berg, E.J., and Yu, C. (2016). Supplement to "Semiparametric quantile regression imputation for a complex survey with application to the conservation effects assessment project". Available at: https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf.

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

Breidt, F.J., and Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 391-403.

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Chen, S., and Yu, C. (2016). Parameter estimation through semiparametric quantile regression imputation. *Electronical Journal of Statistics*, 10, 3621-3647.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B*, 60, 71-87.

D'Arrigo, J., and Skinner, C. (2010). Linearization variance estimation for generalized raking estimators in the presence of nonresponse. *Survey Methodology*, 36, 2, 181-192. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010002/article/11380-eng.pdf.

De Boor, C. (2001). *A Practical Guide to Splines* (Revised Edition), New York: Springer-Verlag.

Fuller, W.A. (1996). *Introduction to Statistical Time Series: Second Edition*. New York: John Wiley & Sons, Inc.

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

Fuller, W.A. (2009b). *Sampling Statistics*. New York: John Wiley & Sons, Inc. Vol. 560.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer. Vol. 2, No. 1.

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

Jang, W., and Wang, J.H. (2015). A semiparametric Bayesian approach for joint-quantile regression with clustered data. *Computational Statistics and Data Analysis*, 84, 99-115.

Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1), 119-132.

Kim, J.K., and Riddles, M.K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38, 2, 157-165. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11754-eng.pdf.

Kim, J.K., and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, Chapman and Hall/CRC, Boca Raton.

Koenker, R., and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Koenker, R. (2005). *Quantile Regression*. Cambridge university press. No. 38.

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

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data missing values. *Applied Statistics*, 37, 23-38.

Mealli, F., and Rubin, D. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102, 995-1000.

Nusser, S.M., and Goebel, J.J. (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3), 181-204.

Nusser, S.M. (2006). National Resources Inventory (NRI), US. *Encyclopedia of Environmetrics Second Edition*, 1-3.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W. and Fuller, W.A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436), 1440-1449.

Pakes, A., and Pollard, D. (1989). Simulation and the asymptotic of optimization estimators. *Econometrica*, 57(4), 1027-1057.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37, 2, 115-136. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11602-eng.pdf.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32, 2, 143-149. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9548-eng.pdf.

Robins, J.M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika*. 87, 113-124.

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

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc. Vol. 81.

Sheather, S.J., and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

Wang, D., and Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 490-517.

Wei, Y., Ma, Y. and Carroll, R.J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99, 423-438.

Yoshida, T. (2013). Asymptotics for penalized spline estimators in quantile regression. *Communications in Statistics - Theory and Methods*, DOI 10.1080/03610926.2013.765477.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 271-294
Statistics Canada, Catalogue No. 12-001-X

Multiple imputation of missing values in household data with structural zeros

Olanrewaju Akande, Jerome Reiter and Andrés F. Barrientos[1]

Abstract

We present an approach for imputation of missing items in multivariate categorical data nested within households. The approach relies on a latent class model that (i) allows for household-level and individual-level variables, (ii) ensures that impossible household configurations have zero probability in the model, and (iii) can preserve multivariate distributions both within households and across households. We present a Gibbs sampler for estimating the model and generating imputations. We also describe strategies for improving the computational efficiency of the model estimation. We illustrate the performance of the approach with data that mimic the variables collected in typical population censuses.

Key Words: Categorical; Census; Edit; Latent; Mixture; Nonresponse.

1 Introduction

In many population censuses and demographic surveys, statistical agencies collect data on individuals grouped within houses. In the U.S. decennial census, for example, the Census Bureau collects the age, race, sex, and relationship to the household head for every individual in the household, as well as whether or not the residents own the house. After collection, agencies share these datasets for secondary analysis, either as tabular summaries, public use microdata samples, or restricted access files.

When creating these data products, agencies typically have to deal with item nonresponse both for individual-level variables and household-level variables. They typically do so using some type of imputation procedure. Ideally, these procedures satisfy three desiderata. First, the imputations preserve the joint distribution of the variables as best as possible. As part of this, the procedure should preserve relationships within households. For example, the missing race of a spouse likely, but certainly not definitely, matches the race of the household head; the imputation procedure should reflect that. Second, the imputations respect structural zeros. For example, a daughter's age cannot exceed her biological mother's age. The imputations should not create impossible combinations of individuals in the same household. Third, the imputation procedure allows for appropriate uncertainty to be propagated in subsequent analyses of the data.

Typical approaches to imputation of missing household items use some variant of hot deck imputation (Kalton and Kasprzyk, 1986; Andridge and Little, 2010). However, depending on how the hot deck is implemented, it may not satisfy one or more of the desiderata. Indeed, we are not aware of any hot deck imputation procedure for household data that satisfies all three explicitly. An alternative is to estimate a model that describes the joint distribution of all the variables, and impute missing values from the implied predictive distributions in the model. For household data, one such model is the nested data Dirichlet process mixture of products of multinomial distributions (NDPMPM) model of Hu, Reiter and Wang (2018), which assumes that (i) each household is a member of a household-level latent class, and (ii) each individual is a

1. Olanrewaju M. Akande is Ph.D. Candidate, Department of Statistical Science, Duke University, Durham, NC 27708. E-mail: olanrewaju.akande@duke.edu; Jerome P. Reiter is Professor of Statistical Science, Duke University, Durham, NC 27708. E-mail: jerry@stat.duke.edu; Andrés F. Barrientos is Postdoctoral Associate, Department of Statistical Science, Duke University, Durham, NC 27708. E-mail: anfebar@stat.duke.edu.

member of an individual-level latent class nested within its household-level latent class. The model assigns zero probability to combinations corresponding to structural zeros, and also handles both household-level and individual-level variables simultaneously. The NDPMPM is appealing as an imputation engine, as it can preserve multivariate associations while avoiding imputations that result in impossible households. The NDPMPM is related to models proposed by Vermunt (2003, 2008) and Bennink, Croon, Kroon and Vermunt (2016), although these are used for regression rather than multivariate imputation and do not deal with structural zeros.

Hu et al. (2018) use the NDPMPM to generate synthetic datasets (Rubin, 1993; Raghunathan and Rubin, 2001; Reiter and Raghunathan, 2007) for statistical disclosure limitation, but they do not describe how to use it for imputation of missing data. We do so in this article. With structural zeros in the NDPMPM, the conditional distributions of the missing values given the observed values are not available in closed form. We therefore add a rejection sampling step to the Gibbs sampler used by Hu et al. (2018), which generates completed datasets as byproducts of the Markov chain Monte Carlo (MCMC) algorithms used to estimate the model. These completed datasets can be analyzed using multiple imputation inferences (Rubin, 1987). We also present two new strategies for speeding up the computations with NDPMPMs, namely (i) turning data for the household head into household-level variables rather than individual-level variables, and (ii) using an approximation to the likelihood function. These scalable innovations are necessary, as the NDPMPM is computationally quite intensive even without missing data. The speed-up strategies also can be employed when using the NDPMPM to generate synthetic data.

The remainder of this article is organized as follows. In Section 2, we review the NDPMPM model in the presence of structural zeros and the MCMC sampler for fitting the model without missing data. In Section 3, we extend the MCMC sampler for the NDPMPM model to allow for missing data. In Section 4, we present the two strategies for speeding up the MCMC sampler. In Section 5, we present results of simulation studies used to examine the performance of the NDPMPM as a multiple imputation engine, using the two strategies for speeding up the run time. In Section 6, we discuss findings, caveats and future work.

2 Review of the NDPMPM model

Hu et al. (2018) present the NDPMPM model including motivation for how it can preserve associations across variables and account for structural zeros. Here, we summarize the model without detailed motivations, referring the reader to Hu et al. (2018) for more information. We begin with notation needed to understand the model and the Gibbs sampler, assuming complete data. The presentation closely follows that in Hu et al. (2018).

2.1 Notation and model specification

Suppose the data contain n households. Each household $i = 1, \ldots, n$ contains n_i individuals, so that there are $\sum_{i=1}^{n} n_i = N$ individuals in the data. Let $X_{ik} \in \{1, \ldots, d_k\}$ be the value of categorical variable k for household i, which is assumed to be identical for all n_i individuals in household i, where $k = p+1, \ldots, p+q$. Let $X_{ijk} \in \{1, \ldots, d_k\}$ be the value of categorical variable k for person j in

household i, where $j = 1, \ldots, n_i$ and $k = 1, \ldots, p$. Let $\mathbf{X}_i = \left(X_{i(p+1)}, \ldots, X_{i(p+q)}, X_{i11}, \ldots, X_{in_i p}\right)$ include all household-level and individual-level variables for the n_i individuals in household i.

Let $\mathcal{H}$ be the set of all household sizes that are possible in the population. For all $h \in \mathcal{H}$, let $\mathcal{C}_h$ represent the set of all combinations of individual-level and household-level variables for households of size h, including impossible combinations; that is, $\mathcal{C}_h = \prod_{k=p+1}^{p+q}\{1, \ldots, d_k\}\prod_{j=1}^{h}\prod_{k=1}^{p}\{1, \ldots, d_k\}$. Let $\mathcal{S}_h \subset \mathcal{C}_h$ represent the set of impossible combinations, i.e., those that are structural zeros, for households of size h. These include combinations of variables within any individual, e.g., a three year old person cannot be a spouse, or across individuals in the same household, e.g., a person cannot be older than his biological parents. Let $\mathcal{C} = \bigcup_{h\in\mathcal{H}} \mathcal{C}_h$ and $\mathcal{S} = \bigcup_{h\in\mathcal{H}} \mathcal{S}_h$.

Although the NDPMPM model we use restricts the support of $\mathbf{X}_i$ to $\mathcal{C} - \mathcal{S}$, it is helpful for understanding the model to begin with no restrictions on the support of $\mathbf{X}_i$. Each household i belongs to one of F classes representing latent household types. For $i = 1, \ldots, n$, let $G_i \in \{1, \ldots, F\}$ indicate the household class for household i. Let $\pi_g = \Pr(G_i = g)$ be the probability that household i belongs to class g. Within any class, all household-level variables follow independent, multinomial distributions. For any $k \in \{p+1, \ldots, p+q\}$ and any $c \in \{1, \ldots, d_k\}$, let $\lambda_{gc}^{(k)} = \Pr(X_{ik} = c \,|\, G_i = g)$ for any class g, where $\lambda_{gc}^{(k)}$ is the same value for every household in class g. Let $\pi = \{\pi_1, \ldots, \pi_F\}$, and $\lambda = \{\lambda_{gc}^{(k)}\colon\ c = 1, \ldots, d_k;\ k = p+1, \ldots, p+q;\ g = 1, \ldots, F\}$.

Within each household class, each individual belongs to one of S individual-level latent classes. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let M_{ij} represent the individual-level latent class of individual j in household i. Let $\omega_{gm} = \Pr\left(M_{ij} = m \,|\, G_i = g\right)$ be the probability that individual j in household i belongs to individual-level class m nested within household-level class g. Within any individual-level class, all individual-level variables follow independent, multinomial distributions. For any $k \in \{1, \ldots, p\}$ and any $c \in \{1, \ldots, d_k\}$, let $\phi_{gmc}^{(k)} = \Pr\left(X_{ijk} = c \,|\, \left(G_i, M_{ij}\right) = (g, m)\right)$ for the class pair (g, m), where $\phi_{gmc}^{(k)}$ is the same value for every individual in the class pair (g, m). Let $\omega = \left\{\omega_{gm}\colon\ g = 1, \ldots, F;\ m = 1, \ldots, S\right\}$, and $\phi = \left\{\phi_{gmc}^{(k)}\colon\ c = 1, \ldots, d_k;\ k = 1, \ldots, p;\ m = 1, \ldots, S;\ g = 1, \ldots, F\right\}$.

For purposes of the Gibbs sampler in Section 2.2, it is useful to distinguish values of $\mathbf{X}_i$ that satisfy all the structural zero constraints from those that do not. Let the superscript "1" indicate that a random variable has support only on $\mathcal{C} - \mathcal{S}$. For example, $\mathbf{X}_i^1$ represents data for a household with values restricted only on $\mathcal{C} - \mathcal{S}$, i.e., not an impossible household, whereas $\mathbf{X}_i$ represents data for a household with any values in $\mathcal{C}$. Let $\mathcal{X}^1$ be the observed data comprising n households, that is, a realization of $(\mathbf{X}_1^1, \ldots, \mathbf{X}_n^1)$. The kernel of the NDPMPM, $\Pr(\mathcal{X}^1 \,|\, \theta)$, is

$$L(\mathcal{X}^1 \mid \theta) = \prod_{i=1}^{n}\sum_{h\in\mathcal{H}} 1\{n_i = h\}\, 1\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}\left[\sum_{g=1}^{F}\pi_g \prod_{k=p+1}^{p+q}\lambda_{gX_{ik}^1}^{(k)}\prod_{j=1}^{h}\sum_{m=1}^{S}\omega_{gm}\prod_{k=1}^{p}\phi_{gmX_{ijk}^1}^{(k)}\right], \tag{2.1}$$

where θ includes all the parameters, and $1\{.\}$ equals one when the condition inside the $\{\}$ is true and equals zero otherwise.

For all $h \in \mathcal{H}$, let $n_{1h} = \sum_{i=1}^{n} 1\{n_i = h\}$ be the number of households of size h in $\mathcal{X}^1$ and $\pi_{0h}(\theta) = \Pr(\mathbf{X}_i \in \mathcal{S}_h \,|\, \theta)$. As stated in Hu et al. (2018), the normalizing constant in the likelihood in (2.1) is $\prod_{h\in\mathcal{H}}(1 - \pi_{0h}(\theta))^{n_{1h}}$. Therefore, the posterior distribution is

$$\Pr(\theta \mid \mathcal{X}^1, T(\mathcal{S})) \propto \Pr(\mathcal{X}^1 \mid \theta) \Pr(\theta) = \frac{1}{\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{1h}}} L(\mathcal{X}^1 \mid \theta) \Pr(\theta) \tag{2.2}$$

where $T(\mathcal{S})$ emphasizes that the density is for the NDPMPM with support restricted to $\mathcal{C} - \mathcal{S}$.

The likelihood in (2.1) can be written as a generative model of the form

$$\begin{aligned} X_{ik} \mid G_i, \lambda \sim\ & \text{Discrete}\left(\lambda^{(k)}_{G_i 1}, \ldots, \lambda^{(k)}_{G_i d_k}\right) \\ & \forall i = 1, \ldots, n \text{ and } k = p+1, \ldots, p+q \end{aligned} \tag{2.3}$$

$$\begin{aligned} X_{ijk} \mid G_i, M_{ij}, \phi, n_i \sim\ & \text{Discrete}\left(\phi^{(k)}_{G_i M_{ij} 1}, \ldots, \phi^{(k)}_{G_i M_{ij} d_k}\right) \\ & \forall i = 1, \ldots, n,\ j = 1, \ldots, n_i \text{ and } k = 1, \ldots, p \end{aligned} \tag{2.4}$$

$$\begin{aligned} G_i \mid \pi \sim\ & \text{Discrete}(\pi_1, \ldots, \pi_F) \\ & \forall i = 1, \ldots, n \end{aligned} \tag{2.5}$$

$$\begin{aligned} M_{ij} \mid G_i, \omega, n_i \sim\ & \text{Discrete}\left(\omega_{G_i 1}, \ldots, \omega_{G_i S}\right) \\ & \forall i = 1, \ldots, n \text{ and } j = 1, \ldots, n_i \end{aligned} \tag{2.6}$$

where the Discrete distribution refers to the multinomial distribution with sample size equal to one. We restrict the support of each $\mathbf{X}_i$ to ensure the model assigns zero probability to all combinations in $\mathcal{S}$ as desired. The model in (2.3) to (2.6) can be used without restricting the support to $\mathcal{C} - \mathcal{S}$. This ignores all structural zeros. While not appropriate for the joint distribution of household data, this model turns out to useful for the Gibbs sampler. We refer to the generative model in (2.3) to (2.6) with support on all of $\mathcal{C}$ as the untruncated NDPMPM. For contrast, we call the model in (2.1) the truncated NDPMPM.

For prior distributions, we follow the recommendations of Hu et al. (2018). We use independent uniform Dirichlet distributions as priors for λ and ϕ, and the truncated stick-breaking representation of the Dirichlet process as priors for π and ω (Sethuraman, 1994; Dunson and Xing, 2009; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014),

$$\lambda^{(k)}_g = \left(\lambda^{(k)}_{g1}, \ldots, \lambda^{(k)}_{gd_k}\right) \sim \text{Dirichlet}(1, \ldots, 1) \tag{2.7}$$

$$\phi^{(k)}_{gm} = \left(\phi^{(k)}_{gm1}, \ldots, \phi^{(k)}_{gmd_k}\right) \sim \text{Dirichlet}(1, \ldots, 1) \tag{2.8}$$

$$\pi_g = u_g \prod_{f<g} (1 - u_f) \text{ for } g = 1, \ldots, F \tag{2.9}$$

$$u_g \sim \text{Beta}(1, \alpha) \text{ for } g = 1, \ldots, F-1,\ u_F = 1 \tag{2.10}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25) \tag{2.11}$$

$$\omega_{gm} = v_{gm} \prod_{s<m} (1 - v_{gs}) \text{ for } m = 1, \ldots S \tag{2.12}$$

$$v_{gm} \sim \text{Beta}(1, \beta_g) \text{ for } m = 1, \ldots, S-1,\ v_{gS} = 1 \tag{2.13}$$

$$\beta_g \sim \text{Gamma}\,(0.25,\ 0.25). \tag{2.14}$$

We set the parameters for the Dirichlet distributions in (2.7) and (2.8) to $\mathbf{1}_{d_k}$ (a d_k – dimensional vector of ones) and the parameters for the Gamma distributions in (2.11) and (2.14) to 0.25 to represent vague prior specifications. We also set $\beta_g = \beta$ for computational expedience. For further discussion on prior specifications, see Hu et al. (2018).

Conceptually, the latent household-level classes can be interpreted as clusters of households with similar compositions, e.g., households with children or households in which no one is related. Similarly, the latent individual-level classes can be interpreted as clusters of individuals with similar characteristics, e.g., older male spouses or young female children. However, for purposes of imputation, we do not care much about interpreting the classes, as they serve mainly to induce dependence across variables and individuals in the joint distribution.

It is important to select F and S to be large enough to ensure accurate estimation of the joint distribution. However, we also do not want to make F and S so large as to produce many empty classes in the model estimation. Allowing many empty classes increases computational running time without any corresponding increase in estimation accuracy. This can be especially problematic in the Gibbs sampler for the truncated NDPMPM, as these empty classes can introduce mass in regions of the space where impossible combinations are likely to be generated. This slows down the convergence of the Gibbs sampler.

We therefore recommend following the strategy in Hu et al. (2018) when setting (F, S). Analysts can start with moderate values for both, say between 10 and 15, in initial tuning runs. After convergence, analysts examine posterior samples of the latent classes to check how many individual-level and household-level latent classes are occupied. Such posterior predictive checks can provide evidence for the case that larger values for F and S are needed. If the numbers of occupied household-level classes hits F, we suggest increasing F. If the number of occupied individual-level classes hits S, we suggest increasing F first but then increasing S, possibly in addition to F, if increasing F alone does not suffice. When posterior predictive checks do not provide evidence that larger values of F and S are needed, analysts need not increase the number of classes, as doing so is not expected to improve the accuracy of the estimation. We note that similar logic is used in other mixture model contexts (Walker, 2007; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014; Murray and Reiter, 2016).

2.2 MCMC sampler for the NDPMPM

Hu et al. (2018) use a data augmentation strategy (Manrique-Vallier and Reiter, 2014) to estimate the posterior distribution in (2.2). They assume that the observed data $\mathcal{X}^1$, which includes only feasible households, is a subset from a hypothetical sample $\mathcal{X}$ of $(n + n_0)$ households directly generated from the untruncated NDPMPM. That is, $\mathcal{X}$ is generated on the support $\mathcal{C}$ where all combinations are possible and structural zeros rules are not enforced, but we only observe the sample of n households $\mathcal{X}^1$ that satisfy the structural zero rules and do not observe the sample of n_0 households $\mathcal{X}^0 = \mathcal{X} - \mathcal{X}^1$ that fail the rules.

We use the strategy of Hu et al. (2018) and augment the data as follows. For each $h \in \mathcal{H}$, we simulate $\mathcal{X}$ from the untruncated NDPMPM, stopping when the number of simulated feasible households in $\mathcal{X}$

directly matches n_{1h} for all $h \in \mathcal{H}$. We replace the simulated feasible households in $\mathcal{X}$ with $\mathcal{X}^1$, thus, assuming that $\mathcal{X}$ already contains $\mathcal{X}^1$ and we only need to generate the part $\mathcal{X}^0$ that fall in $\mathcal{S}$. Given a draw of $\mathcal{X}$, we draw θ from posterior distribution defined by the untruncated NDPMPM, treating $\mathcal{X}$ as the observed data. This posterior distribution can be estimated using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013).

We now present the full MCMC sampler for fitting the truncated NDPMPM. Let $\mathbf{G}^0$ and $\mathbf{M}^0$ be vectors of the latent class membership indicators for the households in $\mathcal{X}^0$ and n_{0h} be the number of households of size h in $\mathcal{X}^0$, with $n_0 = \sum_h n_{0h}$. In each full conditional, let "–" represent conditioning on all other variables and parameters in the model. At each MCMC iteration, we do the following steps.

S1. Set $\mathcal{X}^0 = \mathbf{G}^0 = \mathbf{M}^0 = \varnothing$. For each $h \in \mathcal{H}$, repeat the following:

(a) Set $t_0 = 0$ and $t_1 = 0$.

(b) Sample $G_i^0 \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^{**}, \ldots, \pi_F^{**})$ where $\pi_g^{**} \propto \lambda_{gh}^{(k)} \pi_g$ and k is the index for the household-level variable "household size".

(c) For $j = 1, \ldots, h$, sample $M_{ij}^0 \in \{1, \ldots, S\} \sim \text{Discrete}\left(\omega_{G_i^0 1}, \ldots, \omega_{G_i^0 S}\right)$.

(d) Set $X_{ik}^0 = h$, where X_{ik}^0 corresponds to the variable for household size. Sample the remaining household-level and individual-level values using the likelihoods in (2.3) and (2.4). Set the household's simulated value to $\mathbf{X}_i^0$.

(e) If $\mathbf{X}_i^0 \in \mathcal{S}_h$, let $t_0 = t_0 + 1$, $\mathcal{X}^0 = \mathcal{X}^0 \cup \mathbf{X}_i^0$, $\mathbf{G}^0 = \mathbf{G}^0 \cup G_i^0$ and $\mathbf{M}^0 = \mathbf{M}^0 \cup \{M_{i1}^0, \ldots, M_{ih}^0\}$. Otherwise set $t_1 = t_1 + 1$.

(f) If $t_1 < n_{1h}$, return to step (b). Otherwise, set $n_{0h} = t_0$.

S2. For observations in $\mathcal{X}^1$,

(a) Sample $G_i \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^*, \ldots, \pi_F^*)$ for $i = 1, \ldots, n$, where

$$\pi_g^* = \Pr(G_i = g \mid -) = \frac{\pi_g \left[\prod_{k=p+1}^{q} \lambda_{gX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{gmX_{ijk}^1}^{(k)}\right)\right]}{\sum_{f=1}^{F} \pi_f \left[\prod_{k=p+1}^{q} \lambda_{fX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{fmX_{ijk}^1}^{(k)}\right)\right]}$$

for $g = 1, \ldots, F$. Set $G_i^1 = G_i$.

(b) Sample $M_{ij} \in \{1, \ldots, S\} \sim \text{Discrete}\left(\omega_{G_i^1 1}^*, \ldots, \omega_{G_i^1 S}^*\right)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, where

$$\omega_{G_i^1 m}^* = \Pr\left(M_{ij} = m \mid -\right) = \frac{\omega_{G_i^1 m} \prod_{k=1}^{p} \phi_{G_i^1 m X_{ijk}^1}^{(k)}}{\sum_{s=1}^{S} \omega_{G_i^1 s} \prod_{k=1}^{p} \phi_{G_i^1 s X_{ijk}^1}^{(k)}}$$

for $m = 1, \ldots, S$. Set $M_{ij}^1 = M_{ij}$.

S3. Set $u_F = 1$. Sample

$$u_g \,|- \sim \text{Beta}\left(1 + U_g,\, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g}\left(1 - u_f\right)$$

where

$$U_g = \sum_{i=1}^{n} 1\left(G_i^1 = g\right) + \sum_{i=1}^{n_0} 1\left(G_i^0 = g\right)$$

for $g = 1, \ldots, F-1$.

S4. Set $v_{gM} = 1$ for $g = 1, \ldots, F$. Sample

$$v_{gm} \,|- \sim \text{Beta}\left(1 + V_{gm},\, \beta + \sum_{s=m+1}^{S} V_{gs}\right), \quad \omega_{gm} = v_{gm} \prod_{s<m}\left(1 - v_{gs}\right)$$

where

$$V_{gm} = \sum_{i=1}^{n} 1\left(M_{ij}^1 = m,\, G_i^1 = g\right) + \sum_{i=1}^{n_0} 1\left(M_{ij}^0 = m,\, G_i^0 = g\right)$$

for $m = 1, \ldots, S-1$ and $g = 1, \ldots, F$.

S5. Sample

$$\lambda_g^{(k)} \,|- \sim \text{Dirichlet}\left(1 + \eta_{g1}^{(k)}, \ldots, 1 + \eta_{gd_k}^{(k)}\right)$$

where

$$\eta_{gc}^{(k)} = \sum_{i \,|\, G_i^1 = g}^{n} 1\left(X_{ik}^1 = c\right) + \sum_{i \,|\, G_i^0 = g}^{n_0} 1\left(X_{ik}^0 = c\right)$$

for $g = 1, \ldots, F$ and $k = p+1, \ldots, q$.

S6. Sample

$$\phi_{gm}^{(k)} \,|- \sim \text{Dirichlet}\left(1 + \nu_{gm1}^{(k)}, \ldots, 1 + \nu_{gmd_k}^{(k)}\right)$$

where

$$\nu_{gmc}^{(k)} = \sum_{i,\, j \,|\, G_i^1 = g,\, M_{ij}^1 = m}^{n} 1\left(X_{ijk}^1 = c\right) + \sum_{i,\, j \,|\, G_i^0 = g,\, M_{ij}^0 = m}^{n_0} 1\left(X_{ijk}^0 = c\right)$$

for $g = 1, \ldots, F$, $m = 1, \ldots, S$ and $k = 1, \ldots, p$.

S7. Sample

$$\alpha \,|- \sim \text{Gamma}\left(a_\alpha + F - 1,\, b_\alpha - \sum_{g=1}^{F-1} \log\left(1 - u_g\right)\right).$$

S8. Sample

$$\beta \mid - \sim \text{Gamma}\left(a_\beta + F \times (S-1),\, b_\beta - \sum_{m=1}^{S-1}\sum_{g=1}^{F} \log\left(1 - v_{gm}\right)\right).$$

This Gibbs sampler is implemented in the R software package "NestedCategBayesImpute" (Wang, Akande, Hu, Reiter and Barrientos, 2016). The software can be used to generate synthetic versions of the original data, but it requires all data to be complete.

3 Handling missing data using the NDPMPM

We modify the Gibbs sampler for the truncated NDPMPM to incorporate missing data. For $i = 1, \ldots, n$, let $\mathbf{a}_i = \left(a_{i(p+1)}, \ldots, a_{i(p+q)}\right)$ be a vector with $a_{ik} = 1$ when household-level variable $k \in \{p+1, \ldots, p+q\}$ in $\mathbf{X}_i^1$ is missing, and $a_{ik} = 0$ otherwise. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let $\mathbf{b}_{ij} = \left(b_{ij1}, \ldots, b_{ijp}\right)$ be a vector with $b_{ijk} = 1$ when individual-level variable $k \in \{1, \ldots, p\}$ for individual $j \in \{1, \ldots, n_i\}$ in $\mathbf{X}_i^1$ is missing, and $b_{ijk} = 0$ otherwise. For each household i, let $\mathbf{X}_i^1 = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}})$, where $\mathbf{X}_i^{\text{obs}}$ comprise all data values corresponding to $a_{ik} = 0$ and $b_{ijk} = 0$, and $\mathbf{X}_i^{\text{mis}}$ comprises all data values corresponding to $a_{ik} = 1$ and $b_{ijk} = 1$. We assume that the data are missing at random (Rubin, 1976).

To incorporate missing values in the Gibbs sampler, we need to sample from the full conditional of each variable in $\mathbf{X}_i^{\text{mis}}$, conditioned on the variables for which $a_{ik} = 0$ and $b_{ijk} = 0$, at every iteration. Thus, we add the ninth step,

S9. For $i = 1, \ldots, n$, sample $\mathbf{X}_i^{\text{mis}}$ from its full conditional distribution

$$\Pr\left(\mathbf{X}_i^{\text{mis}} \mid -\right) \propto 1\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}\left(\pi_{G_i^1} \prod_{k \mid a_{ik}=1}^{p+q} \lambda^{(k)}_{G_i^1 X_{ik}^1} \prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k \mid b_{ijk}=1}^{p} \phi^{(k)}_{G_i^1 M_{ij}^1 X_{ijk}^1}\right).$$

Sampling from this conditional distribution is nontrivial because of the dependence among variables induced by the structural zero rules in each $\mathcal{S}_h$. Because of the dependence, we cannot simply sample each variable independently using the likelihoods in (2.3) and (2.4). If we could generate the set of all possible completions for all households with missing entries, conditional on the observed values, then calculating the probability of each one and sampling from the set would be straightforward. Unfortunately, this approach is not practical when the size of each $\mathcal{S}_h$ is large. Even when the size of each $\mathcal{S}_h$ is modest, each household could have different sets of completions, necessitating significant computing, storage, and memory requirements.

However, the full conditional in S9 takes a similar form as the kernel of the truncated NDPMPM in (2.1), so that we can generate the desired samples through a second rejection sampling scheme. Essentially, we sample from an untruncated version of the full conditional $P^*_{\mathbf{X}_i^{\text{mis}}} = \pi_{G_i^1} \prod_{k \mid a_{ik}=1}^{p+q} \lambda^{(k)}_{G_i^1 X_{ik}^1} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k \mid b_{ijk}=1}^{p} \phi^{(k)}_{G_i^1 M_{ij}^1 X_{ijk}^1}\right)$, until we obtain a valid sample that satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$;

see the Appendix for a proof that this rejection sampling scheme results in a valid Gibbs sampler. Notice that since $P^*_{\mathbf{X}_i^{\text{mis}}}$ itself is untruncated, we can generate samples from it by sampling each variable independently using (2.3) and (2.4). We therefore replace step S9 with S9'.

S9'. For $i = 1, \ldots, n$, sample $\mathbf{X}_i^{\text{mis}}$ as follows.

(a) For each missing household-level variable, that is, each variable where $k \in \{p + 1, \ldots, p + q\}$ with $a_{ik} = 1$, sample X^1_{ik} using (2.3).

(b) For each missing individual-level variable, that is, each variable where $j = 1, \ldots, n_i$ and $k \in \{1, \ldots, p\}$ with $b_{ijk} = 1$, sample X^1_{ijk} using (2.4).

(c) Set the sampled household-level and individual-level values to $\mathbf{X}_i^{\text{mis}*}$.

(d) Combine $\mathbf{X}_i^{\text{mis}*}$ with the observed $\mathbf{X}_i^{\text{obs}}$, that is, set $\mathbf{X}_i^{1*} = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}*})$. If $\mathbf{X}_i^{1*} \notin \mathcal{S}_h$, set $\mathbf{X}_i^{\text{mis}} = \mathbf{X}_i^{\text{mis}*}$, otherwise, return to step (9'a).

To initialize each $\mathbf{X}_i^{\text{mis}}$, we suggest sampling from the empirical marginal distribution of each variable k using the available cases for each variable, and requiring that the household satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$.

4 Strategies for speeding up the MCMC sampler

The rejection sampling step in the Gibbs sampler in Section 2.2 can be inefficient when $\mathcal{S}$ is large (Manrique-Vallier and Reiter, 2014; Hu et al., 2018), as the sampler tends to generate many impossible households before getting enough feasible ones. In addition, it takes computing time to check whether or not each sampled household satisfies all the structural zero rules. These computational costs are compounded when the sampler also incorporates missing values. In this section, we present two strategies that can reduce the number of impossible households that the algorithm generates, thereby speeding up the sampler. The Appendix includes simulation studies showing that both strategies can speed up the MCMC significantly.

4.1 Moving the household head to the household level

Many datasets include a variable recording the relationship of each individual to the household head. There can be only one household head in any household. This restriction can account for a large proportion of the combinations in $\mathcal{S}$. As a simple working example, consider a dataset that contains $n = 1{,}000$ households of size two, resulting in a total of $N = 2{,}000$ individuals. Suppose the data contain no household-level variables and two individual-level variables, age and relationship to household head. Also, suppose age has 100 levels while relationship to household head has 13 levels, which include household head, spouse of the household head, etc. Then, $\mathcal{C}$ contains $13^2 \times 100^2 = 1.69 \times 10^6$ combinations. Suppose the rule, "each household must contain exactly one head", is the only structural zero rule defined on the dataset. Then, $\mathcal{S}$ contains 1.45×10^6 impossible combinations, approximately 86% the size of $\mathcal{C}$. If, for example, the model assigns uniform probability to all combinations in $\mathcal{C}$, we would expect to sample about

$(0.86/0.14) * 1{,}000 \approx 6{,}143$ impossible households at every iteration to augment the n feasible households.

Instead, we treat the variables for the household head as a household-level characteristic. This eliminates structural zero rules defined on the household head alone. Using the working example, moving the household head to the household level results in one new household-level variable, age of household head, which has 100 levels. The relationship to household head variable can be ignored for household heads. For others in the household, the relationship to household head variable now has 12 levels, with the level corresponding to "household head" removed. Thus, $\mathcal{C}$ contains $12 \times 100^2 = 1.20 \times 10^5$ combinations, and $\mathcal{S}$ contains zero impossible combinations. We wouldn't even need to sample impossible households in the Gibbs sampler in Section 2.2.

In general, this strategy can reduce the size of $\mathcal{S}$ significantly, albeit usually not to zero as in the simple example here since $\mathcal{S}$ usually contains combinations resulting from other types of structural zero rules. This strategy is not a replacement for the rejection sampler in Section 2.2; rather, it is a data reformatting technique that can be combined with the sampler.

4.2 Setting an upper bound on the number of impossible households to sample

To reduce computation time, we can put an upper bound on the number of sampled cases in $\mathcal{X}^0$. One way to achieve this is to replace n_{1h} in step S1(f) of Section 2.2 with $\lceil n_{1h} \times \psi_h \rceil$, for some ψ_h such that $1/\psi_h$ is a positive integer, so that we sample only approximately $\lceil n_{0h} \times \psi_h \rceil$ impossible households for each $h \in \mathcal{H}$. However, doing so underestimates the actual probability mass assigned to $\mathcal{S}$ by the model. We can illustrate this using the simple example of Section 4.1. Suppose the model assigns uniform probability to all combinations in $\mathcal{C}$ as before. We set $\psi_2 = 0.5$, so that we sample approximately $3{,}072 = \lceil 6{,}143 \times 0.5 \rceil$ impossible households in every iteration of the MCMC sampler. The probability of generating one impossible household is $3{,}072/(1{,}000 + 3{,}072) = 0.75$, a decrease from the actual value of 0.86. Therefore, we would underestimate the true contribution of $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ to the likelihood.

To use the cap-and-weight approach, we need to apply a correction that re-weights the contribution of $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ to the full joint likelihood. We do so using ideas akin to those used by Chambers and Skinner (2003) and Savitsky and Toth (2016), approximating the likelihood of the full unobserved data with a "pseudo" likelihood using weights (the $1/\psi_h$'s). The impossible households only contribute to the full joint likelihood through the discrete distributions in (2.3) to (2.6). The sufficient statistics for estimating the parameters of the discrete distributions in (2.3) to (2.6) are the observed counts for the corresponding variables in the set $\{\mathcal{X}^1, \mathbf{G}^1, \mathbf{M}^1, \mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$, within each latent class for the household-level variables and within each latent class pair for the individual-level variables. Thus, for each $h \in \mathcal{H}$, we can re-weight the contribution of impossible households by multiplying the observed counts for households of size h in $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ by $1/\psi_h$ for the corresponding variable and latent classes. This raises the likelihood contribution of impossible households of size h to the power of $1/\psi_h$. Clearly, $1/\psi_h$ need not be a positive integer. We require that only to make its multiplication with the observed counts free of decimals. We

modify the Gibbs sampler to incorporate the cap-and-weight approach by replacing steps S1, S3, S4, S5 and S6; see the Appendix for the modified steps.

Setting each $\psi_h = 1$ corresponds to the original rejection sampler, so that the two approaches should provide very similar results when ψ_h near 1. Based on our experience, results of the cap-and-weight approach become significantly less accurate than the regular rejection sampler when $\psi_h < 1/4$. The time gained using this speedup approach in comparison to the regular sampler depends on the features of the data and the specified values for the weights $\{\psi_h: h \in \mathcal{H}\}$. To select the ψ_h's, we suggest trying out different values – starting with values close to one – in initial runs of the MCMC sampler on a small random sample of the data. Analysts should examine the convergence and mixing behavior of the chains in comparison to the chain with all the ψ_h's set to one, and select values that offer reasonable speedup while preserving convergence and mixing. This can be done quickly by comparing trace plots of a random set of parameters from the model that are not subject to label switching, such as α and β, or by examining marginal, bivariate and trivariate probabilities estimated from synthetic data generated from the MCMC.

5 Empirical study

To evaluate the performance of the NDPMPM as an imputation method, as well as the speed up strategies, we use data from the public use microdata files from the 2012 ACS, available for download from the United States Census Bureau (http://www2.census.gov/acs2012_1yr/pums/). We construct a population of 764,580 households of sizes $\mathcal{H} = \{2, 3, 4\}$, from which we sample $n = 5{,}000$ households comprising $N = 13{,}181$ individuals. We work with the variables described in Table 5.1, which mimic those in the U.S. decennial census. The structural zeros involve ages and relationships of individuals in the same house; see the Appendix for a full list of rules that we used. We move the household head to the household level as in Section 4.1 to take advantage of the computational gains.

We introduce missing values using the following scenario. We let household size and age of household heads be fully observed. We randomly and independently blank 30% of each variable for the remaining household-level variables. For individuals other than the household head, we randomly and independently blank 30% of the values for gender, race and Hispanic origin. We make age missing with rates 50%, 20%, 40% and 30% for values of the relationship variable in the sets $\{2\}$, $\{3, 4, 5, 10\}$, $\{7, 9\}$ and $\{6, 8, 11, 12, 13\}$, respectively. We make the relationship variable missing with rates 40%, 25%, 10%, and 55% for values of age in the sets $\{x: x \leq 20\}$, $\{x: 20 < x \leq 50\}$, $\{x: 50 < x \leq 70\}$, and $\{x: x > 70\}$, respectively. This results in approximately 30% missing values for both variables. About 8% of the individuals in the sample are missing both the age and relationship variable, and 2% are missing gender, age, and relationship jointly. This mechanism results in data that technically are not missing at random, but we use the NDPMPM approach regardless to examine its potential in a complicated missingness mechanism. Actual rates of item nonresponse in census data tend to be smaller than what we use here, but we use high rates to put the NDPMPM through a challenging stress test. We also introduce missing values

using a missing completely at random scenario with rates in the 10% range across all the variables. In short, the results are similar to those here, though more accurate due to the lower rates of missingness. See the Appendix for the results.

Table 5.1
Description of variables used in the study. "HH" means household head

Description of variable		Categories
Household-level variables	Ownership of dwelling	1 = owned or being bought, 2 = rented
	Household size	2 = 2 people, 3 = 3 people, 4 = 4 people
	Gender of HH	1 = male, 2 = female
	Race of HH	1 = white, 2 = black, 3 = American Indian or Alaska native, 4 = Chinese, 5 = Japanese, 6 = other Asian/Pacific islander, 7 = other race, 8 = two major races, 9 = three or more major races
	Hispanic origin of HH	1 = not Hispanic, 2 = Mexican, 3 = Puerto Rican, 4 = Cuban, 5 = other
	Age of HH	1 = less than one year old, 2 = 1 year old, 3 = 2 years old, ..., 96 = 95 years old
Individual-level variables	Gender	same as "Gender of HH"
	Race	same as "Race of HH"
	Hispanic origin	same as "Hispanic origin of HH"
	Age	same as "Age of HH"
	Relationship to head of household	1 = spouse, 2 = biological child, 3 = adopted child, 4 = stepchild, 5 = sibling, 6 = parent, 7 = grandchild, 8 = parent-in-law, 9 = child-in-law, 10 = other relative, 11 = boarder, roommate or partner, 12 = other non-relative or foster child

We estimate the NDPMPM using two approaches, both using the rejection step S9' in Section 3. The first approach considers $\psi_2 = \psi_3 = \psi_4 = 1$, i.e., without using the cap-and-weight approach, while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = 1/3$. For each approach, we run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in and thinning the remaining samples every five iterations, resulting in 1,000 MCMC post burn-in iterates. We set $F = 30$ and $S = 15$ for each approach based on initial tuning runs. Across the approaches, the effective number of occupied household-level clusters usually ranges from 13 to 16 with a maximum of 25, while the effective number of occupied individual-level clusters across all household-level clusters ranges from 3 to 5 with a maximum of 10. For convergence, we examined trace plots of α, β, and weighted averages of a random sample of the multinomial probabilities in (2.3) and (2.4) (since the multinomial probabilities themselves are prone to label switching).

For both methods, we generate $L = 50$ completed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, using the posterior predictive distribution of the NDPMPM, from which we estimate all marginal distributions, bivariate distributions of all possible pairs of variables, and trivariate distributions of all possible triplets of variables.

We also estimate several probabilities that depend on within household relationships and the household head to investigate the performance of the NDPMPM in estimating complex relationships. We obtain confidence intervals using multiple imputation inferences (Rubin, 1987). As a brief review, let q be the completed-data point estimator of some estimand Q, and let u be the estimator of variance associated with q. For $l = 1, \ldots, L$, let $q^{(l)}$ and $u^{(l)}$ be the values of q and u in completed dataset $\mathbf{Z}^{(l)}$. We use $\overline{q}_L = \sum_{l=1}^{L} q^{(l)} \big/ L$ as the point estimate of Q. We use $T_L = (1 + 1/L)\, b_L + \overline{u}_L$ as the estimated variance of $\overline{q}$, where $b_L = \sum_{l=1}^{L} (q^{(l)} - \overline{q}_L)^2 \big/ (L-1)$ and $\overline{u}_L = \sum_{l=1}^{L} u^{(l)} \big/ L$. We make inference about Q using $(\overline{q}_L - Q) \sim t_v\,(0, T_L)$, where t_v is a $t-$distribution with $v = (L-1)\,(1 + \overline{u}_L \big/ [(1 + 1/L)\, b_L])^2$ degrees of freedom.

Figures 5.1 and 5.2 display the value of $\overline{q}_{50}$ for each estimated marginal, bivariate and trivariate probability plotted against its corresponding estimate from the original data, without missing values. Figure 5.1 shows the results for the NDPMPM with the rejection sampler, and Figure 5.2 shows the results for the NDPMPM using the cap-and-weight approach. For both approaches, the point estimates are close to those from the data before introducing missing values, suggesting that the NDPMPM does a good job of capturing important features of the joint distribution of the variables. Figure 5.2 in particular also shows that the cap-and-weight approach did not degrade the estimates.

Table 5.2 displays 95% confidence intervals for several probabilities involving within-household relationships, as well as the value in the full population of 764,580 households. The intervals include the two based on the NDPMPM imputation engines and the interval from the data before introducing missingness. For the latter, we use the usual Wald interval, $\hat{p} \pm 1.96 \sqrt{\hat{p}\,(1 - \hat{p})/n}$, where $\hat{p}$ is the corresponding sample percentage. For the most part, the intervals from the NDPMPM with the full rejection sampling are close to those based on the data without any missingness. They tend to include the true population quantity. The NDPMPM imputation engine results in noticeable downward bias for the percentages of households where everyone is the same race, with bias increasing as the household size gets bigger. This is a challenging estimand to estimate accurately via imputation, particularly for larger households. Hu et al. (2018) identified biases in the same direction when using the NDPMPM (with household head data treated as individual-level variables) to generate fully synthetic data, noting that the bias gets smaller as the sample size increases. The NDPMPM fits the joint distribution of the data better and better as the sample size grows. Hence, we expect the NDPMPM imputation engine to be more accurate with larger sample sizes, as well as with smaller fractions of missing values.

The interval estimates from the cap-and-weight method are generally similar to those for the full rejection sampler, with some degradation particularly for the percentages of same race households by household size. This degradation comes with a benefit, however. Based on MCMC runs on a standard laptop, the NDPMPM using the cap-and-weight approach and moving household heads' data values to the household level is about 42% faster than the NDPMPM with household heads' data values moved to the household level.

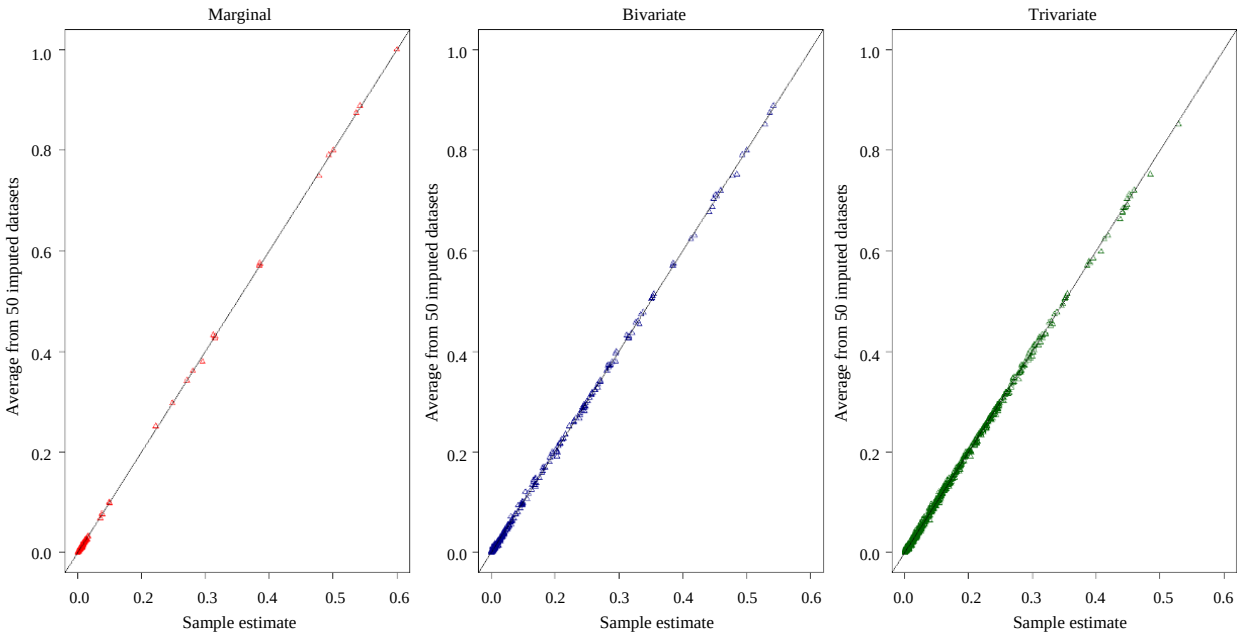


Figure 5.1 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.

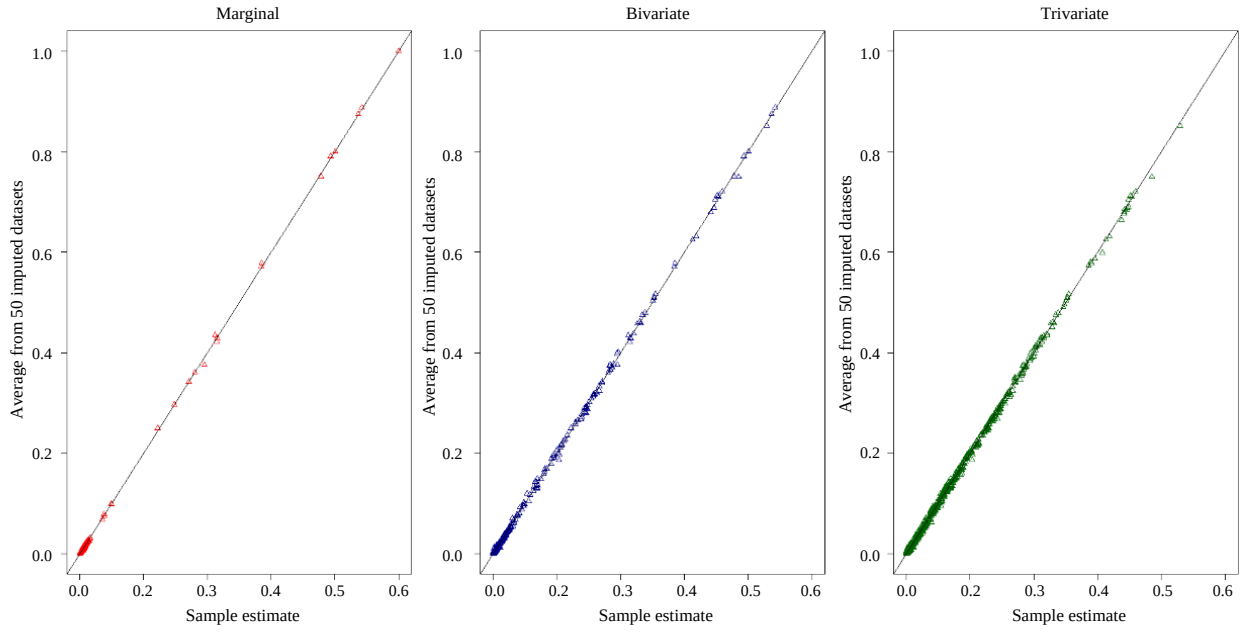


Figure 5.2 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.

Table 5.2
Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "No missing" is based on the sampled data before introducing missing values, "NDPMPM" uses the truncated NDPMPM, moving household heads' data values to the household level, and "NDPMPM Capped" uses the truncated NDPMPM with the cap-and-weight approach and moving household heads' data values to the household level. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. *Q* is the value in the full population of 764,580 households

		Q	No Missing	NDPMPM	NDPMPM Capped
All same race household:	$n_i = 2$	0.942	(0.932, 0.949)	(0.891, 0.917)	(0.884, 0.911)
	$n_i = 3$	0.908	(0.907, 0.937)	(0.843, 0.890)	(0.821, 0.870)
	$n_i = 4$	0.901	(0.879, 0.917)	(0.793, 0.851)	(0.766, 0.828)
SP present		0.696	(0.682, 0.707)	(0.695, 0.722)	(0.695, 0.722)
Same race CP		0.656	(0.641, 0.668)	(0.640, 0.669)	(0.634, 0.664)
SP present, HH is White		0.600	(0.589, 0.616)	(0.603, 0.632)	(0.604, 0.634)
White CP		0.580	(0.569, 0.596)	(0.577, 0.606)	(0.574, 0.604)
CP with age difference less than five		0.488	(0.465, 0.492)	(0.341, 0.371)	(0.324, 0.355)
Male HH, home owner		0.476	(0.456, 0.484)	(0.450, 0.479)	(0.451, 0.480)
HH over 35, no CH present		0.462	(0.441, 0.468)	(0.442, 0.470)	(0.443, 0.471)
At least one biological CH present		0.437	(0.431, 0.458)	(0.430, 0.459)	(0.428, 0.456)
HH older than SP, White HH		0.322	(0.309, 0.335)	(0.307, 0.339)	(0.311, 0.343)
Adult female w/ at least one CH under 5		0.078	(0.070, 0.085)	(0.062, 0.078)	(0.061, 0.077)
White HH with Hisp origin		0.066	(0.064, 0.078)	(0.062, 0.079)	(0.062, 0.078)
Non-White CP, home owner		0.058	(0.050, 0.063)	(0.038, 0.052)	(0.037, 0.051)
Two generations present, Black HH		0.057	(0.053, 0.066)	(0.052, 0.066)	(0.052, 0.067)
Black HH, home owner		0.052	(0.046, 0.058)	(0.044, 0.058)	(0.044, 0.059)
SP present, HH is Black		0.039	(0.032, 0.042)	(0.032, 0.044)	(0.031, 0.043)
White-nonwhite CP		0.034	(0.029, 0.039)	(0.038, 0.053)	(0.043, 0.059)
Hisp HH over 50, home owner		0.029	(0.025, 0.034)	(0.023, 0.034)	(0.024, 0.034)
One grandchild present		0.028	(0.023, 0.033)	(0.024, 0.035)	(0.023, 0.035)
Adult Black female w/ at least one CH under 18		0.027	(0.028, 0.038)	(0.025, 0.036)	(0.025, 0.036)
At least two generations present, Hisp CP		0.027	(0.022, 0.031)	(0.022, 0.032)	(0.023, 0.033)
Hisp CP with at least one biological CH		0.025	(0.020, 0.028)	(0.019, 0.029)	(0.020, 0.030)
At least three generations present		0.023	(0.020, 0.028)	(0.017, 0.026)	(0.017, 0.026)
Only one parent		0.020	(0.016, 0.024)	(0.013, 0.021)	(0.013, 0.021)
At least one stepchild		0.019	(0.018, 0.026)	(0.019, 0.030)	(0.019, 0.030)
Adult Hisp male w/ at least one CH under 10		0.018	(0.017, 0.025)	(0.014, 0.022)	(0.014, 0.022)
At least one adopted CH, White CP		0.008	(0.005, 0.010)	(0.004, 0.010)	(0.004, 0.011)
Black CP with at least two biological children		0.006	(0.003, 0.007)	(0.003, 0.007)	(0.003, 0.007)
Black HH under 40, home owner		0.005	(0.005, 0.009)	(0.006, 0.013)	(0.007, 0.013)
Three generations present, White CP		0.005	(0.004, 0.008)	(0.004, 0.010)	(0.004, 0.009)
White HH under 25, home owner		0.003	(0.002, 0.005)	(0.003, 0.007)	(0.003, 0.007)

6 Discussion

The empirical study suggests that the NDPMPM can provide high quality imputations for categorical data nested within households. To our knowledge, this is the first parametric imputation engine for nested multivariate categorical data. The study also illustrates that, with modest sample sizes, agencies should not expect the NDPMPM to preserve all features of the joint distribution. Of course, this is the case with any imputation engine. For the NDPMPM, agencies may be able to improve accuracy for targeted quantities by recoding the data used to fit the model. For example, one can create a new household-level variable that equals one when everyone has the same race and equals zero otherwise, and replace the individual race variable with a new variable that has levels "1 = race is the same as race of household head", "2 = race is white and differs from race of household head", "3 = race is black and differs from race of household head", and so on. The NDPMPM would be estimated with the household-level same race variable and the new individual-level race variable. This would encourage the NDPMPM to estimate the percentages with the

same race very accurately, as it would be just another household-level variable like home ownership. It also would add structural zeros involving race to the computation. Evaluating the trade offs in accuracy and computational costs of such recodings is a topic for future research.

The NDPMPM can be computationally expensive, even with the speed-ups presented in this article. The expensive parts of the algorithm are the rejection sampling steps. Fortunately, these can be done easily by parallel processing. For example, we can require each processor to generate a fraction of the impossible cases in Section 2.2. We also can spread the rejection steps for the imputations over many processors. These steps should cut run time by a factor roughly equal to the number of processors available.

The empirical study used households up to size four. We have run the model on data with households up to size seven in reasonable time (a few hours on a standard laptop). Accuracy results are similar qualitatively. As the household sizes get large, the model can generate hundreds or even thousands times as many impossible households as there are feasible ones, slowing the algorithm. In such cases, the cap-and-weight approach is essential for practical applications.

Acknowledgements

This research was supported by grants from the National Science Foundation (NSF SES 1131897) and the Alfred P. Sloan Foundation (G-2-15-20166003).

Appendix

This is an Appendix to the paper. It contains proof that the rejection sampling step S9' in Section 3 generates samples from the correct posterior distribution. It also contains the modified Gibbs sampler for the cap-and-weight approach and a list of the structural zero rules used in fitting the NDPMPM model. Finally, we include empirical results for the speedup approaches mentioned in the paper, using synthetic data, and additional results for handling missing data using the NDPMPM under a missing completely at random scenario.

A.1 Proof that the rejection sampling step S9' in Section 3 generates samples from the correct posterior distribution

The X_{ik}^{1} and X_{ijk}^{1} values generated using the rejection sampler in Step S9' are generated from the full conditionals, resulting in a valid Gibbs sampler. The proof follows from the properties of rejection sampling (or simple accept reject). The target distribution is the full conditional for $\mathbf{X}_i^{\text{mis}}$. It can be re-expressed as

$$p\left(\mathbf{X}_i^{\text{mis}}\right) = \frac{1\left\{\mathbf{X}_i^{1} \notin \mathcal{S}_h\right\}}{\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right)} g\left(\mathbf{X}_i^{\text{mis}}\right)$$

where

$$g\left(\mathbf{X}_i^{\text{mis}}\right) = \pi_{G_i^1} \prod_{k\,|\,a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k\,|\,b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)} \right).$$

Our rejection scheme uses $g\left(\mathbf{X}_i^{\text{mis}}\right)$ as a proposal for $p\left(\mathbf{X}_i^{\text{mis}}\right)$. To show that the draws are indeed from $p\left(\mathbf{X}_i^{\text{mis}}\right)$, we need to verify that $w\left(\mathbf{X}_i^{\text{mis}}\right) = p\left(\mathbf{X}_i^{\text{mis}}\right) / g\left(\mathbf{X}_i^{\text{mis}}\right) < M$, where $1 < M < \infty$, and that we are accepting each sample with probability $w\left(\mathbf{X}_i^{\text{mis}}\right) / M$. In our case,

1. $w\left(\mathbf{X}_i^{\text{mis}}\right) = p\left(\mathbf{X}_i^{\text{mis}}\right) / g\left(\mathbf{X}_i^{\text{mis}}\right) = 1\left\{\mathbf{X}_i^1 \notin \mathcal{S}_h\right\} / \Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \,|\, \theta\right) \leq 1 / \Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \,|\, \theta\right)$, and $0 < \Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \,|\, \theta\right) < 1 \Rightarrow 1 < 1 / \Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \,|\, \theta\right) < \infty$ necessarily.
2. By sampling until we obtain a valid sample that satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$, we are indeed sampling with probability $w\left(\mathbf{X}_i^{\text{mis}}\right) / M = 1\left\{\mathbf{X}_i^1 \notin \mathcal{S}_h\right\}$.

A.2 Modified Gibbs sampler for the cap-and-weight approach

The modified Gibbs sampler for the cap-and-weight approach replaces steps S1, S3, S4, S5 and S6 of the Gibbs sampler in the main text as follows.

S1*. For each $h \in \mathcal{H}$, repeat steps S1(a) to S1(e) as before but modify step S1(f) to: if $t_1 < \lceil n_{1h} \times \psi_h \rceil$, return to step (b). Otherwise, set $n_{0h} = t_0$.

S3*. Set $u_F = 1$. Sample

$$u_g \,|\, - \sim \text{Beta}\left(1 + U_g,\, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g} \left(1 - u_f\right)$$

where

$$U_g = \sum_{i=1}^{n} 1\left(G_i^1 = g\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i\,|\,n_i^0 = h} 1\left(G_i^0 = g\right)$$

for $g = 1, \ldots, F-1$.

S4*. Set $v_{gM} = 1$ for $g = 1, \ldots, F$. Sample

$$v_{gm} \,|\, - \sim \text{Beta}\left(1 + V_{gm},\, \beta + \sum_{s=m+1}^{S} V_{gs}\right), \quad \omega_{gm} = v_{gm} \prod_{s<m} \left(1 - v_{gs}\right)$$

where

$$V_{gm} = \sum_{i=1}^{n} 1\left(M_{ij}^1 = m,\, G_i^1 = g\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i\,|\,n_i^0 = h} 1\left(M_{ij}^0 = m,\, G_i^0 = g\right)$$

for $m = 1, \ldots, S-1$ and $g = 1, \ldots, F$.

S5*. Sample

$$\lambda_g^{(k)} \mid - \sim \text{Dirichlet}\left(1+\eta_{g1}^{(k)}, \ldots, 1+\eta_{gd_k}^{(k)}\right)$$

where

$$\eta_{gc}^{(k)} = \sum_{i \mid G_i^1 = g}^{n} 1\left(X_{ik}^{1} = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \mid n_i^0 = h,\, G_i^0 = g} 1\left(X_{ik}^{0} = c\right)$$

for $g = 1, \ldots, F$ and $k = p+1, \ldots, q$.

S6*. Sample

$$\phi_{gm}^{(k)} \mid - \sim \text{Dirichlet}\left(1+\nu_{gm1}^{(k)}, \ldots, 1+\nu_{gmd_k}^{(k)}\right)$$

where

$$\nu_{gmc}^{(k)} = \sum_{i \mid G_i^1 = g,\, M_{ij}^1 = m}^{n} 1\left(X_{ijk}^{1} = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \mid n_i^0 = h,\, G_i^0 = g,\, M_{ij}^0 = m} 1\left(X_{ijk}^{0} = c\right)$$

for $g = 1, \ldots, F$, $m = 1, \ldots, S$ and $k = 1, \ldots, p$.

A.3 List of structural zeros

We fit the NDPMPM model using structural zeros which involve ages and relationships of individuals in the same house. The full list of the rules used is presented in Table A.1. These rules were derived from the 2012 ACS by identifying combinations involving the relationship variable that do not appear in the constructed population. This list should not be interpreted as a "true" list of impossible combinations in census data.

Table A.1
List of structural zeros

Description
Rules common to generating both the synthetic and imputed datasets
1. Each household must contain exactly one head and he/she must be at least 16 years old.
2. Each household cannot contain more than one spouse and he/she must be at least 16 years old.
3. Married couples are of opposite sex, and age difference between individuals in the couples cannot exceed 49.
4. The youngest parent must be older than the household head by at least 4.
5. The youngest parent-in-law must be older than the household head by at least 4.
6. The age difference between the household head and siblings cannot exceed 37.
7. The household head must be at least 31 years old to be a grandparent and his/her spouse must be at least 17. Also, He/she must be older than the oldest grandchild by at least 26.
Rules specific to generating the synthetic datasets
8. The household head must be older than the oldest child by at least 7.
Rules specific to generating the imputed datasets
9. The household head must be older than the oldest biological child by at least 7.
10. The household head must be older than the oldest adopted child by at least 11.
11. The household head must be older than the oldest stepchild by at least 9.

A.4 Empirical study of the speedup approaches

We evaluate the performance of the two speedup approaches mentioned in the main text using synthetic data. We use data from the public use microdata files from the 2012 ACS, available for download from the United States Census Bureau (http://www2.census.gov/acs2012_1yr/pums/) to construct a population of 857,018 households of sizes $\mathcal{H} = \{2, 3, 4, 5, 6\}$, from which we sample $n = 10{,}000$ households comprising $N = 29{,}117$ individuals. We work with the variables described in Table A.2. We evaluate the approaches using probabilities that depend on within household relationships and the household head.

Table A.2
Description of variables used in the synthetic data illustration

Description of variable		Categories
Household-level variables	Ownership of dwelling	1 = owned or being bought, 2 = rented
	Household size	2 = 2 people, 3 = 3 people, 4 = 4 people, 5 = 5 people, 6 = 6 people
Individual-level variables	Gender	1 = male, 2 = female
	Race	1 = white, 2 = black, 3 = American Indian or Alaska native, 4 = Chinese, 5 = Japanese, 6 = other Asian/Pacific islander, 7 = other race, 8 = two major races, 9 = three or more major races
	Hispanic origin	1 = not Hispanic, 2 = Mexican, 3 = Puerto Rican, 4 = Cuban, 5 = other
	Age	1 = less than one year old, 2 = 1 year old, 3 = 2 years old, ..., 96 = 95 years old
	Relationship to head of household	1 = household head, 2 = spouse, 3 = child, 4 = child-in-law, 5 = parent, 6 = parent-in-law, 7 = sibling, 8 = sibling-in-law, 9 = grandchild, 10 = other relative, 11 = partner/friend/visitor, 12 = other non-relative

We consider the NDPMPM using two approaches, both moving the values of the household head to the household level as in Section 4.1 of the main text and also using the cap-and-weight approach in Section 4.2 of the main text. The first approach considers $\psi_2 = \psi_3 = \psi_4 = \psi_5 = \psi_6 = 1$ while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = \psi_5 = \psi_6 = 1/3$. We compare these approaches to the NDPMPM as presented in Hu et al., 2018. For each approach, we create $L = 50$ synthetic datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$. We generate the synthetic datasets so that the number of households of size $h \in \mathcal{H}$ in each $\mathbf{Z}^{(l)}$ exactly matches n_h from the observed data. Thus, $\mathbf{Z}$ comprises partially synthetic data (Little, 1993; Reiter, 2003), even though every released Z_{ijk} is a simulated value. We combine the estimates using using the approach in Reiter (2003). As a brief review, let q be the point estimator of some estimand Q, and let u be the estimator of variance associated with q. For $l = 1, \ldots, L$, let q_l and u_l be the values of q and u in synthetic dataset $\mathbf{Z}^{(l)}$. We use $\bar{q} = \sum_{l=1}^{L} q_l / L$ as the point estimate of Q and $T = \bar{u} + b/L$ as the estimated variance of $\bar{q}$, where $b = \sum_{l=1}^{L} (q_l - \bar{q})^2 / (L-1)$ and $\bar{u} = \sum_{l=1}^{L} u_l / L$. We make inference about Q using $(\bar{q} - Q) \sim t_v(0, T)$, where t_v is a $t-$distribution with $v = (L-1)(1 + L\bar{u}/b)^2$ degrees of freedom.

For each approach, we run the MCMC sampler for 20,000 iterations, discarding the first 10,000 as burn-in and thinning the remaining samples every five iterations, resulting in 2,000 MCMC post burn-in iterates. We create the $L = 50$ synthetic datasets by randomly sampling from the 2,000 iterates. We set $F = 40$ and $S = 15$ for each approach based on initial tuning runs. For convergence, we examined trace plots of α, β and weighted averages of a random sample of the multinomial probabilities in the NDPMPM likelihood. Across the approaches, the effective number of occupied household-level clusters usually ranges from 20 to 33 with a maximum of 38, while the effective number of occupied individual-level clusters across all household-level clusters ranges from 5 to 9 with a maximum of 12.

Based on MCMC runs on a standard laptop, moving household heads' data values to the household level alone results in a speedup of about 63% on the default rejection sampler while the cap-and-weight approach alone results in a speedup of about 40%.

Table A.3 shows the 95% confidence intervals for each approach. Essentially, all three approaches result in similar confidence intervals, suggesting not much loss in accuracy from the speedups. Most intervals also are reasonably similar to confidence intervals based on the original data, except for the percentage of same age couples. The last row is a rigorous test of how well each method can estimate a probability that can be fairly difficult to estimate accurately. In this case, the probability that a household head and spouse are the same age can be difficult to estimate since each individual's age can take 96 different values. All three approaches are thus off from the estimate from the original data in this case. These results suggest that we can significantly speedup the sampler with minimal loss in accuracy of estimates and confidence intervals of population estimands.

Table A.3
Confidence intervals for selected probabilities that depend on within-household relationships in the original and synthetic datasets. "Original" is based on the sampled data, "NDPMPM" is the default MCMC sampler described in Section 2.2 of the main text, "NDPMPM w/ HH moved" is the default sampler, moving household heads' data values to the household level, "NDPMPM capped w/ HH moved" uses the cap-and-weight approach and moving household heads' data values to the household level. "HH " means household head and "SP" means spouse

		Original	NDPMPM	NDPMPM w/ HH moved	NDPMPM capped w/ HH moved
All same race	$n_i = 2$	(0.939, 0.951)	(0.918, 0.932)	(0.912, 0.928)	(0.910, 0.925)
	$n_i = 3$	(0.896, 0.920)	(0.859, 0.888)	(0.845, 0.875)	(0.844, 0.874)
	$n_i = 4$	(0.885, 0.912)	(0.826, 0.860)	(0.813, 0.848)	(0.817, 0.852)
	$n_i = 5$	(0.879, 0.922)	(0.786, 0.841)	(0.786, 0.841)	(0.777, 0.834)
	$n_i = 6$	(0.831, 0.910)	(0.701, 0.803)	(0.718, 0.819)	(0.660, 0.768)
SP present		(0.693, 0.711)	(0.678, 0.697)	(0.676, 0.695)	(0.677, 0.695)
SP with white HH		(0.589, 0.608)	(0.577, 0.597)	(0.576, 0.595)	(0.575, 0.595)
SP with black HH		(0.036, 0.043)	(0.035, 0.043)	(0.034, 0.042)	(0.034, 0.042)
White couple		(0.570, 0.589)	(0.560, 0.580)	(0.553, 0.573)	(0.552, 0.572)
White couple, own		(0.495, 0.514)	(0.468, 0.488)	(0.461, 0.481)	(0.463, 0.483)
Same race couple		(0.655, 0.673)	(0.636, 0.655)	(0.626, 0.645)	(0.625, 0.644)
White-nonwhite couple		(0.028, 0.035)	(0.028, 0.035)	(0.034, 0.041)	(0.036, 0.044)
Nonwhite couple, own		(0.057, 0.067)	(0.047, 0.056)	(0.045, 0.053)	(0.045, 0.054)
Only mother present		(0.017, 0.022)	(0.014, 0.019)	(0.014, 0.019)	(0.013, 0.018)
Only one parent present		(0.021, 0.026)	(0.026, 0.032)	(0.026, 0.033)	(0.027, 0.033)
Children present		(0.507, 0.527)	(0.493, 0.512)	(0.517, 0.537)	(0.511, 0.531)
Siblings present		(0.022, 0.028)	(0.027, 0.034)	(0.027, 0.033)	(0.027, 0.033)
Grandchild present		(0.041, 0.049)	(0.051, 0.060)	(0.049, 0.058)	(0.050, 0.059)
Three generations present		(0.036, 0.044)	(0.037, 0.045)	(0.042, 0.050)	(0.040, 0.048)
White HH, older than SP		(0.309, 0.327)	(0.283, 0.301)	(0.294, 0.313)	(0.302, 0.321)
Nonhisp HH		(0.882, 0.894)	(0.875, 0.888)	(0.879, 0.891)	(0.876, 0.889)
White, Hisp HH		(0.071, 0.082)	(0.074, 0.085)	(0.072, 0.082)	(0.073, 0.084)
Same age couple		(0.087, 0.098)	(0.027, 0.034)	(0.023, 0.029)	(0.024, 0.031)

A.5 Empirical study of missing data imputation under MCAR

We also evaluate the performance of the NDPMPM as an imputation method under a missing completely at random (MCAR) scenario. We use the same data as in Section 5 of the main text. As a reminder, the data contains $n = 5{,}000$ households of sizes $\mathcal{H} = \{2, 3, 4\}$, comprising $N = 13{,}181$ individuals. We introduce missing values using a MCAR scenario. We randomly select 80% households to be complete cases for all variables. For the remaining 20%, we let the variable "household size" be fully observed and randomly – and independently – blank 50% of each variable for the remaining household-level and individual-level variables. We use these low rates to mimic the actual rates of item nonresponse in census data.

Similar to the main text, we estimate the NDPMPM using two approaches, both combining the rejection step in Section 4.1 of the main text with the cap-and-weight approach in Section 4.2 of the main text. The first approach considers $\psi_2 = \psi_3 = \psi_4 = 1$ while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = 1/3$. For each approach, we run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in and thinning the remaining samples every five iterations, resulting in 1,000 MCMC post burn-in iterates. We set $F = 30$ and $S = 15$ for each approach based on initial tuning runs. We monitor convergence as in the main text. For both methods, we generate $L = 50$ completed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, using the posterior predictive distribution of the NDPMPM, from which we estimate the same probabilities as in the main text.

Figures A.1 and A.2 display each estimated marginal, bivariate and trivariate probability $\overline{q}_{50}$ plotted against its corresponding estimate from the original data, without missing values. Figure A.1 shows the results for the NDPMPM with the rejection sampler, and Figure A.2 shows the results for the NDPMPM using the cap-and-weight approach. For both approaches, the NDPMPM does a good job of capturing important features of the joint distribution of the variables as the point estimates are very close to those from the data before introducing missing values. In short, the results are very similar to those in the main text, though more accurate.

Table A.4 displays 95% confidence intervals for selected probabilities involving within-household relationships, as well as the value in the full population of 764,580 households. The intervals include the two based on the NDPMPM imputation engines and the interval from the data before introducing missingness. The intervals are generally more accurate than those presented in the main text. This is expected since we use lower rates of missingness in the MCAR scenario. For the most part, the intervals from the NDPMPM with the two approaches tend to include the true population quantity. Again, the NDPMPM imputation engine results in downward bias for the percentages of households where everyone is the same race. As mentioned in the main text, this is a challenging estimand to estimate accurately via imputation, particularly for larger households.

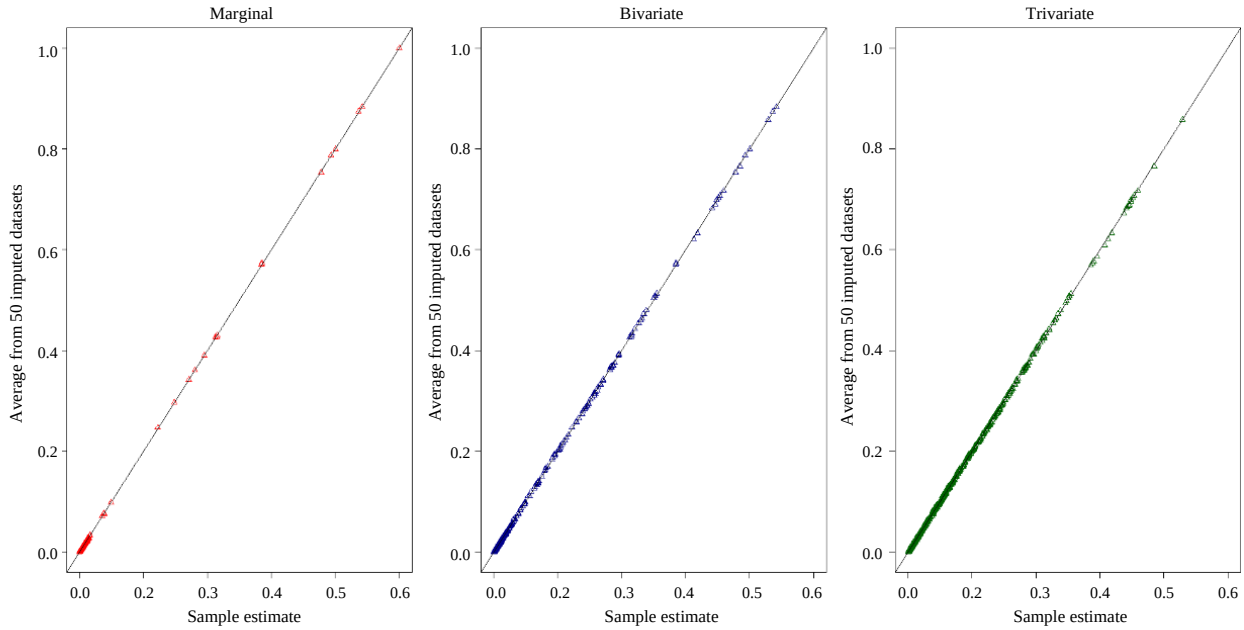


Figure A.1 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.

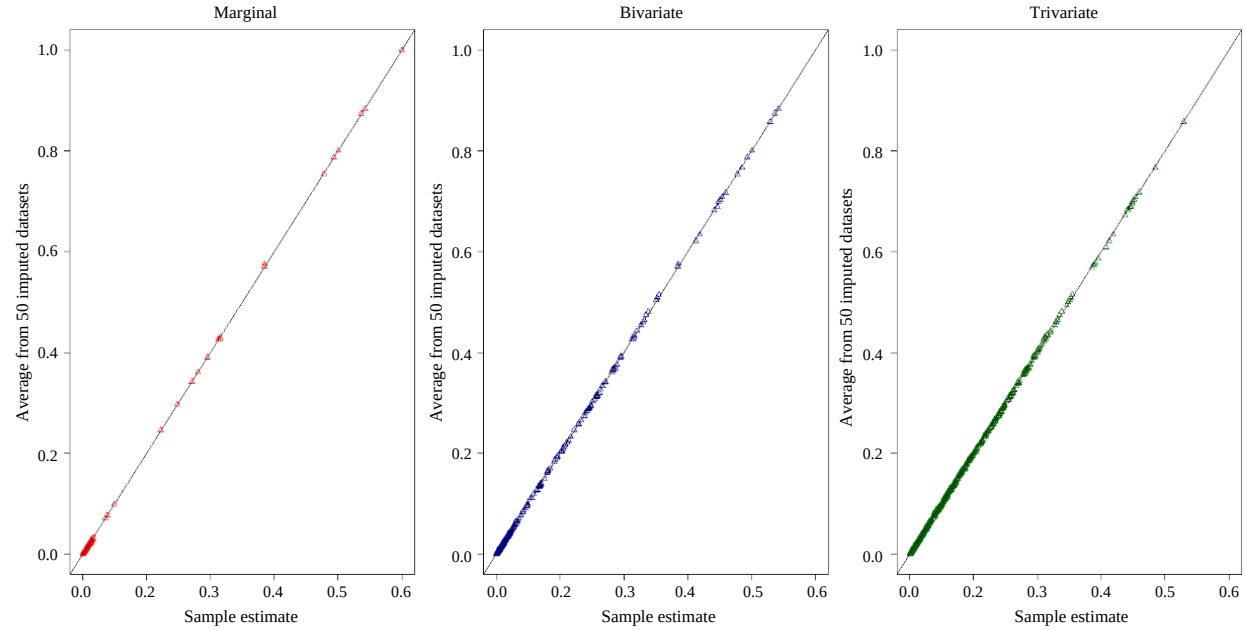


Figure A.2 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.

Table A.4
Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets under MCAR. "No missing" is based on the sampled data before introducing missing values, "NDPMPM" uses the truncated NDPMPM, moving household heads' data values to the household level, and "NDPMPM Capped" uses the truncated NDPMPM with the cap-and-weight approach and moving household heads' data values to the household level. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. *Q* is the value in the full population of 764,580 households

		Q	No Missing	NDPMPM	NDPMPM Capped
All same race household:	$n_i = 2$	0.942	(0.932, 0.949)	(0.924, 0.944)	(0.925, 0.946)
	$n_i = 3$	0.908	(0.907, 0.937)	(0.887, 0.924)	(0.890, 0.925)
	$n_i = 4$	0.901	(0.879, 0.917)	(0.854, 0.900)	(0.855, 0.900)
SP present		0.696	(0.682, 0.707)	(0.683, 0.709)	(0.683, 0.709)
Same race CP		0.656	(0.641, 0.668)	(0.637, 0.664)	(0.638, 0.665)
SP present, HH is White		0.600	(0.589, 0.616)	(0.590, 0.618)	(0.590, 0.618)
White CP		0.580	(0.569, 0.596)	(0.568, 0.596)	(0.568, 0.597)
CP with age difference less than five		0.488	(0.465, 0.492)	(0.422, 0.451)	(0.422, 0.450)
Male HH, home owner		0.476	(0.456, 0.484)	(0.455, 0.483)	(0.456, 0.485)
HH over 35, no CH present		0.462	(0.441, 0.468)	(0.438, 0.466)	(0.438, 0.466)
At least one biological CH present		0.437	(0.431, 0.458)	(0.432, 0.460)	(0.432, 0.460)
HH older than SP, White HH		0.322	(0.309, 0.335)	(0.308, 0.335)	(0.306, 0.333)
Adult female w/ at least one CH under 5		0.078	(0.070, 0.085)	(0.068, 0.084)	(0.067, 0.083)
White HH with Hisp origin		0.066	(0.064, 0.078)	(0.064, 0.079)	(0.064, 0.079)
Non-White CP, home owner		0.058	(0.050, 0.063)	(0.048, 0.061)	(0.048, 0.061)
Two generations present, Black HH		0.057	(0.053, 0.066)	(0.053, 0.066)	(0.053, 0.067)
Black HH, home owner		0.052	(0.046, 0.058)	(0.046, 0.059)	(0.046, 0.059)
SP present, HH is Black		0.039	(0.032, 0.042)	(0.032, 0.043)	(0.032, 0.042)
White-nonwhite CP		0.034	(0.029, 0.039)	(0.032, 0.044)	(0.032, 0.044)
Hisp HH over 50, home owner		0.029	(0.025, 0.034)	(0.025, 0.035)	(0.025, 0.035)
One grandchild present		0.028	(0.023, 0.033)	(0.024, 0.034)	(0.024, 0.034)
Adult Black female w/ at least one CH under 18		0.027	(0.028, 0.038)	(0.027, 0.037)	(0.027, 0.037)
At least two generations present, Hisp CP		0.027	(0.022, 0.031)	(0.022, 0.031)	(0.022, 0.031)
Hisp CP with at least one biological CH		0.025	(0.020, 0.028)	(0.019, 0.028)	(0.019, 0.028)
At least three generations present		0.023	(0.020, 0.028)	(0.019, 0.028)	(0.019, 0.028)
Only one parent		0.020	(0.016, 0.024)	(0.016, 0.024)	(0.016, 0.024)
At least one stepchild		0.019	(0.018, 0.026)	(0.018, 0.027)	(0.018, 0.027)
Adult Hisp male w/ at least one CH under 10		0.018	(0.017, 0.025)	(0.016, 0.025)	(0.016, 0.025)
At least one adopted CH, White CP		0.008	(0.005, 0.010)	(0.005, 0.010)	(0.005, 0.010)
Black CP with at least two biological children		0.006	(0.003, 0.007)	(0.003, 0.007)	(0.003, 0.007)
Black HH under 40, home owner		0.005	(0.005, 0.009)	(0.005, 0.010)	(0.005, 0.011)
Three generations present, White CP		0.005	(0.004, 0.008)	(0.004, 0.010)	(0.004, 0.009)
White HH under 25, home owner		0.003	(0.002, 0.005)	(0.004, 0.009)	(0.004, 0.009)

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40-64.

Bennink, M., Croon, M.A., Kroon, B. and Vermunt, J.K. (2016). Micro-macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*.

Chambers, R., and Skinner, C. (2003). Analysis of Survey Data, Wiley Series in Survey Methodology, Wiley.

Dunson, D.B., and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104, 1042-1051.

Hu, J., Reiter, J.P. and Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis*, 13, 183-200.

Ishwaran, H., and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 161-173.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1, 1-16. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1986001/article/14404-eng.pdf.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Manrique-Vallier, D., and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23, 1061-1079.

Murray, J.S., and Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence (forthcoming). *Journal of the American Statistical Association*.

Raghunathan, T.E., and Rubin, D.B. (2001). Multiple imputation for statistical disclosure limitation. Technical Report.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6785-eng.pdf.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

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

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, Inc.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Savitsky, T.D., and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10.1, 1677-1708.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Si, Y., and Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38.5, 199-521.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 213-239.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 33-51.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 1, 45-54.

Wang, Q., Akande, O., Hu, J., Reiter, J. and Barrientos, A. (2016). NestedCategBayesImpute: Modeling and Generating Synthetic Versions of Nested Categorical Data in the Presence of Impossible Combinations. *The Comprehensive R Archive Network*.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 295-315
Statistics Canada, Catalogue No. 12-001-X

An optimisation algorithm applied to the one-dimensional stratification problem

José André de Moura Brito, Tomás Moura da Veiga and Pedro Luis do Nascimento Silva[1]

Abstract

This paper presents a new algorithm to solve the one-dimensional optimal stratification problem, which reduces to just determining stratum boundaries. When the number of strata *H* and the total sample size *n* are fixed, the stratum boundaries are obtained by minimizing the variance of the estimator of a total for the stratification variable. This algorithm uses the Biased Random Key Genetic Algorithm (BRKGA) metaheuristic to search for the optimal solution. This metaheuristic has been shown to produce good quality solutions for many optimization problems in modest computing times. The algorithm is implemented in the R package *stratbr* available from CRAN (de Moura Brito, do Nascimento Silva and da Veiga, 2017a). Numerical results are provided for a set of 27 populations, enabling comparison of the new algorithm with some competing approaches available in the literature. The algorithm outperforms simpler approximation-based approaches as well as a couple of other optimization-based approaches. It also matches the performance of the best available optimization-based approach due to Kozak (2004). Its main advantage over Kozak's approach is the coupling of the optimal stratification with the optimal allocation proposed by de Moura Brito, do Nascimento Silva, Silva Semaan and Maculan (2015), thus ensuring that if the stratification bounds obtained achieve the global optimal, then the overall solution will be the global optimum for the stratification bounds and sample allocation.

Key Words: Optimal stratification; Genetic algorithm; Integer programming; Nonlinear optimization; BRKGA Metaheuristic.

1 Introduction

Stratified sampling is a widely used approach to achieve efficiency in sampling designs. The substantial literature on optimal stratification (to be reviewed later in this paper) signals to both the importance of this topic for research and to its wide range of applications. Recently, Hidiroglou and Kozak (2017) compared optimization-based and approximate methods for one-dimensional stratification of skewed populations and concluded that optimization methods are superior and should be used in practice.

In this paper, we propose applying a new optimisation algorithm to determine the stratum boundaries, which we coupled with an approach to obtain the globally optimal sample size allocation to the defined strata. The one-dimensional stratification problem is addressed using a global optimisation technique (a metaheuristic) called Biased Random Key Genetic Algorithm (BRKGA), proposed by Gonçalves and Resende (2011). This technique does not ensure achieving the global optimum for the stratum boundaries, but has been shown to produce good quality solutions for many optimization problems in modest computing times (see Gonçalves and Resende, 2004; Gonçalves, Mendes and Resende, 2005; Festa, 2013 and Oliveira, Chaves and Lorena, 2017).

Our approach for sample allocation given a defined stratification (see de Moura Brito et al., 2015), namely a stratification by a specified variable and given number of strata, is based on an integer programming formulation, and always achieves the global optimum for either minimizing the total sample

1. José André de Moura Brito, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106, Centro, Rio de Janeiro RJ, 20231-050. E-mail: jambrito@gmail.com; Tomás Moura da Veiga, SHIN CA 08 Lote 01 Torre 01, 101 Lago Norte, Brasília DF, 71503-508. E-mail: tmvtomas@gmail.com; Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106 Centro, Rio de Janeiro, RJ, 20231-050. E-mail: pedronsilva@gmail.com.

size given precision constraints or minimizing variance for a fixed total sample size, while providing an exact integer sample allocation, and allowing the specification of minimum and maximum sample sizes per strata, as is often required in practical applications. The approach is implemented in the *stratbr* R package (see de Moura Brito et al., 2017a), thus providing a practical alternative to existing approximate methods, which it clearly outperforms in terms of efficiency. It also compares favourably with other optimization methods, which are not guaranteed to provide the optimal allocation given the stratification.

We compared this new approach with methods proposed by Dalenius and Hodges (1959), Gunning and Horgan (2004), Kozak (2004, 2006), Keskintürk and Er (2007), and de Moura Brito, Silva Semaan, Fadel and Brito (2017b), using a set comprising 27 real and artificial survey populations. Our empirical study is much larger than that of Hidiroglou and Kozak (2017), who have used only two populations in their comparison. It is also larger than other studies in earlier literature.

We have not considered, as suggested, comparing our approach with classification or regression trees or other machine learning algorithms that synthesise one or more covariates into groupings that can be used for strata. The main reason for this is that such methods do not consider the variance of the target sample estimator or the sample size given precision constraints as the criteria to optimize. Therefore, they cannot be expected to achieve the optimum for the problem we wish to address. In addition, for classification or regression trees the analyst must also specify a "response variable", in addition to the predictors or auxiliary variables. In many typical sampling situations, the analyst will not have access to data on such a "response variable", and must aim to minimize variance of the estimator for the total of the size or stratification variable instead (as is the case in most of the literature on this topic).

Although we have addressed only the "one-dimensional stratification problem", in that a single size measure is used for stratification, one could always use some predictive model or alternative variable reduction technique to summarise auxiliary variables or covariates into a single "x" or size variable to be used in our proposed approach. Nevertheless, our approach can easily be extended to address multivariate stratification coupled with optimum allocation given the nature of the components of the approach.

The paper is divided as follows: Section 2 contains the key concepts of stratified sampling. Section 3 contains a detailed description of the stratification problem. Section 4 presents the Biased Random Key Genetic Algorithm (BRKGA) and its novel implementation to resolve the stratification problem, in combination with the optimal allocation method proposed by de Moura Brito et al. (2015). Section 5 contains the results of the application of the proposed method compared to those of five other methods available in the literature previously mentioned. Section 6 presents the conclusions of the comparative study.

2 Stratified sampling

In stratified sampling, the first step is to partition the elements of the target population into well defined, preferably homogeneous, mutually exclusive and exhaustive subgroups called strata. Each population

element (unit) is the focus of the survey and provider of the information which it aims to obtain. Survey units can be households, people, farms, business establishments or companies, etc.

Stratified sampling is recommended in practice for several reasons:

- It can improve precision of the overall population estimates for a fixed total cost;
- It enables controlling sample sizes and precision of estimates for the strata, if required;
- It may facilitate balancing workload distribution;
- It may help reduce travelling costs between survey elements, if stratification includes geography.

When the strata are formed such that the intra-stratum variability is small for a key set of variables, stratification is considered successful, since this enables achieving better precision for the estimates relative to other stratification schemes.

Figure 2.1 presents the basic notation to be used in this paper. In stratified sampling, the population U is partitioned into $H > 1$ nonempty subpopulations called strata (Lohr, 2010), of sizes N_1, $N_2, \ldots, N_h, \ldots, N_H$. These subpopulations are non-overlapping and such that, when taken jointly, combine to form the full population, such that:

$$N_1 + N_2 + \ldots + N_H = N. \tag{2.1}$$

$U = \{1, 2, \ldots, N\}$ – Set of elements comprising the target population;

N – Number of population elements, or population size;

n – Number of elements in the sample, or sample size;

$H > 1$ – Total number of strata;

h – Index for the strata;

U_h – Set of elements in stratum h, satisfying $U_h \subset U$ and $U_h \neq \varnothing$;

N_h – Number of population elements, or population size, in stratum h;

s_h – Set of elements sampled in the h^{th} stratum $- s_h \subset U_h$;

n_h – Number of sample elements, or sample size, in stratum h;

y_i – Value of survey variable for population element $i\,(i \in U)$;

x_i – Value of stratification variable for population element $i\,(i \in U)$;

$Y_h = \sum_{i \in U_h} y_i$ – Total of survey variable y in stratum h;

$X_h = \sum_{i \in U_h} x_i$ – Total of stratification variable x in stratum h;

$\bar{Y}_h = Y_h / N_h$ – Population mean of survey variable y in stratum h; and

$\bar{X}_h = X_h / N_h$ – Population mean of stratification variable x in stratum h.

Figure 2.1 Notation to be used in the paper.

Cochran (1977) lists the following factors which affect the efficiency of a stratified sampling design: choice of stratification variable(s); number of strata (H); delimitation of strata; total sample size (n); allocation of the total sample to the strata; and selection method for sampling within strata. The strata are defined using one or more variables for which the values are known for each population element. In the sequence, samples are selected independently within each of the H strata. Samples sizes in the strata are such that $n_1 + n_2 + \ldots + n_H = n$.

Stratified Simple Random Sampling (SSRS) corresponds to the case when sampling within each stratum is carried out using simple random sampling without replacement. Under SSRS, the Horvitz-Thompson (HT) estimator of the overall population total $Y = \sum_{h=1}^{H} Y_h$ is given by Cochran (1977) as:

$$\hat{Y}_{\text{SSRS}} = \sum_{h=1}^{H} N_h \bar{y}_h \tag{2.2}$$

where $\bar{y}_h = \sum_{i \in s_h} y_i / n_h$ is the sample average for elements sampled in stratum h.

The sampling Variance (Var) and Coefficient of Variation (CV) of the estimator $\hat{Y}_{\text{SSRS}}$ are given respectively by:

$$\text{Var}\left(\hat{Y}_{\text{SSRS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hy}^2 / n_h \tag{2.3}$$

$$\text{CV}\left(\hat{Y}_{\text{SSRS}}\right) = \sqrt{\text{Var}\left(\hat{Y}_{\text{SSRS}}\right)} \Big/ Y \tag{2.4}$$

where $S_{hy}^2 = \sum_{i \in U_h} \left(y_i - \bar{Y}_h\right)^2 / \left(N_h - 1\right)$ is the population variance of the survey variable y in stratum h.

Analogous quantities are defined for the HT estimator for the total $X = \sum_{h=1}^{H} X_h$ of the stratification variable x, namely:

$$\hat{X}_{\text{SSRS}} = \sum_{h=1}^{H} N_h \bar{x}_h \tag{2.5}$$

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hx}^2 / n_h \tag{2.6}$$

$$\text{CV}\left(\hat{X}_{\text{SSRS}}\right) = \sqrt{\text{Var}\left(\hat{X}_{\text{SSRS}}\right)} \Big/ X \tag{2.7}$$

where $\bar{x}_h = \sum_{i \in s_h} x_i / n_h$ is the sample average of x for elements sampled in stratum h, and $S_{hx}^2 = \sum_{i \in U_h} \left(x_i - \bar{X}_h\right)^2 / \left(N_h - 1\right)$ is the population variance of the variable x in stratum h.

3 The one-dimensional stratification problem

Consider the population vector $X_U = \{x_1, x_2, \ldots, x_N\}$ corresponding to the stratification variable x. Without loss of generality, we assume that the population elements in U are ordered by the stratification variable such that $x_1 \leq x_2 \leq \ldots \leq x_N$. The stratum boundaries are used to define the H strata according to the rule:

1) $U_1 = \{i \in U \mid x_i \leq b_1\}$;

2) $U_h = \{i \in U \,|\, b_{h-1} < x_i \leq b_h\}$ for $h = 2, 3, \ldots, H-1$;

3) $U_H = \{i \in U \,|\, b_{H-1} < x_i\}$.

The stratification problem corresponds to determining the cut-off points, i.e., the stratum boundaries $b_1 < b_2 < \ldots < b_h < \ldots < b_{H-1}$ such that the variance (or equivalently the CV) of the estimator of total $\hat{Y}_{\text{SSRS}}$ is minimised. In this section, we consider that the total number of strata H is defined before applying the optimal stratification methods considered.

In practice, the values of the survey variable y are not available and hence the variance in expression (2.3) is not computable. A common approach is to minimise instead the variance (or CV) of the estimator $\hat{X}_{\text{SSRS}}$ for the total of the stratification variable x. Several authors have developed methods that focus on this optimization problem, which from now on we call the one-dimensional "stratification problem". We adopted the same approach here.

Finding the boundary points that minimize the variance (2.6) or the CV (2.7) corresponds to a hard problem both from analytic and computational points of view. This is so because the integer population and sample sizes (N_h and n_h, respectively) depend in a nonlinear way on the stratum boundaries. According to de Moura Brito, Ochi, Montenegro and Maculan (2010a), depending on N, H, and the number of distinct population x values, the number of possible choices for the boundary points can be very large.

In view of this difficulty, over the past decades, various methods were developed to search for the optimum stratum boundaries, aiming to provide at least solutions which correspond to local minima of good quality.

Dalenius (1951) tackled the problem for the case $H = 2$ by approximating the variance in (2.6) by ignoring the finite population correction, which is equivalent to assuming that the sampling within strata would have been simple random sampling with replacement. The approximate variance to be minimised is then given by:

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) \cong \sum_{h=1}^{H} N_h^2 S_{hx}^2 / n_h. \tag{3.1}$$

Under the Neyman allocation (Cochran, 1977) using the x variable, and replacing the sample sizes n_h in (3.1) by their theoretical values $n_h = N_h S_{hx} \Big/ \sum_{k=1}^{H} N_k S_{kx}$ leads to the expression used by Dalenius (1951):

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) \cong \left(\sum_{h=1}^{H} N_h S_{hx}\right)^2 \Big/ n. \tag{3.2}$$

Dalenius and Hodges (1959) considered the case when $H > 2$, and offered an analytic solution which relied on approximating the distribution of the x variable by its histogram with a moderate number of classes. Still considering the approximate variance and assuming Neyman's allocation, Ekman (1959) provided a solution using a geometric approach to find the stratum boundaries. Hedlin (2000) further extended Ekman's solution while retaining the original variance (2.6) as the function to be minimised, which he labelled the extended Ekman rule.

Hidiroglou (1986) proposed an approach which pre-specifies the required precision (CV) for the estimator of total, and which divides the population into two strata $(H = 2)$ such that the total sample size n is minimised. In this paper, the second stratum corresponds to a "take-all" or "certainty" stratum, where all elements are included in the sample with probability one $(n_2 = N_2)$. Lavallée and Hidiroglou (1988) generalized the approach to $H > 2$ strata, while retaining the idea that the stratum containing the largest population units is to be sampled completely. Their approach relied on adopting a special type of allocation called Power Allocation (Bankier, 1988). More recently, Rivest (2002) further generalised the approach of Lavallée and Hidiroglou (1988) while considering that the target is minimising the variance for the estimator of a total of a model-based prediction of the survey variable y, instead of the stratification variable x.

Gunning and Horgan (2004) proposed the so-called Geometric method for defining the stratum boundaries. This method assumes that the CVs of the stratification variable x are approximately constant, and that the distribution of the stratification variable is approximately uniform within each stratum. Under these assumptions, the optimum stratum boundaries would form a Geometric progression, thus leading to a very simple analytic solution.

Keskintürk and Er (2007) proposed an approach based on a global optimisation technique called Genetic Algorithms. Following a similar idea, de Moura Brito et al. (2017b) applied another global optimisation technique called GRASP to the stratification problem. Here we followed a route like Keskintürk and Er (2007), but have adopted an efficient choice of Genetic Algorithm, namely the Biased Random Key Genetic Algorithm (BRKGA), described in Section 4.

Kozak (2004) proposed a method called random search, followed later by Kozak and Verma (2006), where this approach was compared to the Geometric method of Gunning and Horgan (2004). Khan, Nand and Ahmad (2008) used ideas of dynamic programming to develop an algorithm that determines the stratum boundaries considering that the stratification variable has either a Triangular or Normal distribution, and that sampling within strata is with replacement. De Moura Brito, Maculan, Lila and Montenegro (2010b) proposed an exact algorithm based on graph theory, where proportional allocation to the strata is assumed.

Er (2011) compared the efficiency of several methods available in the literature, taking the Geometric approach of Gunning and Horgan (2004) as the initial solution. Kozak (2014) compared his random search with the Genetic Algorithm proposed by Keskintürk and Er (2007). Rao, Khan and Reddy (2014) developed a method that tackles the stratum boundary determination and stratum allocation problems simultaneously. Their algorithm relies on the assumption that the stratification variable follows a Pareto distribution. Our approach is more general and does not assume that the size variable follows a particular distribution.

4 Biased Random-Key Genetic Algorithm

The Biased Random-Key Genetic Algorithm (BRKGA, from now on), proposed by Gonçalves and Resende (2011), is a metaheuristic approach which has been applied to address several optimization

problems – see for example Festa (2013) and Oliveira et al. (2017). The principle behind this approach mimics the biologic theory of evolution of species.

The algorithm starts with an initial "population" of feasible solutions to the target problem generated by a specified random mechanism. This population then evolves over successive iterations by preserving the best solutions available at each iteration (elite solutions), and by replacing the other (non-elite) solutions with solutions generated through random perturbation operations that mimic crossing and mutation in natural populations. Over the iterations, solutions are selected to be preserved or to evolve based on the value of the function to be optimized.

In BRKGA, the candidate solutions are encoded, i.e., are represented by vectors where the components are numbers in the $(0; 1)$ interval. Given an observed vector, a decoding procedure must be applied. The decoding procedure maps the value of a vector with a corresponding feasible solution of the target optimization problem. The decoding procedure is what connects BRKGA with the specific optimization problem to be addressed. Figure 4.1 displays the pseudo-code for a generic BRKGA algorithm.

The approach is described and illustrated in detail in Section 4.1 using an example that considers the one-dimensional stratification problem and describes all the steps mentioned in Figure 4.1.

1) Generate the initial population composed of p random vectors (keys) $\mathbf{v}$, where each value is a random draw from the Uniform $[0; 1]$ distribution.
2) Apply the decoding procedure to each vector $\mathbf{v}$ in the population, yielding p feasible solutions to the optimization problem.
3) Compute the value of the objective function for each solution in the population.
4) Select the best $p_e\,(1 < p_e < p)$ solutions (designated elite) based on the values of the objective function and add them to the population that will be considered in the next iteration.
5) Generate $p_m\,(1 < p_m < p)$ new random vectors as in step 1), called *mutants*, and add them to the population that will be considered in the next iteration.
6) Generate the remaining $(p - p_e - p_m)$ vectors, designated *crossed*, to complete the population that will be considered in the next iteration by crossing one of the p_e vectors corresponding to an elite solution with one of the $(p - p_e)$ vectors corresponding to one of the non-elite solutions in the current iteration.
7) Iterate from step 2) while the stopping criteria are not satisfied.

Figure 4.1 Pseudo-code for a BRKGA.

4.1 BRKGA for the one-dimensional stratification problem

First consider the population vector $X_U = \{x_1, x_2, \ldots, x_N\}$, and derive the set $C = \{c_1, c_2, \ldots, c_K\}$ containing the K distinct values of x observed in the population. For example, if $X_U = \{1, 3, 3, 5, 6, 7, 7, 7, 8, 9, 10, 10, 11\}$, then $C = \{1, 3, 5, 6, 7, 8, 9, 10, 11\}$. When $K > 100$, compute the ten largest

percentiles of x to obtain the set $Q = \{q_{90}, q_{91}, \ldots, q_{99}, q_{100}\}$. When $K \leq 100$, compute the selected percentiles of x to obtain the set $Q = \{q_5, q_{10}, \ldots, q_{95}, q_{100}\}$. The cut-off point of 100 for K was chosen after some initial experimentation of the approach with some of the populations considered in the numerical experiments to be described in Section 5. The alternative definitions for the set Q help with achieving larger diversity in the set of feasible solutions to be generated by the BRKGA.

To apply BRKGA to the one-dimensional stratification problem, each solution is represented by a vector $\mathbf{v} = \{v_1, \ldots, v_H\}$ with H positions, where the first $H-1$ positions contain values between 0 and 1, and position H contains the value of a percentile of the distribution of the stratification variable x.

Then take $x_{\min}$ as the smallest value in C, and v_H as an element selected at random from Q. For the first iteration, sample the values in the first $H-1$ positions of each vector $\mathbf{v}$ independently from the Uniform $[0; 1]$ distribution.

The decoding procedure to obtain a solution to the one-dimensional stratification problem from each vector $\mathbf{v}$ generated is defined as:

$$b_h = x_{\min} + v_h (v_H - x_{\min}) \quad \text{for} \quad h = 1, \ldots, H-1. \tag{4.1}$$

After obtaining the $H-1$ values for b_h, these must be sorted in increasing order, such that the elements of the resulting vector $\mathbf{b} = \left(b_{(1)}, b_{(2)}, \ldots, b_{(H-1)}\right)$ form the solution boundary points for the corresponding vector $\mathbf{v}$, where $b_{(h)}$ is the h^{th} order statistic of the values $b_1, \ldots, b_{H-1}$ calculated using (4.1).

To illustrate an example of decoding, suppose that $H = 4$, $x_{\min} = 10$, $K = 300$, $Q = \{200, 215, 280.5, 300, 318, 400, 425, 478, 500, 510\}$ Consider also the vector $\mathbf{v} = (0.48, 0.35, 0.20)$ generated as described above. Then it follows that: $b_1 = 10 + 0.48 \times (200 - 10)$; $b_2 = 10 + 0.35 \times (200 - 10)$; and $b_3 = 10 + 0.20 \times (200 - 10)$. Then it follows, after sorting, that $\mathbf{b} = (48, 76.5, 101.2)$.

Given the vector $\mathbf{b}$, the values of N_h and S_{hx}^2 are easily obtained for each of the H strata. The values of the sample sizes n_h for each of the strata are obtained by applying the approach for optimal allocation proposed by de Moura Brito et al. (2015). This approach computes the sample sizes n_h such that a weighted sum of variances (or CVs) of the estimators of totals of m survey variables is minimized, while the total sample size n is kept fixed.

Since here we consider the variance of the estimator for the total of the stratification variable x as the target for minimization, we set $m = 1$ and use formulation (D) as provided in de Moura Brito et al. (2015) to resolve the one-dimensional optimal allocation problem taking equation (2.6) as the variance to be minimized. Note that the approach used provides the global optimum for the allocation problem.

The algorithm then proceeds as indicated in Figure 4.1 by generating an initial set of p vectors $\mathbf{v}$. In step 2, each of these vectors $\mathbf{v}$ is decoded to obtain a feasible solution $\mathbf{b}$ to the optimum stratification problem. In step 3, the optimum allocation corresponding to $\mathbf{b}$ is obtained and the value of the objective

function is calculated. Steps 4 to 6 are then applied to obtain the next population of feasible solutions, and the process is repeated until the stopping criteria are satisfied. Step 4 identifies the p_e elite solutions and add these to the next population. In Step 5, p_m mutant solutions are generated and added to the next population. In Step 6, $(p - p_e - p_m)$ crossed solutions are generated using the "uniform crossover" operator proposed by Spears and DeJong (1991) to produce a new vector $\mathbf{v}$ from one of the p_e elite solutions and one of the current $(p - p_e - p_m)$ non-elite solutions. The process is as follows: once the two vectors (say $\mathbf{v}_e$ and $\mathbf{v}_n$) to cross are selected, an auxiliary random-key vector $(\mathbf{v}_a)$ is generated with independent draws from the Uniform $[0; 1]$ distribution. Let $r_c > 0.5$ be a pre-specified probability that a value is copied from the elite vector $\mathbf{v}_e$. Then the crossed vector $\mathbf{v}_c$ is formed by taking the values from $\mathbf{v}_e$ in the positions where the corresponding value in $\mathbf{v}_a$ is less than r_c (equal to 0.7 in the example of Figure 4.2) and from $\mathbf{v}_n$ in all other positions.

To produce each one of the $(p - p_e - p_m)$ vectors for the next generation, the algorithm selects a vector v_e at random (using the *sample* function from R) from the p_e elite vectors and another vector v_n from the $p - p_e$ non-elite vectors, and crosses these vectors. The selection of vectors from both subsets is done with replacement, implying that individual elite or non-elite vectors may be selected for crossing more than once.

Vectors\positions	1	2	3
$\mathbf{v}_e$	0.31	0.77	0.65
$\mathbf{v}_n$	0.26	0.18	0.36
$\mathbf{v}_a$	0.58	0.89	0.11
$\mathbf{v}_c$	0.31	0.18	0.65

Figure 4.2 Uniform crossing with $r_c = 0.7$.

Now consider the example where $H = 4$, $x_{\min} = 10$, $K = 300$, $p = 8$, $p_e = 3$, $p_m = 3$, $r_c = 0.7$ and $Q = \{200, 215, 280.5, 300, 318, 400, 425, 478, 500, 510\}$. Figure 4.3 illustrates the application of all the steps in BRKGA to the one-dimensional stratification problem, for two consecutive iterations of the algorithm.

The BRKGA approach described here for the one-dimensional optimal stratification problem was implemented in the R package *stratbr* (see de Moura Brito et al., 2017a), which is available from CRAN. This package was used to obtain all the results presented in Section 5.

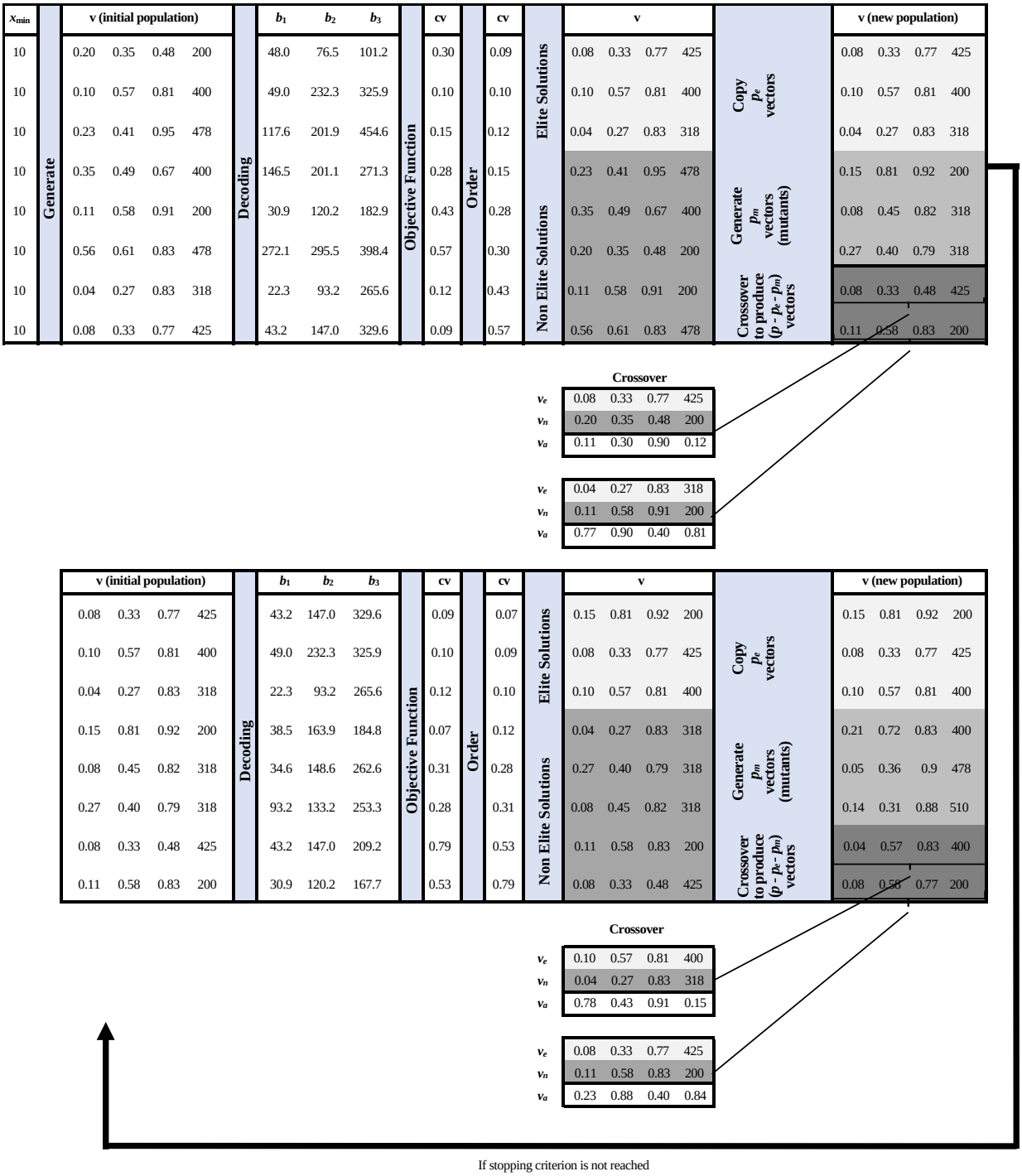


Figure 4.3 Illustration of BRKGA approach for optimal stratification.

5 Computational results

In this section, we present the results of application of six methods to solve the stratification problem, namely: Dalenius and Hodges (DH), Geometric (GH), Kozak (KO), Genetic Algorithm of Keskinturk and Er (KE), GRASP (GR) and the new BRKGA method described in Section 4 (BR). All experiments where carried out using R version 3.3.1. The methods DH, GH and KO are available from the R package *stratification* of Baillargeon and Rivest (2014) (version 2.2-5). With these methods, the Neyman sample allocation method was used. The KE method is available from the R package *GA4Stratification* of Er, Keskintürk and Daly (2010) (version 1.0). With this method, the maximum number of iterations considered was 10,000 and the values of the other parameters required were the same as those reported by Keskintürk and Er (2007), namely using $p = 35$ candidate solutions in each population, a mutation rate of 15% and the sample allocation based also on the Genetic Algorithm. Both the GR and the BR methods were implemented in R by the authors, and the code is provided in package *stratbr* of de Moura Brito et al. (2017a) (version 1.2) available from CRAN.

For the BR method, $p = 100$ candidate solutions were considered in each iteration, with 20% of the solutions being made elite $(p_e = 20)$ and 30% of the solutions being mutant $(p_m = 30)$ in each iteration. The probability of copying a gene from the elite vector was set at $r_c = 0.6$. The total number of iterations was set at 1,500. For the sample allocation, both the BR and the GR methods were coupled with the formulation proposed by de Moura Brito et al. (2015) which is available from the R package *MultAlloc*, also available from CRAN.

To compare the relative efficiency of these methods, they were applied to 27 different populations. Some of these populations are available from the R packages *stratification* and *GA4Stratification*, and were previously used in other comparison studies such as Keskintürk and Er (2007), Er (2011), and de Moura Brito et al. (2017b). Appendix A contains brief descriptions of all these populations, including information on which variable was considered as the "x variable" in each population. Table 5.1 provides some summaries to describe these populations.

The 27 populations considered here form a very diverse set, with total sizes varying from a few hundred (ME84 and P75 with $N = 284$ are the smallest) to several thousand (Coffee with $N = 18{,}570$ is the largest). In the size measure that matters most for efficiency of our optimization algorithm, namely the number K of distinct values of the stratification variable, there's also large variation (from $K = 51$ for Kozak1 to $K = 5{,}453$ to Kozak3). They also display wide variation in the asymmetry of the x variable's distributions, ranging from modestly negative (-0,70 for Beta103 to a substantial 40.04 for CensoCO).

All the calculations for the computational experiment were performed using R in a computer with 24 GB RAM, with 8 processors of 3.40 GHz (I7). Taking advantage of the multicore architecture in modern computers, the *snowfall* R package was used to parallelize the BRKGA algorithm. More specifically, at each iteration, the decoding procedure produces a set of solutions for the boundary points. These boundary points are then supplied to the *MultAlloc* package for optimum allocation, to obtain the sample sizes in each stratum, and then to compute the objective variance function. Since the computation time for this step is

impacted directly by the use of this global optimization formulation, the allocation and calculation of the objective function were parallelized.

Table 5.1
Summaries of the stratification variable for the 27 populations

Populations	N	K	Minimum	Maximum	Skewness
AgrMinas	844	226	5.00	47,800.00	7.32
BeefFarms	430	353	50.00	24,250.00	4.56
Beta103	1,000	1,000	357.98	985.96	-0.70
CensoCO	9,977	79	1.00	911.00	40.04
Chi5	1,000	1,000	0.06	23.43	1.40
Coffee	18,570	538	0.01	13,212.00	19.69
Debtors	3,369	1,129	40.00	28,000.00	6.44
HHinctot	16,025	224	1.00	6,900.00	2.71
Iso2004	487	487	6.36	1,044.66	10.03
Kozak1	4,000	51	72.00	3.00	1.40
Kozak3	2,000	581	2,793.00	6.00	3.55
Kozak4	10,000	5,453	74,400.00	62.00	4.20
ME84	284	264	173.00	47,074.00	8.64
MRTS	2,000	2,000	1.41	4,863.66	8.61
P100e10	1,000	1,000	73.56	127.32	-0.03
P75	284	68	4.00	671.00	8.43
Pop500	500	261	0.01	47,841.42	21.53
Pop800	800	402	0.01	4,735.10	22.13
pop1076	1,076	88	5.00	1,643.00	13.23
pop1616	1,616	165	5.00	2,618.00	11.09
pop2911	2,911	247	5.00	2,497.00	11.50
REV84	284	277	347.00	59,877.00	7.83
SugarCaneFarms	338	101	18.00	280.00	2.26
Swiss	2,896	881	0.00	3,634.00	2.73
USbanks	357	200	70.00	977.00	2.07
UScities	1,038	116	10.00	198.00	2.87
UScolleges	677	576	200.00	9,623.00	2.45

Note: N is the total population size, and K is the number of unique values for the stratification variable.

All six methods considered in the numerical experiment were applied to each of the 27 populations, for numbers of strata H equal to 3, 4, 5 and 6. These values were used since they are often considered in applications, as well as in similar comparative studies available in the literature, such as Er (2011) and Gunning and Horgan (2004). We did not consider larger values for H since the additional gains in efficiency for $H > 6$ are modest. The sample size $n = 100$ (i.e., fixed cost) was used, as in the numerical experiments of Er (2011) and Kozak and Verma (2006).

To assess the efficiency of the methods, the CVs of the estimator for the total of the stratification variable x were calculated for each population and number of strata, leading to $27 \times 4 = 108$ scenarios for each

method. CVs were obtained from equation (2.7) and multiplied by 100, to be presented as percentages. Table 5.2 provides the CVs attained by the six methods. The shaded cells indicate methods providing the best solution (minimum CV) for each of 108 scenarios. The NAs in these tables represent cases where solutions could not be obtained due to problems of the specific stratification method or with the corresponding allocation.

Analyzing the results provided in Table 5.2, and in particular, the shaded cells, it is evident that BR has excellent performance when compared to the five competitors considered. This perception is reinforced by the plots in Figure 5.1, where BR was compared with all competitors. Points above the straight line represent scenarios where the method chosen for comparison is outperformed by BR. It is evident from these plots that the three best performing methods are GR, KO and BR.

Table 5.3 provides the percentage of times that each method produced the best solution over the 108 scenarios. Both BR and KO display performance which is superior to that of the other methods and have tied in the number of times that they have achieved the best solution. DH produced the best solution for only three of the 108 scenarios, and GH has never produced a best solution.

The Geometric method GH, besides leading to high CVs, also often provided infeasible solutions, where the stratum limits lead to allocations where sample sizes were larger than the corresponding population sizes. This method also sometimes partitioned the population such that there were very few population elements in some strata. According to Gunning and Horgan (2004), and as noted by Keskintürk and Er (2007), since the interval widths increase geometrically, the GH method will not perform well when the stratification variable has small values, since this will lead to some narrow strata. This method is also not applicable when the smallest value in the stratification variable is zero.

For most populations, the KE method has produced CVs close to those obtained by the KO, GR and BR methods, which are the most efficient in terms of computing time. Large variation on computing time was observed between different methods. The KE method showed the worst results in this criterion, having displayed computing times much larger than those of the competing methods. The KO method, on the other hand, was the fastest in terms of computing time, while at the same time often achieving the best possible precision (lowest CV). The BR method showed computing time in between those of the KO and KE methods.

The graph in Figure 5.2 shows the percentages of times that each of the methods BR, KO, KE and GR produced the best solution, separated by number of strata. It shows a clear advantage of the BR method when compared to the KE and GR methods. When compared to KO, BR performed better for $H = 3$ and $H = 6,$ while KO was the winner for $H = 4$ and $H = 5.$ GR performed as well as KO for $H = 3$ and $H = 6,$ but was outperformed by both BR and KO for $H = 4$ and $H = 5.$ KE was the clear loser in this analysis, for any number of strata H.

We have also searched for associations between performance and other potential drivers, such as the skewness or the size $(N$ or $K)$ of the populations, but have not found any meaningful association within our limited set of populations.

Table 5.2
CVs for the estimator of total of the stratification variable by scenario

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
AgrMinas	3	4.158	7.187	4.050	4.089	4.050	4.050
	4	2.714	4.965	2.643	2.811	2.645	2.645
	5	2.325	3.828	1.945	2.262	1.945	1.945
	6	1.821	2.975	1.593	1.932	1.580	1.580
BeefFarms	3	2.758	2.491	1.875	2.086	1.875	1.875
	4	1.853	1.825	1.188	1.557	1.188	1.188
	5	1.455	1.369	0.902	1.280	0.902	0.902
	6	1.148	1.167	0.726	0.990	0.726	0.726
Beta103	3	0.561	0.810	0.560	0.560	0.559	0.559
	4	0.413	0.579	0.410	0.408	0.410	0.410
	5	0.337	0.500	0.329	0.329	0.329	0.329
	6	0.280	0.418	0.276	0.275	0.277	0.276
CensoCO	3	NA	4.839	4.334	4.336	4.334	4.334
	4	NA	4.388	3.078	3.062	3.078	3.078
	5	NA	NA	2.401	2.435	2.401	2.401
	6	NA	NA	1.949	1.956	1.943	1.943
Chi5	3	2.522	4.217	2.502	2.489	2.502	2.502
	4	1.897	3.199	1.889	1.881	1.889	1.889
	5	1.518	2.875	1.515	1.538	1.515	1.515
	6	1.258	NA	1.248	1.251	1.248	1.248
Coffe	3	10.049	12.598	6.906	6.876	6.906	6.906
	4	NA	10.450	4.996	5.027	4.996	4.996
	5	NA	8.124	3.877	3.939	3.877	3.877
	6	NA	6.756	3.176	3.477	3.176	3.176
Debtors	3	5.626	6.150	5.554	5.554	5.554	5.554
	4	4.098	4.387	4.049	4.049	4.049	4.049
	5	3.163	3.595	3.131	3.131	3.131	3.131
	6	2.639	2.897	2.562	2.562	2.562	2.562
HHinctot	3	3.206	5.106	3.184	3.184	3.184	3.184
	4	2.436	4.542	2.429	2.430	2.429	2.429
	5	1.993	4.225	1.973	1.979	1.973	1.973
	6	1.676	3.794	1.629	1.629	1.629	1.629
Iso2004	3	2.716	3.330	1.894	1.894	1.894	1.894
	4	2.059	2.154	1.206	1.206	1.207	1.207
	5	1.616	1.839	0.908	0.908	0.909	0.909
	6	1.380	NA	0.702	0.703	0.704	0.703
Kozak1	3	1.695	2.432	1.695	1.695	1.695	1.695
	4	1.305	2.020	1.301	1.301	1.301	1.301
	5	1.051	1.705	1.050	1.052	1.050	1.050
	6	0.904	1.402	0.890	0.917	0.890	0.890
Kozak3	3	3.673	5.049	3.663	3.659	3.663	3.663
	4	2.733	3.980	2.723	2.724	2.723	2.723
	5	2.208	3.199	2.178	2.231	2.178	2.178
	6	1.823	2.733	1.817	1.827	1.819	1.817
Kozak4	3	4.263	5.811	4.257	4.239	4.257	4.257
	4	3.219	4.696	3.204	3.193	3.205	3.204
	5	2.606	3.873	2.589	2.587	2.591	2.589
	6	2.168	3.236	2.155	2.155	2.157	2.158
ME84	3	1.703	2.527	1.296	1.296	1.296	1.296
	4	1.402	1.642	0.870	0.870	0.870	0.870
	5	1.050	1.549	0.661	0.661	0.661	0.661
	6	0.907	1.213	0.521	0.577	0.521	0.521

Table 5.2(continued)
CVs for the estimator of total of the stratification variable by scenario

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
MRTS	3	4.363	5.829	4.167	4.167	4.167	4.167
	4	3.406	5.259	2.960	2.960	2.961	2.960
	5	2.498	4.015	2.297	2.485	2.297	2.297
	6	2.167	3.445	1.836	1.836	1.838	1.836
P100e10	3	0.375	0.444	0.373	0.371	0.373	0.373
	4	0.295	0.346	0.294	0.294	0.294	0.294
	5	0.236	0.288	0.236	0.236	0.236	0.236
	6	0.198	0.242	0.196	0.198	0.196	0.196
P75	3	1.635	2.592	1.459	1.459	1.459	1.459
	4	1.415	1.798	0.966	0.966	0.966	0.966
	5	1.047	1.563	0.829	0.835	0.713	0.713
	6	0.896	1.250	0.769	0.553	0.552	0.552
pop1076	3	4.597	3.715	2.437	2.775	2.437	2.437
	4	NA	2.853	1.624	2.164	1.624	1.624
	5	NA	2.168	1.204	1.869	1.203	1.203
	6	NA	1.827	0.953	1.549	0.951	0.951
pop1616	3	4.989	4.318	3.898	3.921	3.898	3.898
	4	3.823	3.267	2.564	2.716	2.564	2.564
	5	3.187	2.508	1.882	2.183	1.882	1.882
	6	NA	2.050	1.527	1.962	1.496	1.496
pop2911	3	5.925	5.935	5.605	5.569	5.605	5.605
	4	4.070	3.992	3.807	3.807	3.807	3.807
	5	3.262	3.183	2.918	2.943	2.918	2.918
	6	2.632	2.649	2.281	2.418	2.281	2.281
Pop500	3	NA	0.678	0.092	0.127	0.092	0.092
	4	NA	0.178	0.059	0.082	0.060	0.060
	5	NA	0.194	0.043	0.059	0.045	0.046
	6	NA	0.117	0.033	0.046	0.036	0.037
Pop800	3	NA	3.133	1.555	2.448	1.555	1.555
	4	NA	2.755	0.996	1.511	0.996	0.996
	5	NA	1.620	0.701	1.261	0.702	0.702
	6	NA	1.436	0.546	0.823	0.550	0.548
REV84	3	1.901	2.777	1.614	1.776	1.614	1.614
	4	1.500	1.975	1.120	1.120	1.120	1.120
	5	1.235	1.700	0.835	0.836	0.835	0.835
	6	0.881	1.315	0.666	0.666	0.667	0.666
SugarCaneFarms	3	1.640	1.929	1.627	1.628	1.627	1.627
	4	1.152	1.440	1.118	1.122	1.118	1.118
	5	0.912	1.186	0.839	0.858	0.839	0.839
	6	0.707	1.041	0.691	0.732	0.682	0.682
Swiss	3	3.726	NA	3.682	3.683	3.690	3.682
	4	2.830	NA	2.781	2.781	2.787	2.781
	5	2.246	NA	2.227	2.549	2.232	2.228
	6	1.905	NA	1.860	1.880	1.864	1.860
USbanks	3	1.861	1.843	1.802	1.802	1.802	1.802
	4	1.364	1.417	1.270	1.270	1.270	1.270
	5	1.118	1.079	0.861	0.861	0.861	0.861
	6	0.794	0.850	0.718	0.710	0.710	0.710
UScities	3	2.738	2.705	2.655	2.687	2.655	2.655
	4	1.972	1.951	1.927	1.934	1.927	1.927
	5	1.483	1.451	1.436	1.437	1.436	1.436
	6	1.260	1.305	1.228	1.214	1.209	1.209
UScolleges	3	2.928	3.169	2.749	2.749	2.749	2.749
	4	2.106	2.185	2.018	2.018	2.018	2.018
	5	1.707	1.838	1.606	1.607	1.607	1.606
	6	1.486	1.488	1.323	1.323	1.323	1.323

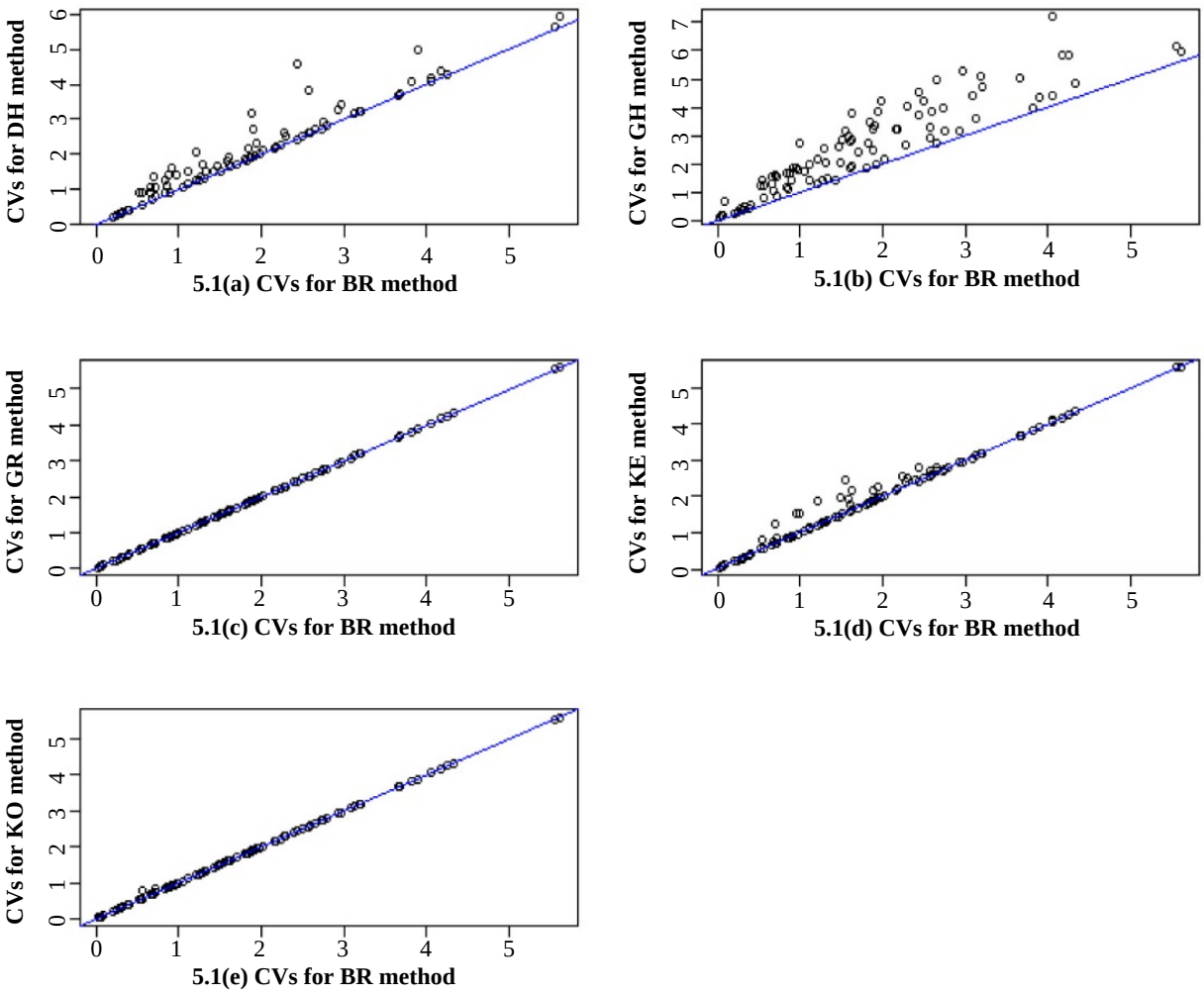


Figure 5.1 Comparing CVs of total estimators under alternative stratification methods, for all populations and numbers of strata (H).

Table 5.3
Percentage of times that method produced the best solution

Method	% Times best
DH	2.8
GH	0.0
KE	42.6
GR	71.3
KO	78.7
BR	78.7

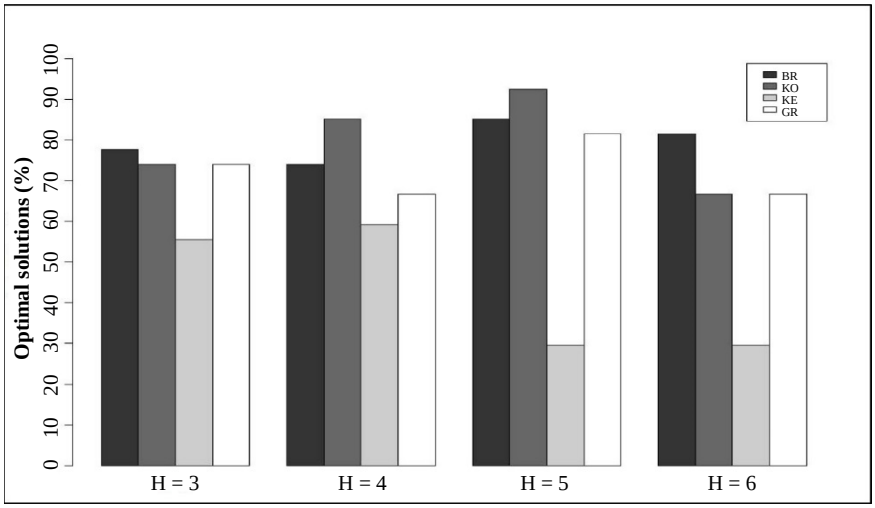


Figure 5.2 Percentage of best solutions yielded by method and number of strata (H).

6 Conclusions

As already mentioned, stratified sampling is a very important idea in survey sampling design, in helping to achieve improved precision of survey estimates for a given sample size or survey budget. This is particularly true for skewed or heterogeneous populations often found in business or establishment surveys. The potential gains due to stratification are strongly dependent on the delimitation of the strata and on the allocation of the sample to the strata, given a specified stratification variable and sample selection method.

The present paper presented a new optimization method for the stratification problem based on the Biased Random-Key Genetic Algorithm (BRKGA). Our approach (named BR) couples BRKGA for the definition of the stratum boundaries with the formulation for optimum sample allocation proposed in de Moura Brito et al. (2015), which is efficient with respect to computing time for large populations (large N).

The results reported here for the comparison of this approach with the five competitors considered suggest that BR offers a good alternative for addressing the stratification and allocation problems in a practical situation.

Our approach can be easily generalised to cases where the stratification variable x is not "measured", but instead represents a summary of several covariates in the form of a predicted y variable. The same is true for generalising to two or more numeric x variables, which can be easily accomplished by changing the decoding function used to retrieve feasible solutions from the BRKGA algorithm with the R package *stratbr* (see de Moura Brito et al., 2017a).

Future work will focus on developing and evaluating alternative decoding procedures which may be used in BR, aiming to produce solutions with superior quality when compared to those produced using the decoding procedure considered here. This research may also focus on solving the dual problem of minimising the total sample size for a specified precision, such as did Lavallée and Hidiroglou (1988).

Finally, additional empirical work may focus on considering varying sample sizes across the various study populations, as did Kozak (2004) and Gunning and Horgan (2004).

Appendix A

Table A1
Description of the 27 populations considered in the numerical experiment

Population	Description
AgrMinas	Agricultural production of municipalities in Minas Gerais State, Brazil, from 2006 Agricultural Census. Stratification variable: planted area.
BeefFarms	Australian beef farms, stratified into seven industrial regions, as considered by (Chambers and Dunstan, 1986). Stratification variable: farm size.
Beta103	Simulated population generated from a Beta distribution with parameters $a = 10$ and $b = 3$ as considered by (Keskintürk and Er, 2007).
Chi5	Simulated population generated from a Chi-square distribution with $df = 5$ as considered by (Keskintürk and Er, 2007).
Coffee	Coffee farms in the state of Paraná, Brazil, in the 1996 Agricultural Census, as considered by (de Moura Brito et al., 2015). Stratification variable: number of coffee trees.
CensoCO	Data from the 2012 school census in Brazil for the mid-west region. Stratification variable: number of classrooms.
Debtors	Population of debtors of an Irish firm as considered by (Er, 2011). Stratification variable: Irish debtors' stated liabilities.
HHinctot	Population of gross family income values (income before tax) from a Family Expenditure Survey 2001 carried out by Statistics Canada, as considered by (Er, 2011).
Iso2004	Data on net sales of 487 Turkish Industrial Enterprises out of the 500 largest enterprises in 2004, obtained by the Istanbul Industrial Chamber, as considered by (Keskintürk and Er, 2007). Stratification variable: net sales.
Kozak1, Kozak3, Kozak4	Populations considered by (Kozak and Verma, 2006). Stratification variable: were generated based on following formula: $X = \exp(Z)$, where Z is a realization of a normal random variable.
ME84	This data is from Särndal, Swensson and Wretman (1992) as considered by (Er, 2011). Stratification variable: number of municipal employees in 1984.
MRTS	Population simulated from the Monthly Survey on Sales in Retail Trade from Statistics Canada, as considered by (Er, 2011). Stratification variable: the size measure used for Canadian retailers in the Monthly Retail Trade Survey (MRTS) carried out by Statistics Canada. This size measure is created using a combination of independent survey data and three administrative variables from the corporation tax return.
P75	Population in thousands of the 284 Swedish municipalities in 1975, as considered by (Er, 2011). Stratification variable: population in thousands.
P100e10	Population simulated from a Normal distribution with $\mu = 100$ and $\sigma = 10$ as considered by (Keskintürk and Er, 2007).
pop1076	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
pop1616	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
pop2911	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
Pop500	Population with $N = 500$ simulated from the Log-Normal Distribution $X = e^z$ where Z is Normal with $\mu = 4$ and $\sigma^2 = 2.7$ as considered by (Hedlin, 2000).
Pop800	Population with $N = 800$ simulated from the Log-Normal Distribution $X = e^z$ where Z is Normal with $\mu = 4$ and $\sigma^2 = 2.7$ as considered by (Hedlin, 2000).
REV84	Value of buildings in million Swedish Crown for the 284 Swedish municipalities in 1984, as considered by (Er, 2011). Stratification variable: the revenues from the 1985 municipal taxation.
SugarCaneFarms	Australian sugar cane farms as considered by (Chambers and Dunstan, 1986). Stratification variable: total cane harvested.
USbanks	Assets in millions of US Dollars for the large north American commercial banks, as considered by (Er, 2011). Stratification variable: the resources in millions of dollars of large commercial US banks.
UScities	Population in thousands for North American cities in 1940, as considered by (Er, 2011). Stratification variable: population in thousands.
UScolleges	Numbers of students in four-year US faculties in 1952-1953 as considered by (Er, 2011). Stratification variable: number of students.
Swiss	Data on Swiss municipalities in 2003, as available from the SamplingStrata package in R. Stratification variable: area under cultivation.

References

Baillargeon, S., and Rivest, L.-P. (2014). Stratification: Univariate stratification of survey populations. R package version 2.2-5. http://CRAN.R-project.org/package=stratification.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Chambers, R., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 3, 597-604.

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

Dalenius, T. (1951). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1-2, 133-148.

Dalenius, T., and Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 285, 54, 88-101.

De Moura Brito, J.A.M., do Nascimento Silva, P.L. and da Veiga, T.M. (2017a). Stratbr: Optimal Stratification in Stratified Sampling. R package version 1.2. https://CRAN.R-project.org/package=stratbr.

De Moura Brito, J.A.M., do Nascimento Silva, P.L., Silva Semaan, G. and Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41, 2, 427-442. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-eng.pdf.

De Moura Brito, J.A.M., Maculan, N., Lila, M. and Montenegro, F. (2010b). An exact algorithm for the stratification problem with proportional allocation. *Optimization Letters*, 4, 185-195.

De Moura Brito, J.A.M., Ochi, L., Montenegro, F. and Maculan, N. (2010a). An iterative local search approach applied to the optimal stratification problem. *International Transactions in Operational Research*, 17, 6, 753-764.

De Moura Brito, J.A.M., Silva Semaan, G., Fadel, A. and Brito, L.R. (2017b). An optimization approach applied to the optimal stratification problem. *Communications in Statistics: Simulation and Computation*, 46, 4419-4451.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 1, 219-229.

Er, S. (2011). Comparison of the efficiency of the various algorithms in stratified sampling when the initial solutions are determined with geometric method. *International Journal of Statistics and Applications*, 1, 1, 1-10.

Er, S., Keskintürk, T. and Daly, C. (2010). GA4Stratification: A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. R package version 1.0. http://CRAN.R-project.org/package=stratification.

Festa, P. (2013). A biased random-key genetic algorithm for data clustering. *SI:BIOCOMP, Math. Biosci.*, 245, 1, 76-85.

Gonçalves, J.F., and Resende, M.G.C. (2004). An evolutionary algorithm for manufacturing cell formation. *Comput. Ind. Eng*, 47, 247-273.

Gonçalves, J.F., and Resende, M. (2011). Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics*, 17, 487-525.

Gonçalves, J.F., Mendes, J.J.M. and Resende, M.G.C. (2005). A hybrid genetic algorithm for the job shop scheduling problem. *Eur. J. Oper. Res*, 167, 77-95.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7749-eng.pdf.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 1, 40, 27-31.

Hidiroglou, M.A., and Kozak, M. (2017). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*, https://doi.org/10.1111/insr.12230.

Keskintürk, T., and Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, 52, 53-67.

Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 2, 205-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10761-eng.pdf.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 5, 797-806.

Kozak, M. (2006). Multivariate sample allocation: Application of a random search method. *Statistics in Transition*, 7, 4, 889-900.

Kozak, M. (2014). Comparison of random search method and genetic algorithm for stratification. *Communications in Statistics – Simulation and Computation*, 43, 2, 249-253.

Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32, 2, 157-163. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9550-eng.pdf.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-eng.pdf.

Lohr, S. (2010). *Sampling: Design and Analysis*, 2nd Ed. Washington: Duxbury Press.

Oliveira, R.M., Chaves, A.A. and Lorena, L.A.N. (2017). A comparison of two hybrid methods for constrained clustering problems. *Applied Soft Computing*, 54, 256-266.

Rao, D.K., Khan, M.G.M. and Reddy, K.G. (2014). Optimum stratification of a skewed population. *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, 8, 3, 497-500.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 2, 191-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-eng.pdf.

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

Spears, W., and De Jong, K. (1991). On the virtues of parameterized uniform crossover. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, 230-236.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 317-337
Statistics Canada, Catalogue No. 12-001-X

An assessment of accuracy improvement by adaptive survey design

Carl-Erik Särndal and Peter Lundquist[1]

Abstract

High nonresponse occurs in many sample surveys today, including important surveys carried out by government statistical agencies. An adaptive data collection can be advantageous in those conditions: Lower nonresponse bias in survey estimates can be gained, up to a point, by producing a well-balanced set of respondents. Auxiliary variables serve a twofold purpose: Used in the estimation phase, through calibrated adjustment weighting, they reduce, but do not entirely remove, the bias. In the preceding adaptive data collection phase, auxiliary variables also play a major role: They are instrumental in reducing the imbalance in the ultimate set of respondents. For such combined use of auxiliary variables, the deviation of the calibrated estimate from the unbiased estimate (under full response) is studied in the article. We show that this deviation is a sum of two components. The *reducible component* can be decreased through adaptive data collection, all the way to zero if perfectly balanced response is realized with respect to a chosen auxiliary vector. By contrast, the *resisting component* changes little or not at all by a better balanced response; it represents a part of the deviation that adaptive design does not get rid of. The relative size of the former component is an indicator of the potential payoff from an adaptive survey design.

Key Words: Nonresponse; Adaptive survey design; Response imbalance; Bias reduction.

1 Introduction

Many important probability sampling surveys have to close the data collection period by declaring a considerable rate of nonresponse. At the estimation stage that follows, the best possible estimates must nevertheless be produced, based on the respondents that a less-than-ideal data collection has delivered. Whatever the techniques used, nonresponse bias remains; it must be kept as low as possible. *Responsive* – or *adaptive – data survey design* for household and other surveys aims at an active control of survey errors and costs, in the planning and data collection phases. One objective is to get a final set of respondents likely to improve prospects for more accurate survey estimates. The concept responsive survey design is due to Groves and Heeringa (2006).

The data collection in a large survey typically extends over a period of time, days or weeks, during which contact attempts are made with the units in the selected probability sample. For some units, a fruitful result is obtained after a few attempts. For other units, despite repeated calls, refusal or non-contact will lead, finally, to declaring them nonrespondents. The background in this article is not that the nonresponse rate will ultimately be reduced to single-digit levels. More realistically, it is that high nonresponse, of the order 30% to 50%, will prevail when the data collection period must necessarily close. Best possible estimates must still be produced.

To improve data collection, a variety of techniques have been suggested and tried: case prioritization, stopping rules, balancing, and others. Case prioritization is considered, for example, in Peytchev, Riley,

1. Carl-Erik Särndal, Statistics Sweden, SE-70185 Örebro, Sweden. E-mail: carl.sarndal@telia.com; Peter Lundquist, Statistics Sweden, SE-10451 Stockholm, Sweden. E-mail: peter.lundquist@scb.se.

Rosen, Murphy and Lindblad (2010), Beaumont, Haziza and Bocci (2014). Stopping rules are discussed in Rao, Glickman and Glynn (2008), and in Wagner and Raghunathan (2010).

Variation between demographic subgroup response rates has been considered in seeking evidence of bias, as in Peytcheva and Groves (2009). Andridge and Little (2011) propose proxy pattern-mixture analysis as a method to assess non-response bias. The nonresponse bias cannot be quantified, but indicators of the risk of bias can be used, as discussed in Wagner (2012), Kreuter, Olson, Wagner, Yan, Ezzati-Rice, Casas-Cordero, Lemay, Peytchev, Groves and Raghunathan (2010), Lohr, Riddles and Morganstein (2016), Nishimura, Wagner and Elliott (2016).

Representativeness and *balance* are terms now frequently used in reference to the set of responding units. Indicators of representativeness, including the R-indicator based on estimated response probabilities, have been developed, see Schouten, Cobben and Bethlehem (2009), Bethlehem, Cobben and Schouten (2011), and Schouten, Shlomo and Skinner (2011), Bianchi, Shlomo, Schouten, da Silva and Skinner (2016).

Balancing is a procedure practiced in data collection, to get in the end a well representative set of respondents. The key is to make auxiliary variable means for the respondents agree closely with corresponding means computable for the probability sample, or known for the population. This lies behind the *response imbalance* statistic IMB, used in Särndal (2011a), Lundquist and Särndal (2013). Methods are available for reducing IMB in adaptive data collection, see for example Särndal and Lundquist (2014).

"Nonresponse adjustment" is an often used term suggesting that an ideal estimator that would perform to satisfaction under full response will need, under nonresponse, "repair" to be a credible surrogate. The goal of the adjustment is to keep the harm - the nonresponse bias - within limits. Current practice in statistical agencies is to use auxiliary variables in the estimation, in computing calibrated nonresponse adjustment weights. Reduced variance and reduced nonresponse bias can be realized. A number of recent contributions, some of them review articles, discuss weighting procedures for the estimation phase, for example, Brick (2013), Fattorini, Franceschi and Maffei (2013), Haziza and Lesage (2016), Little and Vartivarian (2005), Tourangeau, Brick, Lohr and Li (2017). The selection of auxiliary variables for weighting procedures is considered in Särndal and Lundström (2005, 2008, 2010), Särndal (2011b), but is not a topic in the present article.

It is evident that better balanced response could significantly improve accuracy if the estimator in use is rudimentary, such as the response mean expanded by the population size. But the starting point here is that auxiliary variables are used extensively at the estimation stage, and that their use also in adaptive data collection is an added feature that may pay off. Such combined role of auxiliary variables is potentially important.

The question whether balancing the response will significantly reduce nonresponse bias has been raised in the literature, see, for example, Schouten, Cobben, Lundquist and Wagner (2014), Lundquist and Särndal (2013), Särndal and Lundquist (2014), Särndal, Lumiste and Traat (2016). These references point to some, although not strongly pronounced, accuracy improvement from balancing. Gains have appeared modest rather than strong or convincing. Much of the evidence is empirical. The theoretical reasons behind a

"limited success" of adaptive data collection need to be better understood. This article attempts to throw more light on the question: Can a use of auxiliary variables in an adaptive data collection contribute further to improving the estimation, given that such variables are in any case used in a calibrated weighting at the estimation stage? The question is important for research on adaptive survey design. These designs must sustain a clear promise for improved estimates. If little or no improvement is forthcoming, under fairly general conditions, a part of the motivation for adaptive design seems to be lost. We inquire into the theoretical reasons for a certain accuracy gain from better balanced response, more precisely why one might expect "a marginal advantage" from adaptive design, that is, a further improvement of estimates by putting auxiliary variables to work also at the data collection stage. It is perfectly legitimate to re-utilize auxiliary variables at the estimation stage.

The contents are arranged as follows: Notation is introduced (Section 2) for three important population subsets: the probability sample, the response set and its complement, the nonresponse set. The role of the auxiliary vector ($\mathbf{x}$-vector) is emphasized, particularly its role in the response imbalance statistic denoted IMB. To estimate the population y-total, we consider the estimator based on weights calibrated on the auxiliary vector. Its deviation from the unbiased estimator – which is hypothetical because requiring full response – is an indicator of bias that we examine in depth.

We decompose the deviation of the calibration estimator (Section 3) into two terms, the *reducible term* and the *resisting term*. The former can be reduced by adaptive design; in fact, all the way to zero if perfect balance could be realized. The resisting term, by contrast, is little affected by adaptive data collection, hence suggesting that adaptive design is at best a partial remedy for getting rid of nonresponse bias. The special case where the $\mathbf{x}$-vector codes a set of mutually exhaustive and exhaustive sample groups is particularly important (Section 4). Mathematically more tractable, it gives a clearer understanding of the two terms of the decomposition. Empirical evidence using Swedish Labour Force Survey data (Section 5) illustrates and confirms the theoretical conclusions about the two terms. Final comments (Section 6) terminate the article. Proofs are presented in an Appendix.

2 Probability sampling followed by nonresponse

2.1 Population, sample, response set, nonresponse set

We denote by $U = \{1, 2, \ldots, k, \ldots, N\}$ a finite population of size N from which a probability sample s has been drawn, giving unit k the inclusion probability π_k and the sampling weight $d_k = 1/\pi_k$. Let r be the response set, that is, the set of units for which the value y_k of the survey variable y - categorical or continuous – is observed. We thus have observations y_k for $k \in r \subset s \subset U$, but they are missing for the nonresponse set denoted $nr = s - r$. Using the observed y_k we wish to estimate the population y-total $Y = \sum_U y_k$. A summation $\sum_{k \in A}$ over a set of units $k \in A \subseteq U$ is written simply as $\sum_A$.

That "non-respondents are not like respondents" is a well-established fact, and neither group is sufficiently like the whole sample. This causes bias in estimates of the total Y. This article focuses on the

contrast between response set and nonresponse set, something that has been fruitful earlier, in concepts such as the fraction of missing information.

Adaptive data collection is a dynamic process. The response and nonresponse sets, and associated quantities such as means and regression coefficients, are temporarily defined. A more complete notation would distinguish response sets $r^{(a)}$, $a = 1, 2, \ldots,$ in hierarchical sequence,

$$r^{(1)} \subseteq r^{(2)} \subseteq \ldots \subseteq r^{(a)} \subseteq \ldots$$

Here $r^{(a)}$ is the set of units having delivered the value y_k after a call attempts (or, alternatively, after a data collection days), and $nr^{(a)} = s - r^{(a)}$ is the corresponding nonresponse set. But to simplify, we let r refer to any one of the increasingly larger response sets. Tools for data collection, such as Statistics Sweden's WinDATI system, allow us to record all contact attempt and to intervene and redirect the data inflow, so as to get in the end a better balanced final response set.

The *response rate* and corresponding *non-response rate* (both d-weighted) are

$$P = \sum_r d_k \Big/ \sum_s d_k ; \quad Q = 1 - P = \sum_{nr} d_k \Big/ \sum_s d_k .$$

Further notation for response, nonresponse and full sample is given for the auxiliary vector $\mathbf{x}$ in Section 2.2, for the survey variable y and for the regression vector of y on $\mathbf{x}$ in Section 2.3.

2.2 Auxiliary vector and response imbalance

An auxiliary vector $\mathbf{x}$ is chosen following a structured selection of auxiliary variables from a supply of such variables. In the Scandinavian countries, the supply is vast, from administrative sources, in surveys of individuals and households. Much can be said about principles, or "attempts at optimality", that might guide this choice, but this selection procedure is not a topic in this article.

There is a designated auxiliary vector $\mathbf{x}$ of dimension $J \geq 1$. Its value $\mathbf{x}_k$ is assumed known for all units $k \in s$. (The case where the population total of $\mathbf{x}_k$ is known is not considered.) In an important special case, $\mathbf{x}$ is a *group vector* of dimension J, of the form $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, that is, with $J - 1$ entries "0" and a single entry "1" to identify the group that contains unit k. The groups are non-overlapping and exhaustive; the group vector codes J properties of the units. For example, two education categories crossed with three income categories gives $J = 2 \times 3 = 6$ groups. But categorical x-variables used in the vector $\mathbf{x}$ need not all be fully crossed; in many applications they are not.

We use vectors $\mathbf{x}$ with a property that brings mathematical convenience in many derivations: A vector $\boldsymbol{\mu}$ not depending on k exists such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k. \tag{2.1}$$

This is satisfied by a majority of $\mathbf{x}$-vectors of interest, or can be made to hold. For example, if there is a single continuous auxiliary variable x, and $\mathbf{x}_k = (1, x_k)'$, then $\boldsymbol{\mu} = (1, 0)'$ satisfies the requirement. In

the group vector case, the requirement is satisfied by $\boldsymbol{\mu} = (1, \ldots, 1, \ldots, 1)'$. For an example where $\mathbf{x}$ is categorical but not a group vector, suppose two *Education* categories are crossed with three *Income* categories and that *Gender* is added to the vector $\mathbf{x}_k$ as a univariate 0/1 variable, then the dimension is $J = 6 + 1 = 7,$ the number of distinct values $\mathbf{x}_k$ is $6 \times 2 = 12,$ and $\boldsymbol{\mu} = (1, 1, 1, 1, 1, 1, 0)'$ satisfies the requirement (2.1).

The (d-weighted) $\mathbf{x}$-vector means – all computable – are

$$\overline{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \Big/ \sum_r d_k; \quad \overline{\mathbf{x}}_{nr} = \sum_{nr} d_k \mathbf{x}_k \Big/ \sum_{nr} d_k; \quad \overline{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \Big/ \sum_s d_k. \tag{2.2}$$

We need the second order moments; $J \times J$ computable matrices assumed non-singular:

$$\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_r d_k; \quad \boldsymbol{\Sigma}_{nr} = \sum_{nr} d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_{nr} d_k; \quad \boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_s d_k. \tag{2.3}$$

The *imbalance* (IMB) of the response is computed on the auxiliary vector values $\mathbf{x}_k$ known for $k \in s,$ see Särndal (2011a). It is a measure of the contrast between response and full sample, or alternatively between response and nonresponse. For given vector $\mathbf{x}$ and sample $s,$ the imbalance is defined as

$$\text{IMB} = P^2 \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\right).$$

A more telling notation would be $\text{IMB}(r, \mathbf{x} \,|\, s)$ but for simplicity we use just IMB. Because $\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s = (1 - P)\left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr}\right),$ we can express IMB as a contrast between response and nonresponse:

$$\text{IMB} = \{P(1 - P)\}^2 \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr}\right)' \boldsymbol{\Sigma}_s^{-1} \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr}\right). \tag{2.4}$$

An objective for an adaptive data collection design is to create a final response set r with low IMB, for the given probability sample s.

The dimension and the composition of the $\mathbf{x}$-vector are important determinants of the IMB-value. A property of IMB, for fixed r and $s,$ is that it increases when we extend a given $\mathbf{x}$-vector by adding more x-variables to it. The increase reflects the intuition that it is harder to make a higher number of x-variable means come close. The trivial $\mathbf{x}$-vector, $\mathbf{x}_k = 1$ for all $k,$ gives IMB = 0, but it is of virtually no interest in practice. We have $0 \leq \text{IMB} \leq P(1 - P) \leq 0.25$ for any $s,$ any $r,$ and any $\mathbf{x}$-vector. Those are broad conditions; for most survey data sets, the computed IMB-value is much below the upper bound, often in the range 0.01 to 0.05.

2.3 The survey variable and its regression on x

We turn to the survey variable y. Its value y_k is observed for $k \in r$ so linear regression fit of y on $\mathbf{x}$ can be carried out for the set r. Although not feasible in a survey with nonresponse, a linear regression fit of y on $\mathbf{x}$ also exists, conceptually, for $nr = s - r$ and for s. Fitted on the response $r,$ the linear regression coefficient vector (by d-weighted least squares) is $\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$. Analogously, the

other two regression fits give regression vectors $\mathbf{b}_{nr}$ and $\mathbf{b}_s$. Expressed with the second-order moment $\mathbf{x}$-matrices in (2.3), the three regression vectors are

$$\mathbf{b}_r = \boldsymbol{\Sigma}_r^{-1} \sum_r d_k \mathbf{x}_k y_k \Big/ \sum_r d_k;\ \ \mathbf{b}_{nr} = \boldsymbol{\Sigma}_{nr}^{-1} \sum_{nr} d_k \mathbf{x}_k y_k \Big/ \sum_{nr} d_k;\ \ \mathbf{b}_s = \boldsymbol{\Sigma}_s^{-1} \sum_s d_k \mathbf{x}_k y_k \Big/ \sum_s d_k. \quad (2.5)$$

The y-means are

$$\overline{y}_r = \sum_r d_k y_k \Big/ \sum_r d_k; \quad \overline{y}_{nr} = \sum_{nr} d_k y_k \Big/ \sum_{nr} d_k; \quad \overline{y}_s = \sum_s d_k y_k \Big/ \sum_s d_k.$$

The following properties hold as a result of (2.1):

$$\overline{y}_r = \overline{\mathbf{x}}_r' \mathbf{b}_r; \quad \overline{y}_{nr} = \overline{\mathbf{x}}_{nr}' \mathbf{b}_{nr}; \quad \overline{y}_s = \overline{\mathbf{x}}_s' \mathbf{b}_s. \quad (2.6)$$

A rudimentary estimator, under nonresponse, of the population total $Y = \sum_U y_k = N \overline{y}_U$ uses straight expansion (EXP) of the response mean:

$$\hat{Y}_{\text{EXP}} = \hat{N} \sum_r d_k y_k \Big/ \sum_r d_k = \hat{N}\, \overline{y}_r,$$

where $\hat{N} = \sum_s d_k$. In $\hat{Y}_{\text{EXP}}$, the weighting is uniform, without any use of auxiliary information. The bias can be high. A use of the auxiliary values $\mathbf{x}_k$ known for $k \in s$ is usually an improvement; $\sum_s d_k \mathbf{x}_k$ is a known and unbiased (Horvitz-Thompson) estimator of the population total $\sum_U \mathbf{x}_k$, so we seek weights $d_k w_k$ to satisfy the calibration equation $\sum_r d_k w_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. A solution (not a unique one) is

$$w_k = \Big(\sum_s d_k \mathbf{x}_k\Big)' \Big(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1} \mathbf{x}_k = (1/P)\, \overline{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k.$$

The resulting linear calibration (CAL) estimator of $Y = \sum_U y_k$ is $\hat{Y}_{\text{CAL}} = \sum_r d_k w_k y_k$, or equivalently

$$\hat{Y}_{\text{CAL}} = \hat{N}\, \overline{\mathbf{x}}_s' \mathbf{b}_r.$$

A reason why we expect $\hat{Y}_{\text{CAL}}$ to be less biased than $\hat{Y}_{\text{EXP}}$ – especially when y and $\mathbf{x}$ are well related – is that the weights $d_k w_k$ must respect the constraint of an unbiased computable quantity, $\sum_s d_k \mathbf{x}_k$, on the right hand side of the calibration equation. But despite the adjustment weighting, $\hat{Y}_{\text{CAL}}$ has non-negligible, possibly considerable, remaining bias.

As a benchmark we use the unbiased estimator of Y requiring full (FUL) response – therefore hypothetical under nonresponse – namely, the Horwitz-Thompson estimator

$$Y_{\text{FUL}} = \sum_s d_k y_k = \hat{N}\, \overline{y}_s.$$

We refer to the three estimators types as EXP, CAL and FUL. In fact, CAL represents a family of estimators, corresponding to all possible choices of $\mathbf{x}$-vector. For a given suitable $\mathbf{x}$-vector, we shall closely examine the CAL *deviation* Δ_{CAL}, defined as the difference between the biased CAL and the unbiased FUL, scaled by dividing by the (estimated if necessary) population size:

$$\Big(\hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}}\Big) \Big/ \hat{N} = \Delta_{\text{CAL}},$$

where

$$\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s' \mathbf{b}_r - \bar{y}_s = \bar{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s). \tag{2.7}$$

Deviation is not bias. The deviation Δ_{CAL} is an outcome for a response set r from a specific sample s. On the other hand, the bias of the CAL estimator is the expected value of Δ_{CAL} over all r from a given s, and over all s from U. If a response r with zero deviation Δ_{CAL} could be realized whatever the sample s, then the CAL estimator would have zero bias, because it is then always equal to the unbiased Horvitz-Thompson estimator. But to get zero deviation, by adaptive data collection or in some other way, is unrealistic in practice. Nevertheless, it makes good sense to attempt to make the deviation small for the given sample s, since it suggests coming closer to the unbiased estimate. One can say that to get unbiased estimation is not the objective, because impossible when there is nonresponse, but a realistic objective is to reduce deviation from the unbiased estimate.

3 Analyzing the deviation from unbiased estimation

3.1 Decomposing the CAL deviation

We decompose the CAL deviation, $\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$, in order to better understand its behaviour when improved balance is gained by adaptive data collection. We may anticipate that Δ_{CAL} is reduced, if not to zero so at least to a degree, and under conditions that we seek to identify. There are no immediately apparent signs that $\bar{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$ would be reduced when $\bar{\mathbf{x}}_r$ draws closer to the fixed $\bar{\mathbf{x}}_s$. A more in-depth analysis is needed to explain a reduction in Δ_{CAL}, if any.

When the chosen $\mathbf{x}$-vector is of fairly high dimension, the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ (and therefore $\text{IMB} = 0$) is hard to achieve in a data collection; one can strive to come close. Some insight into the nature of Δ_{CAL} is nevertheless gained by setting $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$; using (2.6) we then have

$$\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s) = \bar{\mathbf{x}}_r' \mathbf{b}_r - \bar{y}_s = \bar{y}_r - \bar{y}_s = (1 - P)(\bar{y}_r - \bar{y}_{nr}) \neq 0, \tag{3.1}$$

$$\hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}} = \hat{N} \Delta_{\text{CAL}} = \hat{N} (\bar{y}_r - \bar{y}_s) = \hat{Y}_{\text{EXP}} - \hat{Y}_{\text{FUL}}.$$

Hence, the perfect balance, $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, does not reduce Δ_{CAL} to zero; it gives $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{EXP}} = \hat{N} \bar{y}_r$. This latter equation seems a paradox at first, because $\hat{Y}_{\text{CAL}}$ is supposed to be better than the rudimentary $\hat{Y}_{\text{EXP}}$. But we find that under perfect balance they are equal. The paradox is resolved by noting that $\bar{y}_r - \bar{y}_s$ is likely to be smaller, even considerably so, when the response r is made to be perfectly balanced on $\mathbf{x}$ – in that respect closer to s – than when it is not. And it is not true that auxiliary information is not used; such information is precisely what it takes to get the perfect balance.

Empirical evidence on the relationship between Δ_{CAL} and IMB was given in Särndal and Lundquist (2014). For the survey data analyzed in that article, Δ_{CAL} drops significantly when IMB is reduced but does

not approach zero when IMB tends to zero. These empirical results suggest that Δ_{CAL} levels out at a certain non-zero value.

Under perfect balance, the deviation (3.1) has the same *expression*, $\Delta_{\text{CAL}} = (1-P)(\bar{y}_r - \bar{y}_{nr})$, for any $\mathbf{x}$-vector. But its *value* is not the same, because the set of units in a perfectly balanced response r will be different from one $\mathbf{x}$-vector to the next. For given $\mathbf{x}$-vector, the deviation increases with the rate of nonresponse $1-P$ and with the divergence $\bar{y}_r - \bar{y}_{nr}$ between response and nonresponse y-means. Perfect balance and therefore IMB = 0 hold in particular for the trivial $\mathbf{x}$-vector of no interest in practice, $\mathbf{x}_k = 1$ for all k; in that case the formula $\Delta_{\text{CAL}} = (1-P)(\bar{y}_r - \bar{y}_{nr})$ is a reminder of an elementary but often cited note on the biasing effect of nonresponse: Already Cochran (1977, page 361) gives an analogous expression for the bias of the response set mean, in the setting of a population modeled to consist of a response stratum (units responding with probability one) and nonresponse stratum (units responding with probability zero). The message is: Bias increases with the rate of nonresponse and with the separation between response and nonresponse means.

We seek to decompose $\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s'(\mathbf{b}_r - \mathbf{b}_s)$ into two components such that one of them is reducible to zero under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. Several splits may meet this "null requirement" on one of the two terms. Särndal and Lundquist (2017) examine one such split, $\Delta_{\text{CAL}} = \bar{e}_r + u_r$, where $\bar{e}_r = \sum_r d_k e_k \big/ \sum_r d_k = \bar{\mathbf{x}}_r'(\mathbf{b}_r - \mathbf{b}_s)$ is the mean, over the response r, of the residuals from the regression of y on $\mathbf{x}$ fitted on the sample s, $e_k = y_k - \mathbf{x}_k'\mathbf{b}_s$, and u_r is the remainder: $\Delta_r - \bar{e}_r = u_r = -(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s)$. These authors call $\bar{e}_r$ the *regression inconsistency*, to express that the regression model for the full sample fails – or is inconsistent – when applied to the response: The residuals e_k have zero mean, $\bar{e}_s = 0$, over the sample s, but non-zero mean, $\bar{e}_r \neq 0$, over the response r. The inconsistency is a (not surprising) result of "missing not at random".

We have $u_r = 0$ under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, so then $\Delta_{\text{CAL}} = \bar{e}_r$. Särndal and Lundquist (2017) give empirical evidence that when the imbalance IMB is reduced through adaptive data collection, both Δ_{CAL} and $\bar{e}_r$ are reduced, neither of them to zero, and that the ratio $\bar{e}_r / \Delta_{\text{CAL}}$ tends slowly toward 1 when IMB approaches zero.

3.2 Contrasting the response with the nonresponse

Some insight is gained by seeking a decomposition of $\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s'(\mathbf{b}_r - \mathbf{b}_s)$ so that a first term remains roughly constant when the imbalance is reduced, while a second term tends to zero in that process. Such a decomposition would thus isolate a part of the deviation that is untouched by adaptive design and indicate why adaptive design is at best "partially successful" for bias removal. The decomposition in Result 3.1 is driven by a wish to contrast the response set r with the nonresponse set $nr = s - r$. The notation is given in (2.2) for the $\mathbf{x}$-means and in (2.5) for the regression vectors.

Result 3.1. The CAL deviation $\Delta_{\text{CAL}} = \bar{\mathbf{x}}_s'(\mathbf{b}_r - \mathbf{b}_s)$ has the decomposition $\Delta_{\text{CAL}} = D_1 + D_2$ where

$$D_1 = (1-P)\bar{\mathbf{x}}_s'(\mathbf{b}_r - \mathbf{b}_{nr}); \quad D_2 = -P(1-P)(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})'(\mathbf{b}_r - \mathbf{b}_{nr}).$$

Equivalent expressions are

$$\Delta_{\text{CAL}} = (1-P)\,\bar{\mathbf{x}}'_{nr}\,(\mathbf{b}_r - \mathbf{b}_{nr}); \quad D_2 = P\,(\bar{\mathbf{x}}_s - \bar{\mathbf{x}}_r)'\,(\mathbf{b}_r - \mathbf{b}_{nr}).$$

The Appendix, part 1, shows how the components D_1 and D_2 are derived from the definition (2.7) of Δ_{CAL}. The result invites several comments:

1. Balancing the response during data collection can bring $\bar{\mathbf{x}}_r$ close to $\bar{\mathbf{x}}_s$ and thus a reduction of both IMB and D_2 toward zero. Under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, $D_2 = 0$ but $D_1 \neq 0$.
2. To get $D_1 = 0$ would require $\mathbf{b}_r = \mathbf{b}_{nr}$. This does not happen even under perfect balance. The sample s and its mean $\bar{\mathbf{x}}_s$ are fixed. To eliminate the term D_1 seems out of reach. But does a "better composition" of the response set r – closeness of $\bar{\mathbf{x}}_r$ to $\bar{\mathbf{x}}_s$ and lower IMB – have at least some favourable influence on the difference $\mathbf{b}_r - \mathbf{b}_{nr}$ and therefore on D_1?

Simple general answers are hard to obtain. Intuitively, a transfer of some units, during data collection, from nonresponse $nr = s - r$ to response r suggests that $\mathbf{b}_r - \mathbf{b}_{nr}$ is not much affected, leaving D_1 rather constant. This becomes more explicit when $\mathbf{x}$ is a group vector, that is, of the form $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$; see Section 4. We shall call D_2 *the reducible term* and D_1 *the resisting term* of the deviation $\Delta_{\text{CAL}} = D_1 + D_2$. It should be noted that here we consider a fixed choice of $\mathbf{x}$-vector. Changing that vector may of course have a certain effect both on D_1 and on D_2.

4 The group vector case

4.1 The terms of the decomposition

We consider the important case of a group vector, $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$. In practice, its dimension J is often considerable, 30 or more, as when several categorical x-variables are fully crossed. The diagonal form of matrices requiring inversion in (2.5) brings considerable simplification. For $j = 1, 2, \ldots, J$, denote by s_j the subset of the sample s that falls in group j. Its proportion of the sample is $W_j = \sum_{s_j} d_k \Big/ \sum_s d_k$. The response set in s_j is denoted r_j; the nonresponse set is $nr_j = s_j - r_j$. The group j nonresponse rate is $Q_j = \sum_{nr_j} d_k \Big/ \sum_{s_j} d_k = 1 - \sum_{r_j} d_k \Big/ \sum_{s_j} d_k = 1 - P_j$; the overall rate is $Q = \sum_{j=1}^{J} W_j Q_j = 1 - P$. The imbalance becomes a variance of group response rates, see, for example, Särndal (2011a),

$$\text{IMB} = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)^2 = \sum_{j=1}^{J} W_j \left(P_j - P\right)^2.$$

The y-means are $\bar{y}_{r_j} = \sum_{r_j} d_k y_k \Big/ \sum_{r_j} d_k$; $\bar{y}_{nr_j} = \sum_{nr_j} d_k y_k \Big/ \sum_{nr_j} d_k$; $\bar{y}_{s_j} = \sum_{s_j} d_k y_k \Big/ \sum_{s_j} d_k$. The J-dimensional column vector $\mathbf{b}_r - \mathbf{b}_{nr}$ has elements $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$, $j = 1, 2, \ldots, J$. We call δ_j *the group j divergence*, namely, between response y-mean and nonresponse y-mean. The calibration estimator is $\hat{Y}_{\text{CAL}} = \sum_{j=1}^{J} \hat{N}_j \bar{y}_{r_j}$, where $\hat{N}_j = \sum_{s_j} d_k$, and the unbiased benchmark estimator is $\hat{Y}_{\text{FUL}} = \sum_{j=1}^{J} \hat{N}_j \bar{y}_{s_j}$, so

$$\Delta_{\mathrm{CAL}} = \left(\hat{Y}_{\mathrm{CAL}} - \hat{Y}_{\mathrm{FUL}}\right)\Big/ \hat{N} = \sum_{j=1}^{J} W_j \left(\overline{y}_{r_j} - \overline{y}_{s_j}\right) = \sum_{j=1}^{J} W_j Q_j \delta_j .$$

Straightforward application of Result 3.1 to the group vector case gives the following Result 4.1.

Result 4.1. When $\mathbf{x}$ is a group vector of dimension J, $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, the terms of the decomposition in Result 3.1 take the form

$$D_1 = Q\sum_{j=1}^{J} W_j \delta_j ; \quad D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j ; \quad \Delta_{\mathrm{CAL}} = D_1 + D_2 = \sum_{j=1}^{J} W_j Q_j \delta_j ,$$

where $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$, and $Q_j = \sum_{nr_j} d_k \Big/ \sum_{s_j} d_k$ and $Q = \sum_{j=1}^{J} W_j Q_j$ are, respectively, the group j nonresponse rate and the overall nonresponse rate.

To what degree can adaptive data collection and better balanced response r improve accuracy and reduce the deviation $\Delta_{\mathrm{CAL}} = D_1 + D_2$? In the discussion that follows, "reduction" should be understood as "reduction in absolute value", "smaller" as "smaller in absolute value", because Δ_{CAL} and its components can have either sign.

In an adaptive data collection acting on an $\mathbf{x}$-vector that codes quite many groups, it is hard to arrive at the perfect balance where $Q_j = Q$ for every group j, and IMB = 0. But it may be possible to bring all Q_j in fairly close agreement with the overall nonresponse rate Q; then both IMB and D_2 may come close to zero. We should think in terms of a scenario where an actual data collection, obeying a predefined protocol that remains unchanged to the end, is compared with alternative data collections that strive actively, through interventions, to end in a response with low imbalance. This is the format for the empirical validation (in Section 5), based on an important Statistics Sweden survey: There is a response set as actually recorded in the survey, and several alternative response sets are derived by experimental interventions in the actual response to get successively lower imbalance.

We address the following issues about the terms in Result 4.1:

(a) When the nonresponse differentials $Q_j - Q$ move closer to zero, so that IMB is reduced, the term $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j$ tends to be reduced, possibly to near zero. The prospect is quite different for $D_1 = Q\sum_{j=1}^{J} W_j \delta_j$. A priori it is not excluded that the divergences δ_j will be somewhat reduced in the process. If this were to happen in a number of groups, then D_1 might also be reduced. But under certain conditions, a great change in D_1 is unlikely; we expect D_1 to be little affected by a reduced IMB. We justify this first by a model assisted theoretical argument in Section 4.6, then observe it empirically with actual survey data in Section 5.

(b) It is not evident that D_1 and D_2 always have the same sign. When they do, an obtainable reduction of D_2 brings a reduced deviation $\Delta_{\mathrm{CAL}} = D_1 + D_2$, thus an improved accuracy. Under what conditions do they have the same sign?

(c) When D_1 and D_2 have the same sign and D_1 stays roughly constant when IMB is reduced, then any reduction of $\Delta_{\mathrm{CAL}} = D_1 + D_2$ comes from a reduction of D_2; the relative size of the two

terms then determines whether such gain by adaptive design is considerable or just modest or even trivial.

4.2 Within-group correlation between response and survey variable

Response and survey variable y are correlated, in the whole sample s, as within any sample group s_j. This unavoidable consequence of "missing not at random" is a key to interpreting D_1, D_2 and their sum Δ_{CAL}. In group j, let ρ_j be the coefficient of correlation between response indicator i ($i_k = 1$ for $k \in r_j$, $i_k = 0$ for $k \in nr_j$) and survey variable y. This ρ_j has a simple relation, through a multiplicative factor, to the group divergence $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$:

$$\rho_j = \sqrt{P_j\left(1 - P_j\right)}\ \delta_j \big/ S_{yj}, \tag{4.1}$$

where $S_{yj}^2 = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(y_k - \bar{y}_{s_j}\right)^2$ is the y-variance in group j. The expression (4.1) follows (see Appendix, part 2) from the customary definition of correlation coefficient as "covariance divided by the product of the two standard deviations", $\rho_j = S_{iyj} \big/ \left(S_{ij} S_{yj}\right)$.

In the special case where y is dichotomous $0/1$ (as when $y_k = 1$ if k is "Employed", $y_k = 0$ otherwise), then $\bar{y}_{s_j} = \sum_{s_j} d_k y_k \big/ \sum_{s_j} d_k$ is a proportion, namely, the (design weighted) proportion of "1" in group j (the employment rate in group j), so the within-group y-variance is $S_{yj}^2 = \bar{y}_{s_j}\left(1 - \bar{y}_{s_j}\right)$.

4.3 Analysis of the resisting term

We now further examine the behavior of D_1 and D_2 for the group vector case in Result 4.1. Their (relative) sizes depend on several factors. The resisting term $D_1 = Q\sum_{j=1}^{J} W_j \delta_j$ depends on the sizes and the distribution over the J groups of the divergences $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$, or, in an alternative view, of the within-group correlations $\rho_j = \sqrt{P_j\left(1 - P_j\right)}\,\delta_j \big/ S_{yj}$.

In a typical survey setting, a look at the distribution of δ_j, $j = 1, \ldots, J$, where J may be 30 or more, will typically reveal a mixture of positive and negative values. There may be "critical groups" with uncharacteristically large divergence δ_j, of the same sign, be it positive or negative. Such δ_j are influential for the (weighted) mean $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ and therefore for $D_1 = Q\bar{\delta}$. The majority of the δ_j may be quite small, although non-zero.

Example 4.1. Consider y = *Employed* ($y_k = 1$ if unit k is employed, $y_k = 0$ if not). Positive within-group correlation ρ_j between *Employed* and *Response* ($i_k = 1$ for $k \in r_j$, $i_k = 0$ for $k \in nr_j$) may be expected for groups of persons with one or more of characteristics such as male, young, low education and/or foreign origin. Such groups tend to be low (comparatively speaking) both on *Response* and on *Employed*. Alternatively expressed, employment rate is higher among respondents than among non-respondents, thus δ_j positive. On the other hand, groups with characteristics such as middle aged, well educated and/or home owner tend to be high on both *Response* and *Employed*, so δ_j is likely to be positive for those groups as well. Another y-variable which may have a similar pattern is *Income*: In some sample

groups, *Income* may be higher for respondents than for non-respondents, so δ_j again positive. Thus for both y-variables, several distinctly positive and influential δ_j may make $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ and $D_1 = Q\,\overline{\delta}$ positive. This is confirmed in Section 5 for Swedish Labour Force Survey data. However in other countries, a variable such as *Income* may behave differently: In some groups, income may instead be higher for non-respondents than for respondents, making δ_j negative, as when wealthy or high income persons tend to respond relatively less. This illustrates the difficulty in anticipating the sign of the group divergences for some survey variables.

4.4 Analysis of the reducible term

We can write the reducible term as $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\left(\delta_j - \overline{\delta}\right)$. It is the *covariance* between nonresponse $Q_j = 1 - P_j$ and divergence $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$. Here D_2 is a *covariance across groups*, $j = 1, \ldots, J$, whereas the divergence δ_j given in (4.1) reflects *covariance within group* j between response indicator i_k and survey variable y_k. It is hard to draw general conclusions about the size of D_2 and whether or not it has the same sign as D_1. It depends on a number of factors.

Example 4.1, continued. As argued in Example 4.1, a number of positive divergences δ_j are likely to occur for y = *Employed* ($y_k = 1$ if k is Employed, = 0 if not), for example, for groups characterized by male, young and/or low education. Then $D_1 = Q\overline{\delta} = Q\sum_{j=1}^{J} W_j \delta_j$ may be distinctly positive. But whether high nonresponse rate Q_j tends to go together with high divergence δ_j, so as to make the covariance D_2 also positive, this is less evident. It may happen for variables such as *Employed* and *Income*; it is so for the Swedish data in Section 5.

4.5 Do the two terms have same sign?

General conclusions about the signs of D_1 and D_2 do not seem possible. Too many factors are involved, the population, the survey design, the particular survey variable y, and others. When D_1 and D_2 have the same sign, a reduction of D_2 toward zero can give a (perhaps considerably) reduced deviation $\Delta_{\mathrm{CAL}} = D_1 + D_2$. On the other hand, if the terms are of opposite sign, they counteract each other; striving for low imbalance and low D_2 may not reduce $\Delta_{\mathrm{CAL}} = D_1 + D_2$. In this undesirable and perhaps rare situation, attempts at balancing the response – reducing D_2 – would defeat the purpose of achieving a lower deviation Δ_{CAL}; it may get larger, not smaller. Let us examine this "same sign issue".

Example 4.1 and its continuation in Section 4.4 outlined a situation, for the y-variables *Employed* and *Income*, where the mean $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ and the covariance $\sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j = D_2$ are likely to have the same sign, so that $D_1 = Q\overline{\delta}$ and D_2 have the same sign. But more generally, opposite signs are not excluded. An example is when uncharacteristically low nonresponse, $Q_j - Q < 0$, happens in influential groups with distinctly positive divergence δ_j. This may cause D_2 to be negative, while D_1 is positive. Then the two terms counteract each other. A positive prospect remains, however, when the two terms have the same sign: We can reduce D_2 to low levels by making all group nonresponse rates nearly Q_j equal.

Then, supposing that D_1 changes little, as under conditions suggested below in Section 4.6, a considerable reduction of the deviation Δ_{CAL} can take place. These findings are summarized as follows.

Proposition 4.1. For the group vector case, suppose that the mean divergence $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ and the covariance $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right) \delta_j$ have the same sign, say positive. Then $D_1 = Q\bar{\delta}$ and D_2 are positive terms in the deviation $\Delta_{\text{CAL}} = D_1 + D_2 = \sum_{j=1}^{J} W_j Q_j \delta_j$. An adaptive data collection that reduces the imbalance $\text{IMB} = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)^2$ will reduce D_2 (to zero if IMB = 0), whereas the resisting term D_1 is likely to change little.

The choice of $\mathbf{x}$-vector, that is, the choice of variables entering into it and its dimension, can have important effects on D_1 and D_2. Suppose we double the number of groups coded by the group vector, as when $J = 8 = 2^3$ already existing groups (obtained by crossing three dichotomous x-variables) are doubled in number to $J = 16 = 2^4$, by complete crossing on a fourth dichotomous x-variable. A certain skewness in the distribution of the divergences $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$, $j = 1, \ldots, J$, is likely to persist. For a fixed response r, D_1 and D_2 may be reduced by extending the $\mathbf{x}$-vector. The empirical evidence in Section 5 throws some light on this question.

4.6 Insensitivity of group divergences

For the group vector case in Result 4.1, the resisting term is $D_1 = Q\bar{\delta}$, where $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ is a weighted mean of the divergences δ_j, $j = 1, \ldots, J$. We would like to know more about how the δ_j evolve in a data collection that may extend, in a large survey, over a period of time.

An adaptive data collection is monitored during the period. Assessment and modification of protocol and emphasis may take place, at one or more intervention points. For example, units whose computed response propensity is seen to be low, at a given point, may become henceforth specially targeted, so as to boost their response propensity towards an overall mean propensity. Some so far non-responding units become converted into respondents; their *response status* is changing.

In the group vector case, the sample group s_j consists, at a given point in the data collection period, of a certain response set and a corresponding nonresponse set. Contact attempts continue with units having not yet answered; a change in response status occurs for some units in s_j. A few more become respondents, causing a certain change in the divergence δ_j. Under certain conditions, however, δ_j may be little affected. If this were to hold for most or all groups, D_1 would change little. Let us examine this possibility.

That δ_j would change little only cannot be observed in a real survey, because y_k is missing for the nonresponse, but it can be made plausible by a model for response status change. We consider models where units within one and the same group are *exchangeable* or *substitutable* for one another, with respect to response status: At a given point in the data collection, all units are considered alike in regard to response status change; transfer from nonresponse into response, or vice versa, is equally probable for all units.

Such a model can be justified for a sufficiently detailed breakdown of the sample s into small and homogenous groups s_j, such that the units in one and the same group share important characteristics. It

would be unrealistic in a large heterogeneous group with a mixture of units very different in regard to age, gender, income, education and other important aspects. In a heterogeneous group, some units have high conversion probability, others are unlikely to become converted. For example, if a group contains both older, well-educated persons and younger, less educated persons, conversion probability may be high for the former subgroup, considerably lower for the latter.

At a certain point in the data collection, let us consider two models for response status change in a sample group s_j. We assume for simplicity that s is a simple random sample from U, of size N, so the sampling weights are constant, $d_k = N/n$ for all k.

Model 4.1. Suppose transfer takes place, in the group s_j, from the nonresponse set $nr_j = s_j - r_j$ to the response set r_j (a conversion) in such a way that all transfer sets $tr_j \subset nr_j$ of fixed size q_j are equally likely to occur.

In this model, nr_j and r_j are fixed sets with respective sizes $n_j - m_j$ and m_j, while tr_j is a simple random selection of q_j non-respondents. Any particular transfer set tr_j of q_j units can get converted, and with equal probability. Every unit in the nonresponse set nr_j has the same conversion probability, $q_j/(n_j - m_j)$.

Before transfer, the divergence in group j is $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$. After transfer, indicated by a star * in the indices, the new divergence is $\delta_{*j} = \bar{y}_{*r_j} - \bar{y}_{*nr_j}$. Its expected value under Model 1 satisfies

$$P_{*j} E(\delta_{*j}) = P_j \delta_j, \tag{4.2}$$

where the new group response rate is $P_{*j} = (m_j + q_j)/n_j > m_j/n_j = P_j$. The proof is given in the Appendix, part 3. If the group response rate change is small, as when q_j is small compared with n_j, then $P_{*j} \approx P_j$, and the transfer may bring little change to the divergence. Its new value δ_{*j} is just marginally smaller, in expectation, than the before-transfer value δ_j.

In the empirical Section 5, we experiment with response sets that get smaller instead of larger. Model 4.2 deals with a transfer in that direction, and leads to a similar conclusion.

Model 4.2. Suppose a transfer takes place within group j from the response set r_j to the nonresponse set $nr_j = s_j - r_j$ in such a way that all transfer sets $tr_j \subset r_j$ of fixed size q_j are equally likely to occur.

A derivation analogous to the one that gave (4.2) shows that

$$Q_{*j} E(\delta_{*j}) = Q_j \delta_j. \tag{4.3}$$

The new (higher) nonresponse rate is $Q_{*j} = 1 - (m_j - q_j)/n_j > 1 - m_j/n_j = Q_j$. The new divergence δ_{*j} is just slightly smaller in expected value, when q_j is small compared with n_j. The proof is analogous to that of (4.2).

The results (4.2) and (4.3) state that if it does not matter which particular units in group j change their response status, then the divergence δ_j stays nearly the same, in expectation. But the assumption in

Models 4.1 and 4.2 of equally probable transfer sets is hard or impossible to substantiate in practice. It would be hard to specify a vector $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ such that the groups coded by the vector will obey the models. We cannot claim "truth" of these models, only suggest that they may be plausible under a well-diversified grouping of the sample.

5 Empirical evidence

The empirical work reported in this section illustrates some of the theory in earlier sections. We used survey data from Statistics Sweden's Labour Force Survey (LFS) in its 2012 edition. Results are given in Tables 5.1 to 5.4.

We created the LFS2012 data set by combining the 12 wave-one samples in 2012. The monthly first-wave size counts approximately 2,650 units (persons). This LFS2012 data set is treated as a simple random sample s of size n = 32,265. Response or nonresponse in the actual data collection is recorded and available for all those units. The response rate in the actual data collection was 70.6%. The data and the construction of the experimental response sets in the tables are described in further detail in Särndal and Lundquist (2017).

In analyzing the LFS2012 data set, we used different $\mathbf{x}$-vectors obtained by crossing binary x-variables: *Educ*, equal to 1 for a person with high education, 0 otherwise; *Owner*, equal to 1 for an owner of his/her place of residence, 0 otherwise; *Origin*, equal to 1 for person born in Sweden, 0 otherwise; *Civil*, equal to 1 for married or widower, 0 otherwise; *Gender*, equal to 1 for male, 0 otherwise. Results are presented here for two $\mathbf{x}$-vectors.

The $\mathbf{x}$-vector in Tables 5.1 and 5.3 represents the crossing of the first three binary variables: $\mathbf{x} = \mathbf{x}_1$ = (*Educ* × *Owner* × *Origin*); its dimension, equal to the number of groups, is $J = 2^3 = 8$. The $\mathbf{x}$-vector in Tables 5.2 and 5.4 was obtained crossing also by the binary *Civil*: $\mathbf{x} = \mathbf{x}_2$ = (*Educ* × *Owner* × *Origin* × *Civil*), of dimension $J = 2^4 = 16$.

In this experimental study, we used two y-variables, *Employed* and *Income*. Both are register variables; with values y_k available for all units $k \in s$. They are thus "pseudo y-variables" rather than real survey y-variables. Knowing y_k for $k \in s$ allows us to compute regression coefficients and y-means both for the response and for the nonresponse. The y-variable is *Employed* in Table 5.1 (with $\mathbf{x} = \mathbf{x}_1$) and in Table 5.2 (with $\mathbf{x} = \mathbf{x}_2$). *Employed* is binary, with value $y_k = 1$ if k is an employed person, zero otherwise. The y-variable is *Income* in Table 5.3 (with $\mathbf{x} = \mathbf{x}_1$) and in Table 5.4 (with $\mathbf{x} = \mathbf{x}_2$). *Income* is a continuous register variable, available from the Swedish tax register. We standardized *Income* to have zero mean and unit variance over s. Considerable variability and skewness of *Income* creates some volatility in the results. One individual y_k-value can have considerable impact within a smallish group, as compared with the more stable performance of the $0/1$ y-variable *Employed*.

The four rows of Tables 5.1 to 5.4 refer to a series of four different response sets. The important feature is that they are, by construction, sets with progressively lower IMB. The first, *Act*, is the response set

recorded in the actual 2012 Labour Force Survey data collection for the 32,265 units (persons). The last three response sets, *A65*, *A63*, *A60*, taken from Särndal and Lundquist (2017), are constructed from *Act* by the threshold method to have successively lower imbalance IMB.

To illustrate, the response set *A65* was created from *Act* by dropping, at each of a sequence of intervention points, those responding units in *Act* whose computed response propensity exceeds the threshold 0.65. This tends to even out differences in response propensity, so this construction reduces IMB, and the overall response rate P drops somewhat. The response sets denoted *A63* and *A60* were obtained by setting the threshold at 0.63 and 0.60, respectively; IMB and P are again reduced.

The table columns show: Response rate P, imbalance IMB (multiplied by 10^2), the components D_1 and D_2 of the deviation $\Delta_{\text{CAL}} = D_1 + D_2$ (all three multiplied by 10^2), the D_2 proportion of Δ_{CAL}, $\text{Prop}D_2 = 100 \times D_2/\Delta_{\text{CAL}}$, and finally the D_1 size relative, $\text{Rel}\, D_1 = D_1 \big/ \text{mean}\,(D_1)$, where $\text{mean}\,(D_1)$ is the arithmetic mean of the four table values of D_1. We use $\text{Rel}\, D_1$ to see if it is near one for all rows, in line with the contention that D_1 is little affected by a reduced imbalance IMB.

The results in Tables 5.1 to 5.4 prompt the following observations, of which the second and third are particularly interesting, in that they confirm what theory in earlier sections suggests, namely that when IMB is reduced, $\text{Prop}D_2$ drops quite distinctly, whereas $\text{Rel}\, D_1$ stays very close to one.

1. In all four tables, D_1 and D_2 have the same sign. Both are positive, and the reduced IMB (from first to fourth row) brings a reduction in $\Delta_{\text{CAL}} = D_1 + D_2$, due almost entirely to the drop in D_2.
2. In each table, the relatives $\text{Rel}\, D_1$ are not far from 1. Thus D_1 is remarkably constant over the four rows (response sets), thus insensitive to the reduced IMB.
3. In each table, D_2 and $\text{Prop}D_2$ are decreasing over the four rows, as theory makes us expect. In fact, $\text{Prop}D_2$ tends to zero with IMB. The effect of the y-variable is important; $\text{Prop}D_2$ is considerably greater for *Income* than for *Employed*.
4. The change of $\mathbf{x}$-vector is an important influence on D_1, for both y-variables. Going from the smaller $\mathbf{x} = \mathbf{x}_1$ (Tables 5.1 and 5.3) to the more extensive $\mathbf{x} = \mathbf{x}_2$ (Tables 5.2 and 5.4) brings considerable reduction in D_1, whereas D_2 changes very little.

We also examined the distribution of the J group divergences δ_j, $j = 1, \ldots, J$, for the vectors $\mathbf{x}_1$ (with $J = 8$), $\mathbf{x}_2$ (with $J = 16$) and $\mathbf{x}_3 = (\textit{Educ} \times \textit{Owner} \times \textit{Origin} \times \textit{Civil} \times \textit{Gender})$ (with $J = 32$). For both *Employed* and for *Income*, and for all four response sets, there are, not unexpectedly, a few large positive δ_j and a certain skewness in the distribution. For both variables, $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ is clearly positive. That is, on average over the groups, y-means are, for these data, higher for respondents than for non-respondents. It is a feature of those particular y-variables.

For $\mathbf{x}_3$, a plot of the 32 divergences δ_j against the nonresponse differential $Q_j - Q$ shows a majority of points in the vicinity of zero on both axes, and scattered values in the four quadrants of the plot. The plot does suggest a positive, although not very pronounced, correlation between δ_j and $Q_j - Q$, which is what makes the covariance term D_2 positive.

Table 5.1
Survey variable y = Employed; x-vector: (*Educ* × *Owner* × *Origin*). Rows: Four response sets. Columns: Response rate P, imbalance IMB (multiplied by 10^2), components D_1 and D_2 of $\Delta_{CAL} = D_1 + D_2$ (all three multiplied by 10^2), Prop D_2 and Rel D_1

Resp set	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Act	0.706	0.608	0.558	0.151	0.709	21.3	0.96
A65	0.659	0.135	0.586	0.098	0.684	14.2	1.01
A63	0.648	0.113	0.596	0.086	0.682	12.6	1.03
A60	0.625	0.062	0.579	0.058	0.637	9.3	1.00

Table 5.2
Survey variable y = Employed; x-vector: (*Educ* × *Owner* × *Origin* × *Civil*). Rows: Four response sets. Columns: Response rate P, imbalance IMB (multiplied by 10^2), components D_1 and D_2 of $\Delta_{CAL} = D_1 + D_2$ (all three multiplied by 10^2), Prop D_2 and Rel D_1

Resp set	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Act	0.706	0.672	0.459	0.153	0.612	25.0	0.92
A65	0.659	0.165	0.515	0.101	0.616	16.4	1.03
A63	0.648	0.142	0.524	0.083	0.607	13.7	1.05
A60	0.625	0.088	0.493	0.067	0.560	12.0	0.99

Table 5.3
Survey variable y = Income; x-vector: (*Educ* × *Owner* × *Origin*). Rows: Four response sets. Columns: Response rate P, imbalance IMB (multiplied by 10^2), components D_1 and D_2 of $\Delta_{CAL} = D_1 + D_2$ (all three multiplied by 10^2), Prop D_2 and Rel D_1

Resp set	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Act	0.706	0.608	0.668	0.648	1.316	49.2	1.26
A65	0.659	0.135	0.479	0.261	0.740	35.3	0.90
A63	0.648	0.113	0.449	0.250	0.699	35.8	0.84
A60	0.625	0.062	0.530	0.169	0.699	24.2	1.00

Table 5.4
Survey variable y = Income; x-vector: (*Educ* × *Owner* × *Origin* × *Civil*). Rows: Four response sets. Columns: Response rate P, imbalance IMB (multiplied by 10^2), components D_1 and D_2 of $\Delta_{CAL} = D_1 + D_2$ (all three multiplied by 10^2), Prop D_2 and Rel D_1

Resp set	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Act	0.706	0.672	0.324	0.639	0.963	66.4	0.98
A65	0.659	0.165	0.327	0.247	0.574	43.0	0.99
A63	0.648	0.142	0.313	0.232	0.545	42.6	0.95
A60	0.625	0.088	0.355	0.166	0.521	31.9	1.08

6 Concluding comments

Behind this article lies the question: If calibrated weighting adjustment at the estimation stage removes *some* of the nonresponse bias in the estimates, why cannot a use of auxiliary variables also in a preceding adaptive data collection remove *the rest* of the bias? Grounds for believing so may be that after an adaptive data collection, one can have a final set of respondents that is in so many respects a copy of the selected (but nonresponse ridden) probability sample that no appreciable bias should remain. We have examined the calibrated weighting estimator and its deviation Δ_{CAL} from the unbiased estimator requiring full response. It is a theoretical examination, because in a real survey with nonresponse, the unbiased (Horvitz-Thompson) estimator is not available.

Respondents generally differ systematically from non-respondents. With this difference in mind, we were able to write Δ_{CAL} as a sum of a *resisting term* D_1 and a *reducible term* D_2. For a sample divided into subgroups, the reducible term D_2 is determined by the covariance (over the groups) between group nonresponse rate and within-group correlation between *Response* and y-variable. Thus D_2 can be reduced to zero if all group nonresponse rates can be made equal in an adaptive data collection. But adaptive data collection does not get rid of the resisting term D_1. This is in one sense a sobering message: The deviation from the unbiased estimate is not eliminated. But on the other hand, adaptive data collection can promise a better starting point for the estimation phase beginning after a terminated data collection.

Appendix

Part 1. Derivation of the decomposition $\Delta_{\mathrm{CAL}} = D_1 + D_2$ in Result 3.1. By definition $\Delta_{\mathrm{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s) = \mathbf{x}_s' \mathbf{b}_r - \overline{y}_s$ by a use of (2.6). Substitute $\overline{\mathbf{x}}_s = P\overline{\mathbf{x}}_r + (1-P)\overline{\mathbf{x}}_{nr}$ and $\overline{y}_s = P\overline{y}_r + (1-P)\overline{y}_{nr} = P\overline{\mathbf{x}}_r'\mathbf{b}_r + (1-P)\overline{\mathbf{x}}_{nr}'\mathbf{b}_{nr}$. This gives $\Delta_{\mathrm{CAL}} = (1-P)\overline{\mathbf{x}}_{nr}'(\mathbf{b}_r - \mathbf{b}_{nr})$. Finally, substitute $\overline{\mathbf{x}}_{nr} = \overline{\mathbf{x}}_s - P(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr})$ to arrive at the terms D_1 and D_2 in Result 3.1. That the two expressions for D_2 are equivalent follows from $(1-P)(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr}) = \overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s$.

Part 2. Derivation of the within-group correlation coefficient (4.1) between response indicator i and survey variable y. By definition, the correlation is $\rho_j = S_{iyj} / S_{ij} S_{yj}$, where the covariance is $S_{iyj} = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(i_k - \overline{i}_{s_j}\right)\left(y_k - \overline{y}_{s_j}\right)$ with $\overline{i}_{s_j} = \sum_{s_j} d_k i_k \Big/ \sum_{s_j} d_k = P_j$. A development gives $S_{iyj} = P_j\left(1-P_j\right)\delta_j$ with $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$. The y-variance is $S_{yj}^2 = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(y_k - \overline{y}_{s_j}\right)^2$, and S_{ij}^2 is analogous with $\left(i_k - \overline{i}_{s_j}\right)^2$ replacing $\left(y_k - \overline{y}_{s_j}\right)^2$, so $S_{ij}^2 = P_j\left(1-P_j\right)$. The result $\rho_j = \sqrt{P_j\left(1-P_j\right)}\,\delta_j / S_{yj}$ follows.

Part 3. Proof of (4.2): Under Model 4.1, r_j and nr_j are fixed sets, with respective sizes m_j and $n_j - m_j$ and fixed means $\overline{y}_{r_j}$ and $\overline{y}_{nr_j}$. The transfer set tr_j of fixed size q_j is random, withdrawn by simple random

sampling from the nonresponse $nr_j = s_j - r_j$ and transferred to the response r_j. The new y-means, for response and nonresponse, are

$$\bar{y}_{*r_j} = \left(\sum_{r_j} y_k + \sum_{tr_j} y_k\right) \Big/ \left(m_j + q_j\right); \quad \bar{y}_{*nr_j} = \left(\sum_{nr_j} y_k - \sum_{tr_j} y_k\right) \Big/ \left(n_j - m_j - q_j\right).$$

Because tr_j is a simple random sample from the fixed nr_j, the transfer set mean $\bar{y}_{tr_j} = \sum_{tr_j} y_k \Big/ q_j$ has expected value $\bar{y}_{nr_j}$. Therefore, the expected values of the new y-means are

$$E\left(\bar{y}_{*r_j}\right) = \left(m_j \bar{y}_{r_j} + q_j \bar{y}_{nr_j}\right) \Big/ \left(m_j + q_j\right); \quad E\left(\bar{y}_{*nr_j}\right) = \left(\left(n_j - m_j\right) \bar{y}_{nr_j} - q_j \bar{y}_{nr_j}\right) \Big/ \left(n_j - m_j - q_j\right) = \bar{y}_{nr_j}.$$

The expression (4.2) for $E\left(\delta_{*j}\right) = E\left(\bar{y}_{*r_j}\right) - E\left(\bar{y}_{*nr_j}\right)$ follows.

References

Andridge, R.R., and Little, R.J.A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

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

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Households Surveys*. New York: John Wiley & Sons, Inc.

Bianchi, A., Shlomo, N., Schouten, B., da Silva, D. and Skinner, C. (2016). Estimation of response propensities and R-indicators using population-level information. Discussion paper 2016/21, CBS, Voorburg, The Netherlands.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Cochran, W.G. (1977). *Sampling Techniques*. Third edition. New York: John Wiley & Sons, Inc.

Fattorini, L., Franceschi, S. and Maffei, D. (2013). Design-based treatment of unit nonresponse in environmental surveys. *Biometrical Journal*, 55, 925-943.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

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

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M. and Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response; examples from multiple surveys. *Journal of the Royal Statistical Society A*, 173, 389-407.

Little, R.J.A., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 2, 161-168. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf.

Lohr, S.L., Riddles, M.K. and Morganstein, D. (2016). Tests for evaluating nonresponse bias in surveys. *Survey Methodology*, 42, 2, 195-218. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14677-eng.pdf.

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

Nishimura, R., Wagner, J. and Elliott, M. (2016). Alterative indicators for the risk of non-response bias. *International Statistical Review*, 84, 43-62.

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

Peytcheva, E., and Groves, R.M. (2009). Using variation in response rates of demographic subgroups as evidence on nonresponse bias in survey estimates. *Journal of Official Statistics*, 25, 193-201.

Rao, R.S., Glickman, M.E. and Glynn, R.J. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27, 2196-2213.

Särndal, C.-E. (2011a). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.-E. (2011b). Three factors to signal nonresponse bias, with applications to categorical auxiliary variables. *International Statistical Review*, 79, 233-254.

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

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

Särndal, C.-E., and Lundquist, P. (2017). Inconsistent regression and nonresponse bias: Exploring their relationship as a function of response imbalance. *Journal of Official Statistics*, 33(3), 1-27.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24, 251-260.

Särndal, C.-E., and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 2, 131-144. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010002/article/11376-eng.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology,* 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2014). Theoretical and empirical evidence for balancing of survey response by design. Discussion paper 201415, Statistics Netherlands.

Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

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

Wagner, J. (2012). Research synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76, 555-575.

Wagner, J., and Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29, 1014-1024.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 339-347
Statistics Canada, Catalogue No. 12-001-X

An alternative way of estimating a cumulative logistic model with complex survey data

Phillip S. Kott and Peter Frechtel[1]

Abstract

When fitting an ordered categorical variable with $L > 2$ levels to a set of covariates onto complex survey data, it is common to assume that the elements of the population fit a simple cumulative logistic regression model (proportional-odds logistic-regression model). This means the probability that the categorical variable is at or below some level is a binary logistic function of the model covariates. Moreover, except for the intercept, the values of the logistic-regression parameters are the same at each level. The conventional "design-based" method used for fitting the proportional-odds model is based on pseudo-maximum likelihood. We compare estimates computed using pseudo-maximum likelihood with those computed by assuming an alternative design-sensitive robust model-based framework. We show with a simple numerical example how estimates using the two approaches can differ. The alternative approach is easily extended to fit a general cumulative logistic model, in which the parallel-lines assumption can fail. A test of that assumption easily follows.

Key Words: Parallel-lines assumption; Design-sensitive estimation; Standard model; Extended model.

1 Introduction: Fitting a regression model with complex survey data

The goal of this paper is to show an alternative way of estimating a cumulative logistic model (also called the ordinal logistic model or the ordinal regression model), that is, a regression model with a categorical dependent variable having more than two ordered categories, given complex survey data. The standard estimation methods cannot be implemented with most conventional "design-based" software, such as SAS (SAS Institute Inc., 2015), except when the "parallel line assumption" holds as we shall see.

The standard "design-based" framework for fitting a regression model to survey data was introduced by Fuller (1975) for linear regression and by Binder (1983) more generally. This framework treats the finite population as a realization of independent trials from a conceptual population. A maximum likelihood regression estimator could, in principle, be estimated from the finite-population values. The goal in the Fuller/Binder framework is to estimate the conceptual maximum-likelihood estimator, or its limit as the population grows arbitrarily large, from survey data. Skinner (1989) refers to this as the "pseudo-maximum-likelihood" approach.

Kott (2018) describes an alternative model-based approach to estimating regression models with complex survey data dubbed "design sensitive" robust model-based estimation. Following Kott (2007), the *standard model* is defined in this approach in this manner:

$$y_k = f\left(\mathbf{x}_k^T\boldsymbol{\beta}\right) + \varepsilon_k, \text{ where } E\left(\varepsilon_k \mid \mathbf{x}_k\right) = 0. \tag{1.1}$$

Although apparently very general, there is a key restriction imposed by the standard model in equation (1.1): $E\left(\varepsilon_k\right) = 0$ no matter the value of $\mathbf{x}_k$. This assumption can fail and the standard model not be appropriate in the population being analyzed.

1. Phillip S. Kott and Peter Frechtel, RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

In the *extended model*, $E(\varepsilon_k | \mathbf{x}_k) = 0$ in equation (1.1) is replaced by $E(\mathbf{x}_k \varepsilon_k) = \mathbf{0}$. Unlike the standard model, the robust more general extended model rarely fails.

With an independent identically distributed (iid) population U of N elements, it is easy to see that

$$p\lim\left\{N^{-1}\sum\nolimits_U [y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})]\,\mathbf{x}_k\right\} = \mathbf{0}$$

under the extended model. Given a complex sample S with weights $\{w_k\}$, each (nearly) equal to the inverse of the corresponding element's selection probability,

$$p\lim\left\{N^{-1}\sum\nolimits_S w_k\,[y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})]\,\mathbf{x}_k\right\} = \mathbf{0} \tag{1.2}$$

under mild conditions on the sampling design. The parenthetical "nearly" needs to be added when the weights include adjustments for unit nonresponse or coverage errors in the frame which the analyst assumes have been accounted for in an asymptotically unbiased manner. Calibration weight adjustments for statistical efficiency are another reason to add "nearly".

Whether the analyst assumes the standard or the extended model holds in the population, solving for $\mathbf{b}$ in the *weighted estimating equation* (Godambe and Thompson, 1986)

$$\sum\nolimits_S w_k\,[y_k - f(\mathbf{x}_k^T\mathbf{b})]\,\mathbf{x}_k = \mathbf{0} \tag{1.3}$$

provides a consistent estimator for $\boldsymbol{\beta}$ under mild conditions.

The pseudo-maximum-likelihood estimating equation in Binder is

$$\sum\nolimits_S w_k \frac{f'(\mathbf{x}_k^T\mathbf{b})}{v_k}[y_k - f(\mathbf{x}_k^T\mathbf{b})]\,\mathbf{x}_k = \mathbf{0},$$

where $v_k = E(\varepsilon_k^2 | \mathbf{x}_k)$. For logistic, Poisson, and ordinary least squares (OLS) linear regression, $f'(\mathbf{x}_k^T\boldsymbol{\beta})/v_k = 1$. This equality may not hold for general least squares (GLS) linear regression, however even when the elements are uncorrelated. It also need not hold for a cumulative logistic regression model.

The cumulative logistic model is a multinomial logistic regression model for L categories with a natural ordering (e.g., always, frequently, sometimes, never). Being in the first category is assumed to fit a logistic model. Being in either the first or second category is assumed to fit a logistic model. Being in the first, second, or third category is assumed to fit a logistic model, and so forth.

The *general cumulative logistic model* is (splitting out the intercept from the rest of the covariates)

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k^T\boldsymbol{\beta}_\ell)}{1 + \exp(\alpha_\ell + \mathbf{x}_k^T\boldsymbol{\beta}_\ell)} \text{ for } \ell = 1, \ldots, L-1,$$

where $y_{\ell k} = 1$ when k is in one of the first ℓ categories, 0 otherwise. The *parallel-lines assumption* is that $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ for all integer values of ℓ less than L with each such value having its own intercept (α_ℓ). The

cumulative logistic model under the parallel-lines assumption is often called a *proportional-odds model*. We will call it the "simple cumulative logistic model," although it is more commonly referred to as *the* cumulative logistic model (or *the* ordinal logistic model).

Finding the a_ℓ and $\mathbf{b}_\ell$ that satisfy the estimating equation:

$$\sum_{k\in S} w_k \left[y_{\ell k} - \frac{\exp\left(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell\right)}{1+\exp\left(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell\right)} \right] \begin{bmatrix} 1 \\ \mathbf{x}_k \end{bmatrix} = \mathbf{0} \text{ for } \ell = 1, \ldots, L-1 \tag{1.4}$$

can be used for estimating the general cumulative logistic model. This is *not* the pseudo-maximum-likelihood estimating equation in the *surveylogistic* routine in SAS/STAT 14.1 (An (2002, page 7) discusses the multivariate pseudo-maximum-likelihood estimating equation fit by this procedure), the *logistic* routine in SUDAAN 11 (Research Triangle Institute, 2012) or the *gologit2* routine in STATA (Williams, 2005) for the simple cumulative logistic model. Only the STATA routine allows the $\mathbf{b}_\ell$ to vary.

Given L *nominal* categories and complex survey data, SAS and SUDAAN *can* fit the *general multinomial logistic model*,

$$E\left(y_{\ell k} \mid \mathbf{x}_k\right) = \frac{\exp\left(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell\right)}{1+\sum_{j=1}^{L-1} \exp\left(\alpha_j + \mathbf{x}_k^T \boldsymbol{\beta}_j\right)} \text{ for } \ell = 1, \ldots, L-1,$$

with $y_{\ell k} = 1$ when k is in the ℓ^{th} category, 0 otherwise; this is not the same thing as the general *cumulative* logistic model, which these programs cannot estimate with complex survey data.

In what follows, we introduce a modest example of a simple cumulative logistic model. Given complex survey data, we fit the model both with the pseudo-maximum-likelihood technique and with equation (1.4). The latter is accomplished by creating a data set with $L-1$ observations for each respondent k (note that $y_{1k}, \ldots, y_{L-1k}$ are in the same primary sampling unit). We follow Kott (2018) and call this fitting method the "design-sensitive" technique, even though, strictly speaking, it is model based. Moreover, the pseudo-maximum-likelihood approach is also sensitive to the design weights and other aspects of the sampling design.

The article goes on to test the parallel-lines assumption. A simple example is presented in Section 2. Section 3 concludes with a discussion.

2 A simple example

The National Survey on Drug Use and Health (NSDUH) is an annual survey of the civilian, noninstitutionalized population aged 12 or older living in the United States. Using NSDUH data from 2006 to 2010, we focus on a survey question given to adolescents (12-17) who received depression treatment in the past year:

During the past 12 months, how much has treatment or counseling helped you?

The viable responses were: Not at all (l); A little (2); Some (3); A lot (4); or Extremely (5).

We discarded missing and invalid responses both to this question and to the question of whether the respondent received depression treatment in the past year. We will return to this practice in the discussion section.

Using SAS, we estimated the following simple cumulative logistic model:

$$E(y_{\ell k} \mid \mathbf{x}_k) = \frac{\exp(\alpha_\ell + meds_k \beta)}{1 + \exp(\alpha_\ell + meds_k \beta)} \text{ for } \ell = 1, \ldots, L-1, \tag{2.1}$$

where $meds = 1$ when respondent k was taking medication for depression (0 otherwise), with both pseudo-maximum-likelihood and the design-sensitive technique. For pseudo-maximum-likelihood estimation, we reversed the order of the responses with $y_{1k} = 1$ when k responded that treatment (or counseling) helped extremely, $y_{2k} = 1$ when k responded that treatment helped extremely or a lot, $y_{3k} = 1$ when k responded that treatment helped more than a little, and $y_{4k} = 1$ when k responded that treatment helped at least a little. Finally, $y_{5k} = 1 - y_{4k} = 1$ when k responded that treatment did not help at all. In SAS, this meant dependent variable Y was set equal to 1 when treatment helped extremely, to 2 when treatment helped a lot, …, and to 5 when treatment didn't help at all.

For the design-sensitive technique, we created four observations from k in a new data set. In the i^{th} observation labeled $\text{C} = i$ in SAS, a class (categorical) variable added to the model statement, we created a dependent variable (D) equal to y_{ik} in equation (2.1). We needed to add EVENT = "1" after D in the model statement because we were modeling when D = 1.

SAS code for both estimation techniques are in the appendix. The NSDUH data set we used had 60 variance strata with two variance primary sampling units (PSUs) in each and analysis weights based on the probabilities of selection and unit response.

The parameter estimates from our pseudo-maximum-likelihood and design-sensitive SAS runs are displayed in Tables 2.1 and 2.2, respectively. In Table 2.1, $Intercept = i$ is the pseudo-maximum-likelihood estimate of α_{ik} in equation (2.1). The sum of the *Intercept* and $\text{C} = i$ in Table 2.2 is the design-sensitive estimate for α_{ik} when $i = 1, 2,$ or 3, while the design-sensitive estimate for α_{4k} is the Intercept in Table 2.2 minus the sum: [C = 1] + [C = 2] + [C = 3]. Finally (and more simply), *meds* in both tables estimates β.

In all cases, estimates of the same parameter from the two tables are close. The percent increase in every level of satisfaction with treatment due to having taken drugs for depression (the estimate for β) is roughly 45% (in our discussion of the results of the logistic regressions, we treat differences of the log odds as equal to percent differences in the odds, even though this is only approximately true). That near equality suggests that the parallel-lines assumption is not violated by our NSDUH data.

Table 2.1
Pseudo-maximum-likelihood estimates for the simple cumulative logistic model

Parameter	Estimate	Standard Error	t Value	Pr > | t |
Intercept 1	-2.2917	0.0913	-25.10	< 0.0001
Intercept 2	-0.7617	0.0685	-11.11	< 0.0001
Intercept 3	0.2511	0.0624	4.02	0.0002
Intercept 4	1.3695	0.0739	18.53	< 0.0001
meds	0.4516	0.0965	4.68	< 0.0001

NOTE: The degrees of freedom for the *t* tests is 60.

Table 2.2
Design-sensitive estimates for the simple cumulative logistic model

Parameter	Estimate	Standard Error	t Value	Pr > | t |
Intercept	-0.3591	0.0583	-6.16	< 0.0001
C 1	-1.9329	0.0592	-32.63	< 0.0001
C 2	-0.4039	0.0356	-11.33	< 0.0001
C 3	0.6087	0.0392	15.52	< 0.0001
meds	0.4498	0.0955	4.71	< 0.0001

NOTE: The degrees of freedom for the *t* tests is 60.

The parallel-lines assumption can be tested directly by adding a class variable M to the design-sensitive data set with

$$\mathrm{M}=1 \text{ when } \mathrm{C}=1 \text{ and } meds = 1,$$
$$\mathrm{M}=2 \text{ when } \mathrm{C}=2 \text{ and } meds = 1,$$
$$\mathrm{M}=3 \text{ when } \mathrm{C}=3 \text{ and } meds = 1, \text{ and}$$
$$\mathrm{M}=4 \text{ otherwise.}$$

When added to the model statement in SAS, the class variable M captures the differing impacts of taking medication for depression in the previous year on the levels of satisfaction with treatment. For example, the estimated percent increase in the odds of being extremely pleased by treatment due to having taken drugs for depression during the year is, according to Table 2.3, 0.3816 (from $meds$) plus 0.0717 (from M = 1) or 45.33%. The other percent increases are lower, but none are significantly different from the others. We see that from the extremely low F value for M in Table 2.4. In addition, none of the t-values for an M in Table 2.3 is significant even at the 0.5 level (10 times larger than the standard 0.05 level).

Table 2.3
Estimating the general cumulative logistic model

Parameter	Estimate	Standard Error	t Value	Pr > | t |
Intercept	-0.2919	0.1270	-2.30	0.0251
C 1	-1.9636	0.0806	-24.37	< 0.0001
C 2	-0.4104	0.0440	-9.33	< 0.0001
C 3	0.6202	0.0490	12.66	< 0.0001
Meds	0.3816	0.1452	2.63	0.0109
M 1	0.0717	0.1273	0.56	0.5754
M 2	0.0234	0.0652	0.36	0.7215
M 3	-0.0236	0.0719	-0.33	0.7439

NOTE: The degrees of freedom for the *t* tests is 60.

Table 2.4
F tests for the general cumulative logistic model

Effect	F Value	Num DF	Den DF	Pr > F
C	280.39	3	58	< 0.0001
Meds	6.91	1	60	0.0109
M	0.16	3	58	0.9239

3 Discussion

When there is more than one explanatory variable in the cumulative logistic model then each one needs to be tested like *meds* was in the previous section by adding an analogous class variable for each. A general F test can be used for testing whether every class variable is not significant (say at the 0.05 level). A better approach with complex survey data may be to follow Korn and Graubard (1990) and use the simple Bonferroni-adjusted t-test. For significance at the 0.05 level, one would compute the t-values for every tested component of each added class variable (there are three such in Table 2.3), then compare the p-value of the smallest of these to 0.05/the number of components tested.

An advantage of the design-sensitive model-based approach to fitting a simple cumulative logistic model over the pseudo-maximum-likelihood approach is not apparent with our NSDUH data. When the parallel-lines assumption doesn't hold, and an extended model is being fit, satisfying the first "equation" in (1.4) assures us that

$$\sum_{k\in S} w_k y_{\ell k} = \sum_{k\in S} w_k \frac{\exp\left(a_\ell + \mathbf{x}_k\mathbf{b}\right)}{1+\sum_{j=1}^{L-1}\exp\left(a_j + \mathbf{x}_k\mathbf{b}\right)} \text{ for } \ell = 1, \ldots, L-1. \tag{3.1}$$

When $\mathbf{x}_k$ itself is a multi-level categorical variable (so that one and only one component of $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kQ}\right)$ is 1 while the other components are 0), equation (3.1) assures that the weighted mean of $y_{\ell k}$ for each $\mathbf{x}$-category (i.e., component of $\mathbf{x}_k$) and cumulative level ℓ equals its predicted value described by

$$\hat{y}_{\ell k} = \exp\left(a_\ell + \mathbf{x}_k^T\mathbf{b}\right) \Big/ \left[1+\sum_{j=1}^{L-1}\exp\left(a_j + \mathbf{x}_k^T\mathbf{b}\right)\right],$$

which is a reasonable property. Equation (1.4) is simply an extension of the property to more general $\mathbf{x}_k$.

In our NSDUH example, although not generally, using the design-sensitive approach was slightly more efficient than using the pseudo-maximum-likelihood approach. This can be seen by comparing the t-values of *meds* (the inverses of their respective estimated coefficients of variation) in Tables 2.1 and 2.2. When we ignore the analysis weights, the strata, and the clustering (by setting the weights and strata to 1, and treating each respondent as a primary sampling unit), this result reverses as expected. The point here is that pseudo-maximum likelihood with complex survey data is indeed "pseudo" (in this case that is likely because of the impact of the weights on the estimates).

Finally, the data set we created dropped responding observations with missing values of the dependent and *meds* variables. When fitting the *extended* model, this is only valid (i.e., resulting estimates are asymptotically unbiased) when an in-scope respondent – an adolescent who had treatment for depression in the previous year – being dropped occurred completely at random. When fitting the *standard* model, the probability of being dropped can be a function only of whether an in-scope adolescent has taken medication for depression in the previous year but nothing else. This suggests it may have been prudent to add variables to the model that are never missing even when they are not significant. If we add class variables for age, sex, race/ethnicity, urbanicity, and family income (all of which have values imputed for them when missing in the NSDUH) to our simple cumulative logistic model, none are significant at the 0.05 level. The major results do not change meaningfully (the estimate for β increases from roughly 0.45 to 0.50), although that the t-value for meds using the design-sensitive approach $(b_{meds} = 0.4948;\ t_{meds} = 5.49)$ is slightly smaller than that from using the pseudo-maximum-likelihood approach $(b_{meds} = 0.4987;\ t_{meds} = 5.52)$.

Appendix

/* PML is a data set of adolescents NSDUH respondents in the 2006 to 2010 survey years who reported having treatment for depression and whether they had taken drugs for depression. Variables include:

Y = 1 treatment was extremely helpful; Y = 2 treatment helped a lot; Y = 3 some; Y = 4 a little;
Y = 5 not at all
meds = 1 had taken drugs for depression, 0 otherwise
VESTR variance stratum
VEPSU variance primary sampling unit
IDNUM respondent identification number
ANALWT the analysis weight

This set is employed for pseudo-maximum-likelihood estimation of the simple cumulative logistic model and to create the DS_SIMPLE data set, which is used for design-sensitive estimation of the simple cumulative logistic model, and it is employed to create DS_GENERAL data set, which is used for design-sensitive estimation of the general cumulative logistic model. */

```
DATA DS_SIMPLE; SET PML; BY VESTR VEPSU IDNUM;
D = 0;
C = 1; IF Y < 2 THEN D = 1; OUTPUT;
C = 2; IF Y < 3 THEN D = 1; OUTPUT;
C = 3; IF Y < 4 THEN D = 1; OUTPUT;
C = 4; IF Y < 5 THEN D = 1; OUTPUT;

DATA DS_GENERAL; SET DS_SIMPLE;
M = 4;
IF C = 1 AND MEDS = 1 THEN M = 1;
IF C = 2 AND MEDS = 1 THEN M = 2;
IF C = 3 AND MEDS = 1 THEN M = 3;
```

```
/*The PROC below is used to produce Table 2.1*/

PROC SURVEYLOGISTIC DATA = PML; CLUSTER VEPSU;
MODEL Y = MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

/*The PROC below is used to produce Table 2.2*/

PROC SURVEYLOGISTIC DATA = DS_SIMPLE; CLASS C;
CLUSTER VEPSU;
MODEL D(EVENT = '1') = C MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

/*The PROC below is used to produce Tables 2.3 and 2.4*/

PROC SURVEYLOGISTIC DATA =DS_GENERAL; CLASS M C;
CLUSTER VEPSU;
MODEL D(EVENT = '1') = C MEDS M;
STRATA VESTR; WEIGHT ANALWT; RUN;
```

References

An, A. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. In *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*, Cary, NC: SAS Institute Inc. (http://www2.sas.com/proceedings/sugi27/p258-27.pdf).

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā-The Indian Journal of Statistics, Series C*, 37, 117-132.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54(2). 127-138.

Korn, E L., and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *American Statistician*, 44, 270-276.

Kott, P.S. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*, 1, 11-18.

Kott, P.S. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12, 1-17.

Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds., C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: John Wiley & Sons, Inc., 59-87.

Williams, R. (2005). Gologit2: A Program for Generalized Logistic Regression/Partial Proportional Odds Models for Ordinal Variables. Retrieved January 3, 2016 (http://www.nd.edu/~rwilliam/stata/gologit2.pdf).

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 349-364
Statistics Canada, Catalogue No. 12-001-X

On combining independent probability samples

Anton Grafström, Magnus Ekström, Bengt Gunnar Jonsson, Per-Anders Esseen and Göran Ståhl[1]

Abstract

Merging available sources of information is becoming increasingly important for improving estimates of population characteristics in a variety of fields. In presence of several independent probability samples from a finite population we investigate options for a combined estimator of the population total, based on either a linear combination of the separate estimators or on the combined sample approach. A linear combination estimator based on estimated variances can be biased as the separate estimators of the population total can be highly correlated to their respective variance estimators. We illustrate the possibility to use the combined sample to estimate the variances of the separate estimators, which results in general pooled variance estimators. These pooled variance estimators use all available information and have potential to significantly reduce bias of a linear combination of separate estimators.

Key Words: Horvitz-Thompson estimator; Inclusion probabilities; Linear combination estimator; Variance estimation.

1 Introduction

The idea of using all available information to produce better estimates is very appealing, but it is seldom clear how to proceed to achieve the best results. There is a vast literature on what has become known as meta-analysis, that builds on the idea of combining results of multiple studies. Cochran and Carroll (1953) and Cochran (1954) are two early papers that treat combination of estimates from different experiments. Koricheva, Gurevitch and Mengersen (2013) and Schmidt and Hunter (2014) are two books that provide an updated and more comprehensive treatment of meta-analysis. In this paper we do not treat combination of results from traditional experiments, but rather from multiple probability samples. We present all required design elements, such as inclusion probabilities of first and second order, for a general combination of multiple independent samples from different sampling designs. We also present new estimators for the variance of separate estimators based on the design of the combined samples. These suggested variance estimators can be thought of as general pooled variance estimators using all available information. In particular such pooled variance estimators can be used in a linear combination of separate estimators to reduce the mean square error (MSE) compared to using the separate, and thus independent, variance estimators.

A restriction is that we only treat combination of independent probability samples selected from the same population at the same point in time, or under the assumption that there has been a non-significant change in the target variable. Further, we assume that each sampling design is known to the extent that inclusion probabilities of first and second order are known for all units. In general we will also need to be able to

1. Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden. E-mail: anton.grafstrom@slu.se; Magnus Ekström, Department of Statistics, USBE, Umeå University, SE-90187 Umeå, Sweden, and Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden. E-mail: magnus.ekstrom@umu.se; Bengt Gunnar Jonsson, Department of Natural Sciences, Mid Sweden University, SE-85170 Sundsvall, Sweden. E-mail: bengt-gunnar.jonsson@miun.se; Per-Anders Esseen, Department of Ecology and Environmental Science, Umeå University, SE-90187 Umeå, Sweden. E-mail: per-anders.esseen@umu.se; Göran Ståhl, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden. E-mail: goran.stahl@slu.se.

uniquely identify each unit so that we can detect if the same unit is selected in more than one sample, or multiple times in the same sample. At least some of these assumptions may be quite restrictive as they may not hold in some practical circumstances.

Let $U = \{1, 2, \ldots, N\}$ be the set of labels of the N units in the population. Our objective is to estimate the total of a target variable y, that takes value y_i for unit $i \in U$. Thus we wish to estimate $Y = \sum_{i=1}^{N} y_i$. We assume access to k independent probability samples $S^{(\ell)}$, $\ell = 1, \ldots, k$, from U, where the samples may be from different sampling designs. Under these assumptions, we investigate different options for estimating the population total by use of all available information. Knowledge of what units have been included in multiple different samples is required in some cases. Such knowledge is more readily available today in environmental monitoring and natural resource surveys, following the widespread use of accurate satellite-based positioning systems (Næsset and Gjevestad, 2008). In environmental studies the units can often be considered as locations with given coordinates, so the situation is different from surveys of e.g., people that may be anonymous or unidentifiable. Further, in several countries landscape and forest monitoring programmes are performed (Tomppo, Gschwantner, Lawrence and McRoberts, 2009; Ståhl, Allard, Esseen, Glimskår, Ringvall, Svensson, Sundquist, Christensen, Gallegos Torell, Högström, Lagerqvist, Marklund, Nilsson and Inghe, 2011; Fridman, Holm, Nilsson, Nilsson, Ringvall and Ståhl, 2014) which sometimes need to be augmented by special sampling programmes in order to reach specific accuracy targets for certain regions or years (Christensen and Ringvall, 2013).

In Section 2 we first recall the theory for an optimal linear combination of separate independent estimators. Then, in Section 3, we present the theory for combining independent samples. As a unit may be included in more than one sample or multiple times in the same sample we need to choose between using single or multiple count of inclusion. By using single count the resulting design becomes a without replacement design and multiple count results in a form of with replacement design. Two examples comparing different alternatives for estimation are presented in Section 4. We end with a discussion in Section 5.

2 Combining separate estimates

We assume that we have $k \geq 2$ estimators, $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$ of a population total Y, resulting from k independent samples from the same population. Our options greatly depend on what information is available. If we have estimates and corresponding variance estimates, then a linear combination based on weights calculated from estimated variances may be an interesting option. We could also weight the estimators with respect to sample size, if available, but that is known to be far from optimal in some situations. We recall the theory for an optimal linear combination of independent unbiased estimators. The linear combination of $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$ with the smallest variance is

$$\hat{Y}_L = \alpha_1 \hat{Y}_1 + \alpha_2 \hat{Y}_2 + \ldots + \alpha_k \hat{Y}_k,$$

where

$$\alpha_i = \frac{1/V_i\left(\hat{Y}_i\right)}{\sum_{j=1}^{k} 1/V_j\left(\hat{Y}_j\right)}$$

are positive weights that sum to 1. The variance of $\hat{Y}_L$ is

$$V\left(\hat{Y}_L\right) = \frac{1}{\sum_{j=1}^{k} 1/V_j\left(\hat{Y}_j\right)}.$$

It is common that variance estimates are used in place of the unknown variances when calculating the α-weights, see Cochran and Carroll (1953) and Cochran (1954). If the variance estimators are consistent, that approach will asymptotically provide the optimal weighting. Moreover, under the assumption that the variance estimators are independent of the estimators $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$, the resulting estimator

$$\hat{Y}_L^* = \hat{\alpha}_1\hat{Y}_1 + \hat{\alpha}_2\hat{Y}_2 + ... + \hat{\alpha}_k\hat{Y}_k,$$

is unbiased and its variance depends only on the variance of $\hat{Y}_L$ and the MSEs of the $\hat{\alpha}_i$'s, see Rubin and Weisberg (1974). However, as we soon will illustrate, the assumption of independence is likely to be violated in many sampling applications. In case of positive correlations between the estimators and their variance estimators, we will on average put more weight on small estimates because they tend to have smaller estimated variances. Thus the combined estimator (using weights based on estimated variances) will be negatively biased and the negative bias can increase as the number of independent surveys we combine increases, see Example 1. The opposite holds as well, in case of negative correlation, but that is likely a rarer situation in sampling applications.

Example 1: *A very simplistic example that illustrates that the bias can increase as the number of independent surveys we combine increase. Let the unbiased estimator* $\hat{Y}$ *for one sample take the values 1 or 2 with equal probabilities and let the variance estimator take values* c *times the estimator* (*perfectly correlated*) *and let it be unbiased* $(c = 1/6)$. *Clearly the expected value of* $\hat{Y}$ *is 1.5. Next, we consider the linear combination of two independent estimators* $\left(\hat{Y}_1, \hat{Y}_2\right)$ *of the same type as* $\hat{Y}$ *using estimated variances. The pair* $\left(\hat{Y}_1, \hat{Y}_2\right)$ *has the following four possible outcomes* (*1,1*), (*1,2*), (*2,1*), (*2,2*), *each with probability 1/4. The corresponding outcomes for the linear combination* $\hat{Y}_L^*$ *with estimated variances are 1, 4/3, 4/3, 2 with expectation* $17/12 \approx 1.4167$. *It is negatively biased. If a third independent estimator of the same type is added we have the eight outcomes* (*1,1,1*), (*1,1,2*), (*1,2,1*), (*1,2,2*), (*2,1,1*), (*2,1,2*), (*2,2,1*), (*2,2,2*), *each with equal probability 1/8. The corresponding outcomes for* $\hat{Y}_L^*$ *are 1, 6/5, 6/5, 3/2, 6/5, 3/2, 3/2, 2, with expectation* $111/80 = 1.3875$. *It is even more negatively biased, and the bias continues to grow as more independent estimators of the same type are added in the combination.*

2.1 Why positive correlation between estimator and variance estimator is common in sampling applications

The issue of positive correlation between the estimator of a total and its variance estimator has previously been noticed by e.g., Gregoire and Schabenberger (1999) when sampling skewed biological populations, but we show that a high correlation may appear in more general sampling applications. Assume that the

target variable is non-negative and that $y_i > 0$ for exactly N' units. The proportion of non-zero (positive) y_i's is denoted by $p = N'/N$. This is a very common situation in sampling and we get such a target variable if we estimate a domain total ($y_i = 0$ outside of the domain) or if only a subset of the population has the property of interest.

The design-based unbiased Horvitz-Thompson (HT) estimator is given by

$$\hat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i},$$

where S denotes the random set of sampled units and $\pi_i = \Pr(i \in S)$. Under fixed size designs the variance of $\hat{Y}$ is

$$V(\hat{Y}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where $\pi_{ij} = \Pr(i \in S, j \in S)$ is the second order inclusion probability. The corresponding variance estimator is

$$\hat{V}(\hat{Y}) = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$

Provided that all π_{ij} are strictly positive, it follows that the variance estimator is an unbiased estimator of $V(\hat{Y})$.

The number of non-zero y_i's in S (and hence in $\hat{Y}$) is here denoted by n' and it will usually be a random number. It can be shown that the number of non-zero elements in $\hat{V}(\hat{Y})$ is approximately proportional to n' if p is small, which indicates that there might be a strong correlation between $\hat{Y}$ and $\hat{V}(\hat{Y})$ in general if p is small. To show that the number of non-zero terms in $\hat{V}(\hat{Y})$ is approximately proportional to n' we look at three cases, where the third case is the most general.

Case 1: Assume that all the non-zero y_i/π_i's are different, i.e., $y_i/\pi_i \neq y_j/\pi_j$ for $i \neq j$, and $\pi_{ij} \neq \pi_i \pi_j$ for all i, j. The double sum in $\hat{V}(\hat{Y})$ then contains $2n'(n - n')$ non-zero terms of the form

$$\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_k}{\pi_k} \right)^2,$$

where k is equal to i or j and $i \neq j$. There are $n'(n' - 1)$ non-zero terms of the form

$$\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where $i \neq j$. In total the number of non-zero terms is $n'(2n - n' - 1)$. If n is fairly large and p is small, then $n' << n$ and roughly we have $n'(2n - n' - 1) \approx 2n'n$. The number of non-zero terms is approximately proportional to n'.

Case 2: Assume that all the non-zero y_i/π_i's are equal, e.g., y_i is an indicator variable and $\pi_i = n/N$, and $\pi_{ij} \neq \pi_i\pi_j$ for all i, j. Then the double sum in $\hat{V}\left(\hat{Y}\right)$ contain $2n'(n-n')$ non-zero terms of the form

$$\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\left(\frac{y_k}{\pi_k}\right)^2,$$

where k is equal to i or j and $i \neq j$. If n is fairly large and p is small, then $n' << n$ and roughly we have $2n'(n-n') \approx 2n'n$. Thus, the number of non-zero terms is still approximately proportional to n'.

Case 3: If some of the non-zero y_i/π_i's are equal and the rest are different, then the number of non-zero terms will be between $2n'(n-n')$ (case 2) and $n'(2n-n'-1)$ (case 1). Thus, the number of non-zero terms in $\hat{V}\left(\hat{Y}\right)$ is always approximately proportional to n' if p is small.

If $\pi_{ij} \leq \pi_i\pi_j$ for all $i \neq j$, then all non-zero terms are positive. This condition holds e.g., for simple random sampling (SRS) and high entropy unequal probability designs such as Conditional Poisson, Sampford and Pareto. More discussion about entropy of sampling designs can be found in e.g., Grafström (2010). The average size of the positive terms in $\hat{V}\left(\hat{Y}\right)$, or $\hat{Y}$, is not likely to depend much on n'. Thus, if $\hat{Y}$ contains n' positive terms, and $\hat{V}\left(\hat{Y}\right)$ contains a number of positive terms that is proportional to n', their sizes are mainly determined by n'. A high relative variance in n' can cause a high correlation between $\hat{Y}$ and $\hat{V}\left(\hat{Y}\right)$, see Example 2.

Commonly used designs can produce a high relative variance for n'. If we do simple random sampling without replacement we get $n' \sim \text{Hyp}(N, N', n)$ and $V(n')/E(n') = (1-p)(N-n)/(N-1) \approx (1-p)(1-n/N)$, which means that we need a large p or a large sample fraction n/N in order to achieve a small relative variance for n'. In many applications we will have a rather small p and a small sampling fraction n/N and, thus, for many designs (that do not use prior information which can explain to some extent if $y_i \neq 0$ or not) there will be a high relative variance for n'. To illustrate the magnitude of the resulting correlation between the estimator and its variance estimator an example for simple random sampling without replacement follows.

Example 2: *For this example we first simulate a population of size* $N = 1{,}000$ *where* $N' = 100$, *i.e.,* $p = 0.1$. *The 100 non-zero* y*-values are simulated from* $N(\mu, \sigma^2)$ *with* $\mu = 10$ *and* $\sigma = 2$. *We select samples of size* $n = 200$ *with simple random sampling, so* $\pi_i = n/N$ *and* $\pi_{ij} = n(n-1)/(N(N-1))$ *for* $i \neq j$. *The observed correlation between* $\hat{Y}$ *and* $\hat{V}\left(\hat{Y}\right)$ *was 0.974 for* 10^6 *samples, see Figure 2.1 for the first 1,000 observations of* $\left(\hat{Y}, \hat{V}\left(\hat{Y}\right)\right)$. *If we increase* p *to 0.3, the correlation is still above 0.9. The results remain unchanged if the ratio* σ/μ *remains unchanged, e.g., we get the same correlations if* $\mu = 100$ *and* $\sigma = 20$.

Now, assume we have access to more than one sample for the estimation of Y. As previously noted, with high positive correlations between the estimators and their corresponding variance estimators there is a risk of severe bias if we use a linear combination with estimated variances. The interest of using combined

information may be the largest for small domains or rare properties, in which case the problem of high correlation is the most likely. Next, we turn to alternative options for using combined information from multiple samples.

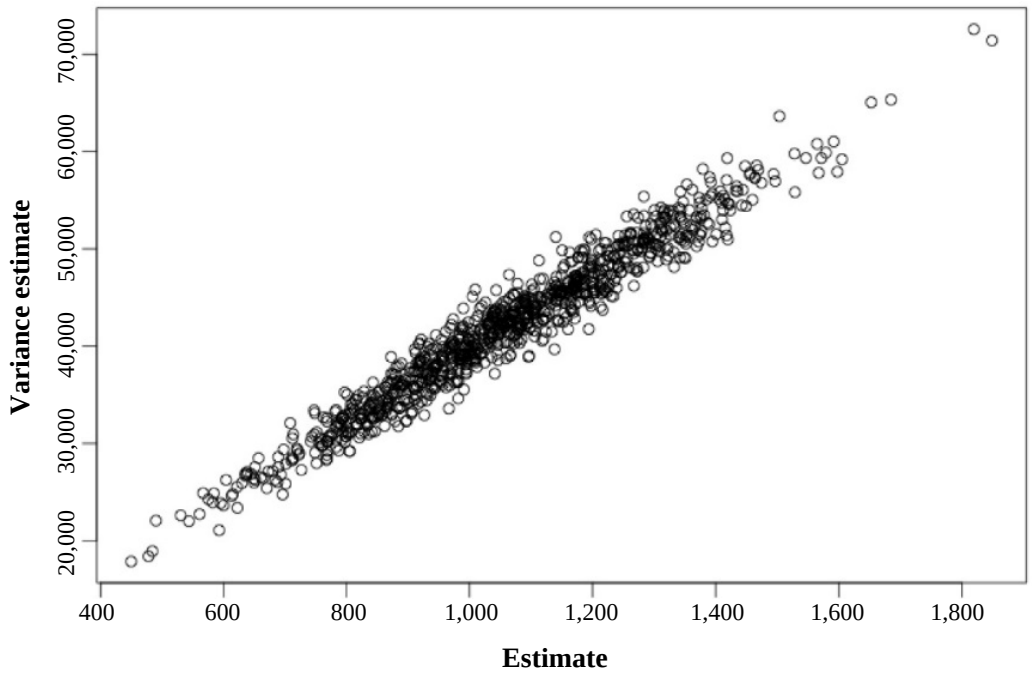


Figure 2.1 Relationship between Horvitz-Thompson estimator and its variance estimator for a variable with 90% zeros.

3 Combining samples

Here we derive the design elements (e.g., inclusion probabilities of first and second order) for the combined sample. There are however different options to combine samples. We must e.g., choose between multiple or single count for the combined design. When combining independent samples selected from the same population we need to know the inclusion probabilities of all units in the samples, for all designs. Second order inclusion probabilities are needed for variance estimation. In some cases we also need to have unique identifiers (labels) for the units so they can be matched, e.g., when we use single count or when at least one separate design has unequal probabilities. Bankier (1986) considered the single count approach for the special case of combining two independently selected stratified simple random samples from the same frame. Roberts and Binder (2009) and O'Muircheartaigh and Pedlow (2002) discussed different options for combining independent samples from the same frame, but not with general sampling designs.

A somewhat similar problem is estimation based on samples from multiple overlapping frames, see e.g., the review articles by Lohr (2009, 2011) and the referenced articles therein. Even though having the same

frame can be considered as a special case of multiple frames, we have not found derivations of the design elements (in particular second order inclusion probabilities and second order of expected number of inclusions) for the combination of general sampling designs. Below we present, for general probability sampling designs, in detail two main ways to combine probability samples and derive corresponding design features needed for unbiased estimation and unbiased variance estimation.

3.1 Combining with single count

Here we first combine two independent samples $S^{(1)}$ and $S^{(2)}$ selected from the same population, and look at the union of the two samples as our combined sample. Thus, the inclusion of a unit is only counted once even if it is included in more than one sample. The first order inclusion probabilities are

$$\pi_i^{(1,\,2)} = \pi_i^{(1)} + \pi_i^{(2)} - \pi_i^{(1)}\pi_i^{(2)}, \tag{3.1}$$

where $\pi_i^{(1,\,2)} = \Pr\left(i \in S^{(1)} \cup S^{(2)}\right)$ and $\pi_i^{(\ell)} = \Pr\left(i \in S^{(\ell)}\right)$ for $\ell = 1, 2$. We let $I_i^{(1)}$, $I_i^{(2)}$ and $I_i^{(1,\,2)}$ be the inclusion indicator for unit i in $S^{(1)}$, $S^{(2)}$ and $S^{(1)} \cup S^{(2)}$ respectively. The resulting design is no longer a fixed size design (even if the separate designs are of fixed size). The expected size of the union $S^{(1)} \cup S^{(2)}$ is given by $E\left(n^{(1,\,2)}\right) = \sum_{i=1}^{N} \pi_i^{(1,\,2)}$, where $n^{(1,\,2)} = \sum_{i=1}^{N} I_i^{(1,\,2)}$ denotes the random size of the union. If we are interested in how much the samples will overlap on average, the expected size of the overlap is given by the sum $\sum_{i=1}^{N} \pi_i^{(1)}\pi_i^{(2)}$.

The second order inclusion probabilities $\pi_{ij}^{(1,\,2)}$ for the union $S^{(1)} \cup S^{(2)}$ can be written in terms of first and second order inclusion probabilities of the two respective designs. Let $B = (i \in S^{(1)} \cup S^{(2)},\ j \in S^{(1)} \cup S^{(2)})$, then $\pi_{ij}^{(1,\,2)} = \Pr(B)$. By conditioning on the outcomes for i and j in $S^{(1)}$ we get the following four cases

m	A_m	$\Pr(A_m)$	$\Pr(B \mid A_m)$
1	$i \in S^{(1)},\ j \in S^{(1)}$	$\pi_{ij}^{(1)}$	1
2	$i \in S^{(1)},\ j \notin S^{(1)}$	$\pi_i^{(1)} - \pi_{ij}^{(1)}$	$\pi_j^{(2)}$
3	$i \notin S^{(1)},\ j \in S^{(1)}$	$\pi_j^{(1)} - \pi_{ij}^{(1)}$	$\pi_i^{(2)}$
4	$i \notin S^{(1)},\ j \notin S^{(1)}$	$1 - \pi_i^{(1)} - \pi_j^{(1)} + \pi_{ij}^{(1)}$	$\pi_{ij}^{(2)}$

where $\pi_{ij}^{(\ell)} = \Pr\left(i \in S^{(\ell)},\ j \in S^{(\ell)}\right)$ for $\ell = 1, 2$. The events A_m, $m = 1, 2, 3, 4$, are disjoint and $\sum_{m=1}^{4} \Pr(A_m) = 1$. Thus, by the law of total probability, we have $\pi_{ij}^{(1,\,2)} = \Pr(B) = \sum_{m=1}^{4} \Pr(B \mid A_m) \Pr(A_m)$. This gives us

$$\pi_{ij}^{(1,\,2)} = \pi_{ij}^{(1)} + \pi_j^{(2)}\left(\pi_i^{(1)} - \pi_{ij}^{(1)}\right) + \pi_i^{(2)}\left(\pi_j^{(1)} - \pi_{ij}^{(1)}\right) + \pi_{ij}^{(2)}\left(1 - \pi_i^{(1)} - \pi_j^{(1)} + \pi_{ij}^{(1)}\right). \tag{3.2}$$

The equations (3.1) and (3.2) can be generalized to recursively obtain first and second order inclusion probabilities of the union of an arbitrary number k of independent samples. After having derived probabilities for the union of the first two samples, we can combine the result with the probabilities of the third design using the same formulas and so on. To exemplify, let $\pi_i^{(1,\ldots,\ell)}$ be the first order inclusion probability of unit i in the union of the first ℓ samples. Then we have

$$\pi_i^{(1,\ldots,\ell+1)} = \pi_i^{(1,\ldots,\ell)} + \pi_i^{(\ell+1)} - \pi_i^{(1,\ldots,\ell)}\pi_i^{(\ell+1)},$$

as the first order inclusion probability of unit i in the union of the first $\ell+1$ samples. Similarly, for the second order inclusion probabilities we get the recursive formula

$$\begin{aligned}\pi_{ij}^{(1,\ldots,\ell+1)} &= \pi_{ij}^{(1,\ldots,\ell)} + \pi_j^{(\ell+1)}\left(\pi_i^{(1,\ldots,\ell)} - \pi_{ij}^{(1,\ldots,\ell)}\right) + \pi_i^{(\ell+1)}\left(\pi_j^{(1,\ldots,\ell)} - \pi_{ij}^{(1,\ldots,\ell)}\right) \\ &\quad + \pi_{ij}^{(\ell+1)}\left(1 - \pi_i^{(1,\ldots,\ell)} - \pi_j^{(1,\ldots,\ell)} + \pi_{ij}^{(1,\ldots,\ell)}\right).\end{aligned}$$

Henceforth, for the combination of k independent samples, we use the simplified notation $\pi_i = \pi_i^{(1,\ldots,k)}$, $\pi_{ij} = \pi_{ij}^{(1,\ldots,k)}$ and $I_i = I_i^{(1,\ldots,k)}$. Since the individual samples may overlap, the resulting design is not of fixed size. The unbiased combined single count (SC) estimator, which has Horvitz-Thompson form, is given by

$$\hat{Y}_{\mathrm{SC}} = \sum_{i\in S}\frac{y_i}{\pi_i}.$$

The variance is

$$V\left(\hat{Y}_{\mathrm{SC}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j},$$

and an unbiased variance estimator is

$$\hat{V}\left(\hat{Y}_{\mathrm{SC}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{I_iI_j}{\pi_{ij}}.$$

For the combination of independent samples with positive first order inclusion probabilities we always have $\pi_{ij} > 0$ for all pairs (i, j), which is the requirement for the above variance estimator to be unbiased. In terms of MSE it may be beneficial not to use the single count estimator, but instead use an estimator that accounts for the random sample size. However, here we restrict ourselves to using only unbiased estimators.

3.2 Combining with multiple count

We first look at how to combine two independent samples $S^{(1)}$ and $S^{(2)}$ selected from the same population, where we allow for each unit to possibly be included multiple times. The number of inclusions of unit i in the combined sample is denoted by $S_i^{(1,2)}$, and it is the sum of the number of inclusions of unit

i in the two samples we combine, i.e., $S_i^{(1,\,2)} = S_i^{(1)} + S_i^{(2)}$, where $S_i^{(\ell)}$ is the number of inclusions of unit i in sample ℓ. The expected number of inclusions of unit i in the combination is given by

$$E\left(S_i^{(1,\,2)}\right) = E_i^{(1,\,2)} = E_i^{(1)} + E_i^{(2)}, \tag{3.3}$$

where $E_i^{(\ell)} = E\left(S_i^{(\ell)}\right)$ is the expected number of inclusions for unit i in sample $S^{(\ell)}$, $\ell = 1, 2$. The (possibly random) sample size is the sum $\sum_{i=1}^{N} S_i^{(1,\,2)}$ of all individual inclusions and the expected sample size is the sum $\sum_{i=1}^{N} E_i^{(1,\,2)}$ of all individual expected number of inclusions. It can be shown that

$$E\left(S_i^{(1,\,2)} S_j^{(1,\,2)}\right) = E_{ij}^{(1,\,2)} = E_{ij}^{(1)} + E_i^{(1)} E_j^{(2)} + E_i^{(2)} E_j^{(1)} + E_{ij}^{(2)}, \tag{3.4}$$

where $E_{ij}^{(\ell)} = E\left(S_i^{(\ell)} S_j^{(\ell)}\right)$, $\ell = 1,\, 2$ are the second order of expected number of inclusions in sample ℓ. Obviously $E_{ij}^{(\ell)} = \pi_{ij}^{(\ell)}$ if the design for sample ℓ is without replacement. Note that as $S_i^{(\cdot)}$ may take other values than 0 or 1 we have that $E_{ii}^{(\cdot)}$ is generally not equal to $E_i^{(\cdot)}$, but $\pi_{ii}^{(\cdot)} = \pi_i^{(\cdot)}$. The equations (3.3) and (3.4) can be used recursively to obtain $E_i^{(\cdot)}$ and $E_{ij}^{(\cdot)}$ for the combination of an arbitrary number k of independent samples. We then get the recursive formulas

$$E_i^{(1,\,\ldots,\,\ell+1)} = E_i^{(1,\,\ldots,\,\ell)} + E_i^{(\ell+1)}$$

and

$$E_{ij}^{(1,\,\ldots,\,\ell+1)} = E_{ij}^{(1,\,\ldots,\,\ell)} + E_i^{(1,\,\ldots,\,\ell)} E_j^{(\ell+1)} + E_j^{(1,\,\ldots,\,\ell)} E_i^{(\ell+1)} + E_{ij}^{(\ell+1)}.$$

The previous results and (3.4) follow from the fact that $S_i^{(1,\,\ldots,\,\ell+1)} = S_i^{(1,\,\ldots,\,\ell)} + S_i^{(\ell+1)}$ and that $S_i^{(1,\,\ldots,\,\ell)}$ and $S_i^{(\ell+1)}$ are independent. For example, we have

$$\begin{aligned} E_{ij}^{(1,\,\ldots,\,\ell+1)} = E\left(S_i^{(1,\,\ldots,\,\ell+1)} S_j^{(1,\,\ldots,\,\ell+1)}\right) &= E\left(\left(S_i^{(1,\,\ldots,\,\ell)} + S_i^{(\ell+1)}\right)\left(S_j^{(1,\,\ldots,\,\ell)} + S_j^{(\ell+1)}\right)\right) = \\ & E\left(S_i^{(1,\,\ldots,\,\ell)} S_j^{(1,\,\ldots,\,\ell)} + S_i^{(1,\,\ldots,\,\ell)} S_j^{(\ell+1)} + S_j^{(1,\,\ldots,\,\ell)} S_i^{(\ell+1)} + S_i^{(\ell+1)} S_j^{(\ell+1)}\right) = \\ & E_{ij}^{(1,\,\ldots,\,\ell)} + E_i^{(1,\,\ldots,\,\ell)} E_j^{(\ell+1)} + E_j^{(1,\,\ldots,\,\ell)} E_i^{(\ell+1)} + E_{ij}^{(\ell+1)}. \end{aligned}$$

For the combination of k independent samples we now use the simplified notation $E_i = E_i^{(1,\,\ldots,\,k)}$, $E_{ij} = E_{ij}^{(1,\,\ldots,\,k)}$, and $S_i = S_i^{(1,\,\ldots,\,k)}$. The total Y can be estimated without bias with the multiple count (MC) estimator, of which the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) is a special case. It is given by

$$\hat{Y}_{\mathrm{MC}} = \sum_{i=1}^{N} \frac{y_i}{E_i} S_i.$$

We get the Hansen-Hurwitz estimator if $E_i = np_i$, where n is the number of units drawn and p_i, with $\sum_{i=1}^{N} p_i = 1$, are probabilities for a single independent draw. The variance of $\hat{Y}_{\mathrm{MC}}$ can be shown to be

$$V\left(\hat{Y}_{\mathrm{MC}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(E_{ij} - E_i E_j\right)\frac{y_i}{E_i}\frac{y_j}{E_j}.$$

A variance estimator is

$$\hat{V}\left(\hat{Y}_{\mathrm{MC}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(E_{ij} - E_i E_j\right)\frac{y_i}{E_i}\frac{y_j}{E_j}\frac{S_i S_j}{E_{ij}}.$$

It follows directly that the above variance estimator is unbiased, because when combining independent samples with positive first order inclusion probabilities we always have $E_{ij} > 0$ for all pairs (i, j).

3.3 Comparing the combined and separate estimators

Two examples that illustrate that the combined estimator is not necessarily as good as the best separate estimator.

Example 3: *Assume that the first sample, $S^{(1)}$, is of fixed size with $\pi_i^{(1)} \propto y_i$, and that the second is a simple random sample with $\pi_i^{(2)} = n/N$. Then the Horvitz-Thompson estimator $\hat{Y}_1 = \sum_{i\in S^{(1)}} y_i/\pi_i^{(1)}$, has zero variance, but the combined single count estimator with $\pi_i = \pi_i^{(1)} + \pi_i^{(2)} - \pi_i^{(1)}\pi_i^{(2)}$ has positive variance. Thus the combined estimator is worse than the best separate estimator.*

Example 4: *Assume that the design for the first sample is stratified in such a way that there is no variation within strata. Then the separate estimator $\hat{Y}_1 = \sum_{i\in S^{(1)}} y_i/\pi_i^{(1)}$ has zero variance. If the first sample is combined with a non-stratified second sample, then the resulting design does not have fixed sample sizes for the strata. Thus, the combined estimator has a positive variance.*

These examples tell us that we need to be careful before combining very different designs, such as an unequal probability design with an equal probability design or a stratified with a non-stratified sampling design. Especially, we need to be careful if we plan to estimate the total directly based on the combined sample. When combining samples from relatively similar designs, it is however likely that the combined estimator becomes better than the best of the separate estimators.

Next, we investigate how to use the combined approach for estimation of the separate variances and then use the linear combination estimator. In fact, as we will see later, using the combined approach for variance estimation of separate variances can act stabilizing for the weights in the linear combination with weights based on estimated variances. There is a sort of pooling effect for the variance estimators when they are estimated with the same set of information.

3.4 Using the combined sample for estimation of variances of separate estimators

An alternative to estimating directly the total Y based on the combined design is to use the combined design to estimate the variances of the separate estimators, and then proceed with a linear combination of the separate estimators. We assume access to k independent samples and that we want to estimate the variance of a separate estimator, whose variance is a double sum over the population units. There are two main options for the variance estimator; multiply by

$$\frac{I_i I_j}{\pi_{ij}} \quad \text{or} \quad \frac{S_i S_j}{E_{ij}}$$

in the variance formula to obtain an unbiased estimator of the variance based on the combination of all the k samples $S^{(\ell)}$, $\ell = 1, \ldots, k$. For example, assuming that the variance of $\hat{Y}_1$ is

$$V\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}},$$

we can use the combination of $S^{(\ell)}$, $\ell = 1, \ldots, k$, to estimate $V\left(\hat{Y}_1\right)$ by the single count estimator

$$\hat{V}_{\text{SC}}\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{I_i I_j}{\pi_{ij}}$$

or the multiple count estimator

$$\hat{V}_{\text{MC}}\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{S_i S_j}{E_{ij}}.$$

Note that $\pi_{ij} = \pi_{ij}^{(1, \ldots, k)}$, $I_i = I_i^{(1, \ldots, k)}$, $E_{ij} = E_{ij}^{(1, \ldots, k)}$ and $S_i = S_i^{(1, \ldots, k)}$, so the above variance estimators use all available information on the target variable. Hence, these variance estimators can be thought of as general pooled variance estimators. It follows directly that both estimators are unbiased because all designs have positive first order inclusion probabilities, which imply that all π_{ij} and all E_{ij} are strictly positive. Interestingly, the above variance estimators are unbiased even if the separate design 1 has some second order inclusion probabilities that are zero, which prevent unbiased variance estimation based on the sample $S^{(1)}$ alone.

Despite the appealing property of producing an unbiased variance estimator for any design, the above variance estimators cannot be recommended for designs with a high degree of zero second order inclusion probabilities (such as systematic sampling). The estimators can be very unstable for such designs and can produce a high proportion of negative variance estimates.

As we will see, if we intend to use a linear combination estimator, it is important that all variances are estimated in the same way. Then it is likely that the ratios, e.g.,

$$\frac{\hat{V}_{\text{SC}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{SC}}\left(\hat{Y}_2\right)} \quad \text{and} \quad \frac{\hat{V}_{\text{MC}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{MC}}\left(\hat{Y}_2\right)}$$

become stable (have small variance). The ratios become more stable because the estimators in the numerator and denominator are based on the same information and are estimated with the same weights for all the pairs (i, j) in all estimators. With estimated variances we get

$$\hat{\alpha}_i = \left[\sum_{j=1}^{k}\frac{\hat{V}\left(\hat{Y}_i\right)}{\hat{V}\left(\hat{Y}_j\right)}\right]^{-1},$$

so if the ratios of variance estimators have small variance then $\hat{\alpha}_i$ has small variance. The weighting in the linear combination $\hat{Y}_L^*$ then becomes stabilized. As the following example demonstrates, the ratio of the variance estimators can even have zero variance. Thus it can sometimes provide the optimal weighting even if the variances are unknown.

Example 5: *Assume we want to combine estimates resulting from two simple random samples of different sizes. This can of course be done optimally without estimating the variances, but as an example we will use the above approach to estimate the separate variances by use of the combined sample. In this case the use of the estimators* $\hat{V}_{\text{SC}}\left(\hat{Y}_1\right)$ *and* $\hat{V}_{\text{SC}}\left(\hat{Y}_2\right)$ *provides the optimal weighting, and so does* $\hat{V}_{\text{MC}}\left(\hat{Y}_1\right)$ *and* $\hat{V}_{\text{MC}}\left(\hat{Y}_2\right)$. *This result follows from the fact that if both designs are simple random sampling we have*

$$\frac{\hat{V}_{\text{SC}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{SC}}\left(\hat{Y}_2\right)} = \frac{\hat{V}_{\text{MC}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{MC}}\left(\hat{Y}_2\right)} = \frac{V\left(\hat{Y}_1\right)}{V\left(\hat{Y}_2\right)},$$

which is straightforward to verify. For two simple random samples the situation corresponds to using a pooled estimate for S^2 (*the population variance of* y) *in the expressions for the variance estimates, and this pooled estimate is then cancelled out in the calculation of the weights.*

The conclusion is that this procedure is likely to provide a more stable weighting also for designs that deviate from simple random sampling as long as the involved designs have large entropy (a high degree of randomness). The problem of bias for the linear combination estimator with estimated variances will be reduced compared to using separate and thus independent variance estimators.

We believe that this can be a very interesting alternative, because the estimator of the total based on the combined design does not necessarily provide a smaller variance than the best of the separate estimators. With this strategy we can improve the separate variance estimators, especially for a smaller sample (if data is available for a larger sample). Hence the resulting linear combination with jointly estimated variances can be a very competitive strategy.

With single count we might use a ratio type variance estimator such as the following

$$\hat{V}_R\left(\hat{Y}_1\right) = \frac{N^2}{\gamma_{1,\ldots,k}} \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{I_i I_j}{\pi_{ij}},$$

where $\gamma_{1,\ldots,k} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{I_i I_j}{\pi_{ij}}$. For multiple count we can replace $I_i I_j / \pi_{ij}$ with $S_i S_j / E_{ij}$. This ratio estimator uses the known size of the population of pairs $(i, j) \in \{1, 2, \ldots, N\}^2$, which is N^2, and divides by the sum of the sample weights for the pairs. Note that $E\left(\gamma_{1,\ldots,k}\right) = N^2$. This correction is useful because the number of pairs in the estimator may be random (since the union of the samples may have random size). This rescales the sample (of pairs) weights to sum to N^2. This will introduce some bias (as usual for ratio estimators), but the idea is that this will reduce the variance of the variance estimator. However, this approach is only useful if we are interested in the separate variance as the correction term will be the same for all separate variance estimators. Hence it does not change the weighting of a linear combination estimator with estimated variances.

4 Simulation examples

Two Monte-Carlo simulation examples are presented here. In the first example we combine two Poisson samples using inclusion probabilities approximately proportional to the target variable. In the second example we combine an unstratified simple random sample with a stratified simple random sample.

4.1 Combining two Poisson samples

We generate a population of size $N = 200$, with an auxiliary variable $X_i \sim N(\mu = 20, \sigma^2 = 16)$. The target variable is generated as $(Y_i \mid X_i = x_i) = x_i + \epsilon_i$, where $\epsilon_i \sim N\left(0, (x_i/20)^2\right)$. Two sets of inclusion probabilities are generated $\pi_i^{(1)} \propto \pi_i^{(2)} \propto x_i$, where $\sum_{i=1}^{N} \pi_i^{(1)} = n_1$ and $\sum_{i=1}^{N} \pi_i^{(2)} = n_2$. We let the expected sample sizes be $n_1 = 15$ and $n_2 = 25$. For simplicity we let both designs be Poisson designs (where units are selected independently). This allows us to calculate exactly the variances for both separate estimators (and thus the optimal linear combination) and for the combined samples with single and multiple count. For the strategies with linear combination using estimated variances, we performed a Monte-Carlo simulation with 1,000,000 repeated sample selections. True variances for the two separate HT estimators, the SC/MC estimators for the combined samples and the optimal linear combination of the separate estimators are presented (Table 4.1). Simulation results for the different linear combinations with estimates variances are also presented (Table 4.1).

Table 4.1
Results for the combination of Poisson samples. True variances for the two separate HT estimators, the SC/MC estimators for the combined samples and the optimal linear combination of the separate estimators. Simulation results, in terms of estimated bias and MSE, for three linear combination estimators with estimated variances

Estimator	Bias (Rel. bias)	MSE
$\hat{Y}_1$	0	1,053,083
$\hat{Y}_2$	0	596,069
$\hat{Y}_{\text{SC}}$	0	361,088
$\hat{Y}_{\text{MC}}$	0	380,929
$\hat{Y}_L$ Optimal	0	380,626
$\hat{Y}_L^*$ Separate	-92.8 (-2.24%)	412,248
$\hat{Y}_L^*$ Pooled SC	1.6 (+0.04%)	381,106
$\hat{Y}_L^*$ Pooled MC	1.6 (+0.04%)	381,106

Using combined (pooled) variance estimators reduced both the bias and the variance for a linear combination in comparison to using separate variance estimators. For this example, the linear combination with pooled variance estimation came very close to the optimal linear combination in performance. The negative bias with separate variance estimators is mainly due to a positive correlation between the total

estimator and its variance estimator under the Poisson design. For this setting, the best result was obtained by combining the samples using a single count.

4.2 Combining an unstratified SRS with a stratified SRS

Here we generated a population of size $N = 1{,}000$, with two strata of sizes $N_1 = 600$ and $N_2 = 400$. The target variable y_i, $i = 1, \ldots, N$, was generated as follows. In stratum 1 there were 500 y_i's equal to zero and the other 100 y_i's were drawn from $N(\mu_1 = 10, \sigma^2 = 4)$. In stratum 2 there were 300 y_i's equal to zero and the other 100 y_i's were drawn from $N(\mu_2 = 15, \sigma^2 = 4)$. The first sample is an unstratified simple random sample of size $n = 50$ and the second sample is a stratified simple random sample with stratum sample sizes $n_1 = 30$ and $n_2 = 20$. The variances for both separate HT estimators and for the combined samples with single and multiple count were calculated exactly. A Monte-Carlo simulation with 10,000 repetitions was performed to evaluate the performance of a linear combination estimator with estimated variances. The results are presented in Table 4.2. Bias is reduced by using a linear combination with pooled variance estimators compared with using separate variance estimators. Also for this setting, the best result was obtained by combining the samples using a single count.

Table 4.2
Results for the combination of a SRS and a stratified SRS. True variances for the two separate HT estimators, the SC/MC estimators for the combined sample and the optimal linear combination of the separate estimators. Simulation results, in terms of estimated bias and MSE, for three linear combination estimators with estimated variances

Estimator	Bias (Rel. bias)	MSE
$\hat{Y}_1$	0	516,835
$\hat{Y}_2$	0	498,321
$\hat{Y}_{\text{SC}}$	0	248,888
$\hat{Y}_{\text{MC}}$	0	253,789
$\hat{Y}_L$ Optimal	0	253,704
$\hat{Y}_L^*$ Separate	-77 (-3%)	287,680
$\hat{Y}_L^*$ Pooled SC	9 (+0.4%)	257,229
$\hat{Y}_L^*$ Pooled MC	9 (+0.4%)	257,217

5 Discussion

The simulation examples in the previous section are only intended to demonstrate the different approaches and we make no claim of the generality of the result. However, we find it very likely that using pooled variance estimators is better than using separate variance estimators in a linear combination estimator, especially in cases where the separate total estimators are highly correlated to their variance estimators.

We have presented in detail how to combine independent probability samples and derived corresponding design features needed to do unbiased estimation and variance estimation. The danger of using the combined sample approach for very different designs has been illustrated. Moreover, we have shown that there is often a risk for a strong positive correlation between the HT estimator and its variance estimator. Such dependence can be a source for bias if estimated variances are used in a linear combination. Thus, as an alternative approach, we have shown how to use the combined sample to estimate separate variances. This alternative approach can lead to more stable weights in a linear combination of separate estimators, and has potential to reduce both bias and variance.

There are of course limitations to when this methodology can be applied due to our assumption of fully known designs and use of the same frame with identifiable units. Sensitivity for deviations from some of these assumptions, such as having unidentifiable units or using approximate second order inclusion probabilities, needs further investigation.

In particular, knowledge of this methodology is important if an initial sampling effort was proven insufficient. Such situations are common in e.g., environmental monitoring (Christensen and Ringvall, 2013). Then a complementary sample may be designed in such a way that it allows for a combination with improved efficiency.

Acknowledgements

We thank an anonymous reviewer and an associate editor for helpful comments that improved the paper. This work was funded by the Swedish Science Council grant 340-2013-5076.

References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81(396), 1074-1079.

Christensen, P., and Ringvall, A.H. (2013). Using statistical power analysis as a tool when designing a monitoring program: Experience from a large-scale Swedish landscape monitoring program. *Environmental Monitoring and Assessment*, 185(9), 7279-7293.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101-129.

Cochran, W.G., and Carroll, S.P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, 9(4), 447-459.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. and Ståhl, G. (2014). Adapting National Forest Inventories to changing requirement-the case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica*, 48(3), article id 1095.

Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statistical Methodology*, 7(2), 84-97.

Gregoire, T.G., and Schabenberger, O. (1999). Sampling-skewed biological populations: Behavior of confidence intervals for the population total. *Ecology*, 80, 1056-1065.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4), 333-362.

Koricheva, J., Gurevitch, J. and Mengersen, K. (Eds.) (2013). Handbook of meta-analysis in ecology and evolution. Princeton University Press.

Lohr, S.L. (2009). Multiple-frame surveys. *Handbook of Statistics*, 29, 71-88.

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.

Næsset, E., and Gjevestad, J.G. (2008). Performance of GPS precise point positioning under conifer forest canopies. *Photogrammetric Engineering & Remote Sensing*, 74(5), 661-668.

O'Muircheartaigh, C., and Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. In *Proceedings of the 2002 Joint Statistical Meetings*, American Statistical Association, 2557-2562.

Roberts, G., and Binder, D. (2009). Analyses based on combining similar information from multiple surveys. In *Joint Statistical Meetings 2009 Survey Research Methods Section*, 2138-2147.

Rubin, D.B., and Weisberg, S. (1974). The variance of a linear combination of independent estimators using estimated weights. *ETS Research Bulletin Series*, 1974(2), i-5.

Schmidt, F.L., and Hunter, J.E. (2014). Methods of meta-analysis: Correcting error and bias in research findings. Sage publications.

Ståhl, G., Allard, A., Esseen, P.-A., Glimskår, A., Ringvall, A., Svensson, J., Sundquist, S., Christensen, P., Gallegos Torell, Å., Högström, M., Lagerqvist, K., Marklund, L., Nilsson, B. and Inghe, O. (2011). National Inventory of Landscapes in Sweden (NILS)-scope, design, and experiences from establishing a multiscale biodiversity monitoring system. *Environmental Monitoring and Assessment*, 173(1-4), 579-595.

Tomppo, E., Gschwantner, T., Lawrence, M. and McRoberts, R.E. (Eds.) (2009). National Forest Inventories: Pathways for Common Reporting. *Springer Science & Business Media*.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 365-390
Statistics Canada, Catalogue No. 12-001-X

Bayesian benchmarking of the Fay-Herriot model using random deletion

Balgobin Nandram, Andreea L. Erciulescu and Nathan B. Cruze[1]

Abstract

Benchmarking lower level estimates to upper level estimates is an important activity at the United States Department of Agriculture's National Agricultural Statistical Service (NASS) (e.g., benchmarking county estimates to state estimates for corn acreage). Assuming that a county is a small area, we use the original Fay-Herriot model to obtain a general Bayesian method to benchmark county estimates to the state estimate (the target). Here the target is assumed known, and the county estimates are obtained subject to the constraint that these estimates must sum to the target. This is an external benchmarking; it is important for official statistics, not just NASS, and it occurs more generally in small area estimation. One can benchmark these estimates by "deleting" one of the counties (typically the last one) to incorporate the benchmarking constraint into the model. However, it is also true that the estimates may change depending on which county is deleted when the constraint is included in the model. Our current contribution is to give each small area a chance to be deleted, and we call this procedure the random deletion benchmarking method. We show empirically that there are differences in the estimates as to which county is deleted and that there are differences of these estimates from those obtained from random deletion as well. Although these differences may be considered small, it is most sensible to use random deletion because it does not give preferential treatment to any county and it can provide small improvement in precision over deleting the last one benchmarking as well.

Key Words: Constraint; Direct estimates; Fay-Herriot model; Multivariate normal density; Official statistics; Small area estimation.

1 Introduction

In official statistics, it is important for lower level estimates to sum to upper level estimates. For example, the National Agricultural Statistics Service (NASS) often uses a "top-down" sequence in the release of its official estimates in which national and state estimates, e.g., estimated corn acreage totals, are published prior to the completion of supplemental data collection and estimation of corresponding county estimates (Cruze, Erciulescu, Nandram, Barboza and Young, 2019). Within these small administrative areas, the survey data often become sparse. Several popular modeling techniques give rise to more reliable small area estimates. However, the small area estimates may not automatically satisfy relationships with estimates at other levels of aggregation, and benchmarking procedures may be applied to enforce consistency among estimates.

There is a considerable history on benchmarking techniques which have been used to impose agreement among multiple levels and to protect against possible model misspecification. These procedures can be broadly classified in two categories: internal benchmarking, in which a target is derived from current survey data, and external benchmarking, in which a desired target may be taken from other sources such as administrative data or previously established estimates. We discuss external benchmarking, in accordance with NASS's "top-down" procedure, of the Fay-Herriot (FH) model (Fay and Herriot, 1979).

1. Balgobin Nandram, Worcester Polytechnic Institute and USDA National Agricultural Statistics Service, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609. E-mail: balnan@wpi.edu; Andreea L. Erciulescu, Westat, 1600 Research Boulevard, Rockville, MD 20850. E-mail: alerciulescu@gmail.com; Nathan B. Cruze, USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Room 6412 A, Washington, DC 20250-2054. E-mail: nathan.cruze@nass.usda.gov.

The most recent review of small area estimation is given in Pfeffermann (2013), but see Rao and Molina (2015) for the most updated textbook on small area estimation. Earlier Jiang and Lahiri (2006) gave an extensive review of the classical inferential approach for linear and generalized linear mixed models that are used in small area estimation. There are discussions of benchmarking in these works as well, but the latter review was not on the hierarchical Bayes approach that is of primary interest in this paper.

Within the hierarchical Bayes framework, You, Rao and Dick (2004) studied benchmarked estimators for small area estimation based on unmatched sampling and linking models proposed earlier by You and Rao (2002). They applied this approach to undercoverage estimation for the ten provinces across Canada for the 1991 Canadian Census. Wang, Fuller and Qu (2008) gave a characterization of the best linear unbiased predictor (BLUP) for small area means under an area level model that satisfies a benchmarking constraint and minimizes the loss function criterion that all linear unbiased predictors satisfy. They also presented an alternative way of imposing the benchmarking constraint such that the BLUP estimator would have a self-calibrated property (discussed in You and Rao, 2002). Wang et al. (2008) characterized a class of benchmarked estimators as the predictors that minimize a quadratic loss function subject to a benchmarking restriction. Their proposed self-calibrated augmented model reduces bias both at the overall and small area level. Other benchmarking procedures are given by Bell, Datta and Ghosh (2013), Ghosh and Steorts (2013), Pfeffermann, Sikov and Tiller (2014) and Pfeffermann and Tiller (2006).

Whether fitting unit-level or area-level models, incorporating a fixed, external target amounts to imposing the general constraint $\sum_{i=1}^{\ell} w_i \theta_i = a$, where a is a known constant and the θ_i denote small area estimates to be benchmarked; for totals, the weights w_i are all equal to 1. One way to do so is by using the following transformation, $\phi = a - \sum_{i=1}^{\ell} \theta_i$, keeping θ_i, $i = 1, \ldots, \ell - 1$, unchanged and "deleting" the last small area, replacing it with $\theta_\ell = \phi - \left(a - \sum_{i=1}^{\ell-1} \theta_i\right)$. Janicki and Vesper (2017) introduced a slightly different transformation, $\phi_i = \theta_i$, $i = 1, \ldots, \ell - 1$, $\phi_\ell = \sum_{i=1}^{\ell} \theta_i$, which is essentially an internal benchmarking that preserves the sum of all ℓ estimates. If that sum (of all ℓ unbenchmarked estimates) were prescribed as an external target, then $\theta_\ell = \phi_\ell - \sum_{i=1}^{\ell-1} \theta_i$, and Janicki and Vesper's transformation becomes equivalent to deleting last small area estimate.

External benchmarking procedures, which deleted the last small area estimate, were explored by Nandram and Sayit (2011) and by Nandram, Berg, and Barboza (2014) for the purposes of benchmarking binomial probabilities and forecasts of crop yield, respectively. (In both of these contexts the constraint was actually imposed on the weighted sum of small area estimates.) Erciulescu, Cruze, and Nandram (2019) considered a variety of external benchmarking techniques including deletion, difference benchmarking, and ratio benchmarking in the context of hierarchical Bayesian small area models. Collectively, the external benchmarking constraint has been inserted in the likelihood function (e.g., Toto and Nandram, 2010), the joint density of the area effects (e.g., Nandram and Sayit, 2011), or in the posterior density of the area effects (Janicki and Vesper, 2017), although the latter choice is using the prior knowledge or requirements embodied in the constraint a posteriori rather than on the prior distributions themselves.

Datta, Ghosh, Steorts and Maples (2011), henceforth DGSM, proposed a general class of constrained Bayes estimators to provide benchmarked estimates. Referring specifically to the method of Toto and

Nandram (2010) for unit level models, DGSM wrote the following: *"A disadvantage to such an approach is that results can differ depending on which unit is dropped"*. This statement also applies to Nandram and Toto (2010), Nandram, Toto and Choi (2011), Janicki and Vesper (2017) and others. It also applies in the same way to an area-level model subject to an external constraint. The procedures of DGSM depend on an important area-specific parameter (see Section 4). This parameter also has several different specifications, and it can be argued that the resulting estimates could also be affected by the choice of specification. Moreover, the procedures of DGSM do not provide posterior standard errors or credible intervals.

In response to DGSM's comment on the last area deletion benchmarking, we introduce a random deletion benchmarking, giving a chance to each area to be deleted, and not just the last one. The random deletion benchmarking method is motivated mathematically in Appendix A. Empirical results show that there are slight differences between the last one deletion benchmarking and the random deletion benchmarking.

In this paper, we discuss random deletion benchmarking in the context of a Bayesian FH (BFH) model. In Section 2, the BFH model without constraint is introduced. The methodology for imposing an external target on the BFH model through random deletion is developed in Section 3. In Section 4, we describe the empirical studies to assess features of estimates obtained from random deletion benchmarking, including related measures of uncertainty. Finally, Section 5 has concluding remarks; more technical details are provided in several appendices.

2 Bayesian Fay-Herriot model

Assume that the observed data are $\left(\hat{\theta}_i,\, s_i\right)$, $i = 1, \ldots, \ell$, where $\hat{\theta}_i$ and s_i are respectively an estimate and its standard error (for simplicity, assumed known) of a quantity under study, e.g., the i^{th} area total θ_i. The BFH model is

$$\hat{\theta}_i \,\big|\, \theta_i \overset{\text{ind}}{\sim} \text{Normal}\left(\theta_i,\, s_i^2\right),$$

$$\theta_i \,|\, \boldsymbol{\beta},\, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta},\, \sigma^2\right),$$

$$\pi\left(\boldsymbol{\beta},\, \sigma^2\right), \tag{2.1}$$

where $i = 1, \ldots, \ell$, $\mathbf{x}_i$ are a set of covariates with p components (including intercept) and $\pi\left(\boldsymbol{\beta},\, \sigma^2\right)$ is the joint prior distribution for $\left(\boldsymbol{\beta},\, \delta^2\right)$. A priori it is assumed that $\pi\left(\boldsymbol{\beta},\, \sigma^2\right) = \pi\left(\boldsymbol{\beta}\right)\pi\left(\sigma^2\right)$, i.e., $\boldsymbol{\beta}$ and σ^2 are independent, with

$$\pi\left(\boldsymbol{\beta}\right) \propto 1 \quad \text{and} \quad \pi\left(\sigma^2\right) = 1\big/\left(1+\sigma^2\right)^2. \tag{2.2}$$

By Bayes' theorem, the joint posterior density of $\ell + p + 1$ model parameters is

$$\pi\left(\boldsymbol{\theta},\, \boldsymbol{\beta},\, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right) \propto \frac{1}{\left(1+\sigma^2\right)^2}\left(\frac{1}{\sigma^2}\right)^{\ell/2} \prod_{i=1}^{\ell}\left\{\exp\left[-\frac{1}{2}\left\{\frac{1}{s_i^2}\left(\hat{\theta}_i - \theta_i\right)^2 + \frac{1}{\sigma^2}\left(\theta_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right]\right\}. \tag{2.3}$$

In (2.2), $\pi(\sigma^2)$ is a proper prior distribution, flatter than $\pi(\sigma^2) \propto 1/\sigma^2$ near zero, with no moments. In fact, any proper prior for σ^2 is fine; an improper prior on σ^2 may lead to improper posterior density. Because $\pi(\boldsymbol{\beta})$ is improper, the product $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2)$ *is improper*, and this could cause the joint posterior density of $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)$ to be improper, an undesirable scenario. Theorem 1 below establishes the propriety of the joint posterior density (2.3).

More details about the BFH model are presented in Appendix B. Specifically, letting $\lambda_i = \frac{\sigma^2}{s_i^2+\sigma^2}$, $i = 1, \ldots, \ell$, we have shown that

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma^2, \hat{\boldsymbol{\theta}} \overset{\text{ind}}{\sim} \text{Normal}\left\{\lambda_i\hat{\theta}_i + (1-\lambda_i)\mathbf{x}_i'\boldsymbol{\beta}, (1-\lambda_i)\sigma^2\right\}, \; i = 1, \ldots, \ell, \tag{2.4}$$

$$\boldsymbol{\beta} \,|\, \sigma^2, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\hat{\boldsymbol{\beta}}, \hat{\Sigma}\right), \tag{2.5}$$

$$\pi_3\left(\sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right) \propto Q(\sigma^2)\frac{1}{(1+\sigma^2)^2}, \tag{2.6}$$

where

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}\sum_{i=1}^{\ell}\frac{\hat{\theta}_i\mathbf{x}_i}{s_i^2+\sigma^2}, \; \hat{\Sigma}^{-1} = \sum_{i=1}^{\ell}\frac{\mathbf{x}_i\mathbf{x}_i'}{s_i^2+\sigma^2},$$

$$Q(\sigma^2) = \left|\hat{\Sigma}\right|^{1/2}\prod_{i=1}^{\ell}\frac{1}{(s_i^2+\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{\ell}\frac{1}{s_i^2+\sigma^2}\left(\hat{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2\right\}.$$

Theorem 1

The joint posterior density (2.3) is proper provided the design matrix is full rank.

Proof of Theorem 1

Because the design matrix is full rank, $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$ are well defined for all σ^2. This implies that $Q(\sigma^2)$ is bounded in σ^2. Therefore, as $\pi(\sigma^2) = \frac{1}{(1+\sigma^2)^2}$ is proper, the posterior density $\pi_3\left(\sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right)$ is proper. Then, by applying the multiplication rule of probability, it follows that the joint posterior density, $\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right)$, is proper.

With propriety assured, sampling from the joint posterior density and inference about θ_i can be achieved through a simple Monte Carlo procedure. Using the multiplication rule and drawing samples from (2.6), (2.5) and (2.4), the procedure follows. First, draw a sample from $\pi_3\left(\sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right)$ in (2.6). Draws from this distribution can be made using a grid method; see Appendix B. It is then easy to draw samples from the conditional posterior density of $\boldsymbol{\beta}$ in (2.5). Finally, samples of the θ_i can be drawn independently from (2.4). Sampling in this manner from the joint posterior density under the unconstrained BFH model does not need monitoring (unlike Markov chain Monte Carlo methods). The unconstrained BFH model will

provide a basis for comparison of the estimates and related measures of uncertainty obtained under the proposed random deletion benchmarking method.

3 Random deletion methodology

As remarked earlier, the random deletion methodology is obtained by introducing a new variable which takes the values $1, \ldots, \ell$ with equal, possibly different, probabilities (weights). In Section 3.1, we show how to construct the joint posterior density of the parameters of the BFH model under random deletion, and in Section 3.2, we show how to sample from the joint posterior density.

3.1 Construction of the joint posterior density

The Basic Benchmarking Theorem is motivated by the work reviewed in Section 1. The next goal is to construct the joint prior density of $\theta_1, \ldots, \theta_\ell$ subject to the benchmarking constraint that $\sum_{i=1}^{\ell} \theta_i = a$, where a is a known external target. The joint prior density of $\theta_1, \ldots, \theta_\ell$ will be used to complete the constrained Fay-Herriot model for the last one deletion benchmarking (see Section 1).

Theorem 2

Let $\theta_i \overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{u}_i'\boldsymbol{\beta}, \delta^2\right)$, $i = 1, \ldots, \ell$. Then, under the constraint, $\sum_{i=1}^{\ell} \theta_i = a$, where a is constant, letting $\boldsymbol{\theta}_{(\ell)}$ be the vector of all the θ_i except the last one, the joint density of θ_i, $i = 1, \ldots, \ell$, is

$$\boldsymbol{\theta}_{(\ell)} \sim \text{Normal}\left\{\left(I - \frac{1}{\ell} J\right)\mathbf{c},\ \delta^2\left(I - \frac{1}{\ell} J\right)\right\},\ \theta_\ell = a - \sum_{i=1}^{\ell-1} \theta_i, \tag{3.1}$$

$\mathbf{c}' = a\mathbf{j}' + \left((\mathbf{u}_1 - \mathbf{u}_\ell)'\boldsymbol{\beta}, \ldots, (\mathbf{u}_{\ell-1} - \mathbf{u}_\ell)'\boldsymbol{\beta}\right)$, J and $\mathbf{j}$ are respectively a $(\ell - 1) \times (\ell - 1)$ matrix and a $(\ell - 1)$ vector of ones.

Proof of Theorem 2

See Appendix C.

The proof of Theorem 2 uses the multivariate normal distribution, and it will be used to prove the more general theorem when the prior is adjusted to delete any area.

The constrained BFH model is

$$\hat{\theta}_i \,|\, \theta_i \overset{\text{ind}}{\sim} \text{Normal}\left(\theta_i, s_i^2\right),\ i = 1, \ldots, \ell, \tag{3.2}$$

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\right),\ \sum_{i=1}^{\ell} \theta_i = a,$$

$$\pi\left(\boldsymbol{\beta}, \sigma^2\right).$$

The constraint on the prior on $\theta_1, \ldots, \theta_\ell$ essentially adjusts the joint prior density. However, to incorporate the constraint, we will use Theorem 2 to adjust the posterior density under the unconstrained model.

For $i = 1, \ldots, \ell$, let $\lambda_i = \frac{\sigma^2}{\sigma^2 + s_i^2}$ and $\hat{\hat{\theta}}_i = \lambda_i \hat{\theta}_i + (1 - \lambda_i)\, \mathbf{x}_i' \boldsymbol{\beta}$. Also, let $\hat{\sigma}_i^2 = \sigma^2 (1 - \lambda_i)$, $i = 1, \ldots, \ell$. Then, under the unconstrained model the joint posterior density is

$$\pi_{nb}\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right) \propto \pi(\boldsymbol{\beta}, \sigma^2) \prod_{i=1}^{\ell} \left[\text{Normal}_{\hat{\theta}_i}\left(\mathbf{x}_i' \boldsymbol{\beta}, \frac{\sigma^2}{\lambda_i}\right) \text{Normal}_{\theta_i}\left(\hat{\hat{\theta}}_i, \hat{\sigma}_i^2\right) \right].$$

We now extend the result of Theorem 2, to reflect our interest in the constrained BFH model. Theorem 3, below, is used to construct the random deletion benchmarking method.

Theorem 3

Using a general notation, let $y_i \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, \ell$. Let $\mathbf{y}_{(i)}$ denote the vector of all the y_i except the i^{th} one. Also, let $v_i = \sigma_i^2 \Big/ \sqrt{\sum_{j=1}^{\ell} \sigma_j^2}$, $i = 1, \ldots, \ell$, and $v_i^* = \sigma_i^2 \Big/ \sum_{j=1}^{\ell} \sigma_j^2$, $i = 1, \ldots, \ell$. Under the constraint $\sum_{i=1}^{\ell} y_i = a$,

$$\mathbf{y}_{(i)} \sim \text{Normal}\left\{ \mathbf{u}_{(i)} - \left(\sum_{j=1}^{\ell} u_j - a \right) \mathbf{v}_{(i)}^*,\ \text{diagonal}\left(\boldsymbol{\sigma}_{(i)}^2\right) - \mathbf{v}_{(i)} \mathbf{v}_{(i)}' \right\} \tag{3.3}$$

with $y_i = a - \sum_{j \neq i} y_j$. Then,

$$p(\mathbf{y},\, z = r \,|\, \phi = 0) = \delta_{y_r}\left(a - \sum_{j \neq r} y_j \right) \text{Normal}_{\mathbf{y}_{(r)}} \left\{ \boldsymbol{\mu}_{(r)} - \left(\sum_{j=1}^{\ell} \mu_j - a \right) \mathbf{v}_{(r)}^*,\ \text{diagonal}\left(\boldsymbol{\sigma}_{(r)}^2\right) - \mathbf{v}_{(r)} \mathbf{v}_{(r)}' \right\}, \tag{3.4}$$

for $r = 1, \ldots, \ell$ and $\psi = a - \sum_{i=1}^{\ell} y_i$.

Proof of Theorem 3

The proof of Theorem 3 is similar to Theorem 2. See Appendix C.

In what follows, one of the ℓ area parameters will be deleted randomly (i.e., with probability $1/\ell$). Let $z = 1, \ldots, \ell$ represent the county that is deleted. That is, $P(z = r) = 1/\ell$, $r = 1, \ldots, \ell$. Then, under the constrained BFH model, using Theorem 3, the joint posterior density is

$$\begin{aligned} \pi_b\left(\boldsymbol{\theta}, z = r, \boldsymbol{\beta}, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}\right) \ \propto\ & \pi(\boldsymbol{\beta}, \sigma^2) \prod_{i=1}^{\ell} \left[\text{Normal}_{\hat{\theta}_i}\left(\mathbf{x}_i' \boldsymbol{\beta}, \frac{\sigma^2}{\lambda_i}\right) \right] \times \delta_{\theta_r}\left(a - \sum_{j \neq r} \theta_j \right) \\ & \times \text{Normal}_{\boldsymbol{\theta}_{(r)}} \left\{ \hat{\hat{\boldsymbol{\theta}}}_{(r)} - \left(\sum_{j=1}^{\ell} \hat{\hat{\theta}}_j - a \right) \mathbf{v}_{(r)}^*,\ \text{diagonal}\left(\hat{\boldsymbol{\sigma}}_{(r)}^2\right) - \mathbf{v}_{(r)} \mathbf{v}_{(r)}' \right\}, \end{aligned} \tag{3.5}$$

where $r = 1, \ldots, \ell$, and for $i = 1, \ldots, \ell$, $v_i^* = \hat{\sigma}_i^2 \Big/ \sum_{j=1}^{\ell} \hat{\sigma}_j^2$ and $v_i = \hat{\sigma}_i^2 \Big/ \sqrt{\sum_{j=1}^{\ell} \hat{\sigma}_j^2}$.

3.2 Sampling the joint posterior density

Unlike the BFH model, the constrained model in (3.2) cannot be fit using random draws; we use a Gibbs sampler. The joint conditional posterior density (cpd) of $(\boldsymbol{\theta}, z)$ is

$$\pi\left(\boldsymbol{\theta},\, z = r \,|\, \boldsymbol{\beta},\, \sigma^2,\, \hat{\boldsymbol{\theta}}\right) \;\propto\; \delta_{\theta_r}\left(a - \sum_{j \neq r} \theta_j\right)$$
$$\times\; \text{Normal}_{\boldsymbol{\theta}_{(r)}}\left\{\hat{\hat{\boldsymbol{\theta}}}_{(r)} - \left(\sum_{j=1}^{\ell} \hat{\hat{\theta}}_j - a\right)\mathbf{v}^*_{(r)},\, \text{diagonal}\left(\hat{\boldsymbol{\sigma}}^2_{(r)}\right) - \mathbf{v}_{(r)}\mathbf{v}'_{(r)}\right\}, \qquad (3.6)$$

where $r = 1, \ldots, \ell$, $\boldsymbol{\theta}_{(r)}$ denotes the vector of θ_i with the r^{th} component deleted, and $\mathbf{v}$ and $\mathbf{v}^*$ are defined in Theorem 3. Then, the joint conditional posterior density of $(\boldsymbol{\beta},\, \sigma^2)$ is

$$\pi\left(\boldsymbol{\beta},\, \sigma^2 \,|\, \boldsymbol{\theta},\, \hat{\boldsymbol{\theta}},\, z = r\right) \;\propto\; \pi(\boldsymbol{\beta},\, \sigma^2) \prod_{i=1}^{\ell} \text{Normal}_{\hat{\theta}_i}\left(\mathbf{x}'_i\boldsymbol{\beta},\, \frac{\sigma^2}{\lambda_i}\right)$$
$$\times\; \text{Normal}_{\boldsymbol{\theta}_{(r)}}\left\{\hat{\hat{\boldsymbol{\theta}}}_{(r)} - \left(\sum_{j=1}^{\ell} \hat{\hat{\theta}}_j - a\right)\mathbf{v}^*_{(r)},\, \text{diagonal}\left(\hat{\boldsymbol{\sigma}}^2_{(r)}\right) - \mathbf{v}_{(r)}\mathbf{v}'_{(r)}\right\}. \qquad (3.7)$$

It is straight forward to sample the cpd's of $\boldsymbol{\theta}$ and z. However, it is not so simple to sample the cpd's of $\boldsymbol{\beta}$ and σ^2 that we will next discuss.

First, to obtain the cpd's of $\boldsymbol{\beta}$ and σ^2, we define

$$\mathbf{g}_1 = \frac{1}{\sigma^2}\sum_{i=1}^{\ell} \lambda_i \hat{\theta}_i \mathbf{x}_i \quad \text{and} \quad A_1 = \frac{1}{\sigma^2}\sum_{i=1}^{\ell} \lambda_i \mathbf{x}_i \mathbf{x}'_i.$$

Second, for $i = 1, \ldots, \ell$, let $d_i = \lambda_i\hat{\theta}_i - \left(\sum_{j=1}^{\ell} \lambda_j \hat{\theta}_j - a\right) v_i^*$ and $\mathbf{x}_i = (1 - \lambda_i)\,\mathbf{x}_i - v_i^* \sum_{j=1}^{\ell} (1 - \lambda_j)\,\mathbf{x}_j$. Let $\mathbf{d}_{(r)}$ and $\tilde{X}_{(r)}$ respectively denote the vector with entries d_i excluding the r^{th} component and the matrix with columns $\mathbf{x}_i$ excluding $\mathbf{x}_r$. Now, let

$$\mathbf{g}_2 = \left(\boldsymbol{\theta}_{(r)} - \mathbf{d}_{(r)}\right)' \Sigma_{(r)} \tilde{X}'_{(r)} \quad \text{and} \quad A_2 = \tilde{X}_{(r)} \Sigma_{(r)} \tilde{X}'_{(r)},$$

where $\Sigma_{(r)}$ is the covariance matrix without the r^{th} row and column. Third, with a multivariate normal prior on $\boldsymbol{\beta}$ of the form $\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\beta}_0,\, \Sigma_0)$, where $\boldsymbol{\beta}_0$ and Σ_0 are specified, we let

$$\mathbf{g}_0 = \boldsymbol{\beta}_0 A_0 \quad \text{and} \quad A_0 = \Sigma_0^{-1};$$

thereby offering some protection against posterior impropriety. It follows that

$$\boldsymbol{\beta} \,|\, \boldsymbol{\theta},\, z = r,\, \sigma^2,\, \hat{\boldsymbol{\theta}} \sim \text{Normal}\left\{\left(\sum_{s=0}^{2} A_s\right)^{-1} \sum_{s=0}^{2} A_s \mathbf{g}_s,\, \left(\sum_{s=0}^{2} A_s\right)^{-1}\right\}.$$

We can eliminate the prior of $\boldsymbol{\beta}$ by letting $\Sigma_0 \to \infty$ (i.e., noninformative prior) to get $\pi(\boldsymbol{\beta}) = 1$ as in the BFH model.

Finally, we consider the cpd of σ^2. Let

$$U_i = \lambda_i\hat{\theta}_i + (1 - \lambda_i)\,\mathbf{x}'_i\boldsymbol{\beta} - \left\{\sum_{j=1}^{\ell}\left\{\lambda_j\hat{\theta}_j + (1 - \lambda_j)\,\mathbf{x}'_j\boldsymbol{\beta}\right\} - a\right\} v_i^*,\; i = 1, \ldots, \ell,$$

and

$$\Sigma = \sigma^2\left\{\text{diagonal}\,(1 - \lambda_1, \ldots, 1 - \lambda_\ell) - \left[(1 - \lambda_i)(1 - \lambda_{i'})\right] \Big/ \sum_{j=1}^{\ell}(1 - \lambda_j)\right\}.$$

Then, the cpd of σ^2 is

$$\pi\left(\sigma^2 \,|\, \boldsymbol{\theta},\, z = r,\, \boldsymbol{\beta},\, \hat{\boldsymbol{\theta}}\right) \propto \pi(\sigma^2)\left[\prod_{i=1}^{\ell} \text{Normal}_{\hat{\theta}_i}\left\{\mathbf{x}_i'\boldsymbol{\beta},\, \sigma^2 + s_i^2\right\}\right] \text{Normal}_{\boldsymbol{\theta}_{(r)}}\left\{\mathbf{U}_{(r)},\, \Sigma_{(r)}\right\},$$

where $\mathbf{U}_{(r)}$ denotes the vector of the U_i excluding the r^{th} component and $\pi(\sigma^2)$ is a prior on σ^2. As in the BFH model (see Section 2), we assign the prior density $\pi(\sigma^2) = 1/(1+\sigma^2)^2$, $\sigma^2 > 0$ to σ^2. Because the baseline BFH posterior density is proper, the constraint BFH posterior density will also be proper.

4 Empirical studies

The purposes of these empirical studies are twofold. First, it is demonstrated that the BFH model can be fit as stated in Section 2 and the deleting the last one benchmarking and random benchmarking methods are performed. Second, the benchmarking methods are compared in a simulation study that uses a well-used dataset in the small area literature.

In the data generation process, we use the data on corn and soybean acres in Battese, Harter and Fuller (1988), available for 12 counties (areas) in Iowa. The resulting county-level corn and soybean acreages are constructed using a number of segments sampled from the population (known number of segments). Landsat satellite data on the number of pixels of corn and soybean in the sampled segments (i.e., two covariates) are also available. The finite population means of the number of pixels classified as corn and soybean for each county are also reported. Starting with this dataset, we construct new datasets with any number of areas.

The data generation process has two steps. In the first step, the unit-level model $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + e_{ij}$, $i = 1, \ldots, \ell,\ j = 1, \ldots, n_i$, where $e_{ij} \overset{\text{iid}}{\sim} (0,\, \sigma^2)$, is fit to the data available for the $\ell = 12$ counties in Iowa. The area sample sizes are $n_1 = n_2 = n_3 = 1$, $n_4 = 2$, $n_5 = n_6 = n_7 = n_8 = 3$, $n_9 = 4$, $n_{10} = n_{11} = 5$, and $n_{12} = 6$. Using least squares, we estimate $\boldsymbol{\beta}$ and σ^2 by $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, respectively. For the areas with sample size greater than one, we set s_i^2 equal to the estimated variance of the sample mean $\bar{y}_i$ $\left(\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i\right)$ and we let S^2 be their geometric mean. For the areas with sample size equal to one, we set s_i^2 equal to S^2. The vector of covariates $\bar{\mathbf{X}}_i$ has three elements, the integer one (for the intercept), followed by the population means of pixels classified as corn and soybean.

In the second step, the data generation process for any desired number ℓ of small areas is illustrated. The covariates $\mathbf{x}_i$, $i = 1, \ldots, \ell$, are sampled with replacement from $\bar{\mathbf{X}}_i$, $i = 1, \ldots, 12$. Then, the area-level means are drawn using

$$\theta_i \overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}},\, \hat{\sigma}^2\right),\ i = 1, \ldots, \ell,$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the least squares estimates defined above. The sample variances s_i^2 are generated in two steps. First, the sample sizes are drawn from a uniform distribution, $n_i \overset{\text{iid}}{\sim} \text{Uniform}\,(5,\, 25)$, $i = 1, \ldots, \ell$. Second, let $s_i^2 = S^2 V_i/(n_i - 1)$, where $V_i \overset{\text{ind}}{\sim} \chi^2_{n_i - 1}$ and S^2 defined above. Finally, the small area survey estimates are drawn using $\hat{\theta}_i \overset{\text{ind}}{\sim} \text{Normal}\,(\theta_i,\, s_i^2)$, $i = 1, \ldots, \ell$. The benchmarking target is set equal to the sum of the $\hat{\theta}_i$ and variants of this value, $\sum_{i=1}^{\ell} \hat{\theta}_i$ scaled up or down by 50%. In NASS's practice,

for crop county estimates, this target is an already set state value. To evaluate the benchmarking methods in extreme cases, we consider additional simulation scenarios, where an area sample size is set to 2 or 50, or where the factor S^2 is multiplied by ten.

In what follows, we report empirical results mostly for a simulation scenario using 12 areas. Examples using larger number of areas are briefly discussed. For example, Iowa has 99 counties, and one of NASS's interests is in benchmarking county estimates for planted acres, harvested acres and production (bushels) to the predefined state-level total. For such small numbers of areas, no adjustment is needed to the benchmarking procedures, deleting the last one or random deletion, introduced in the previous sections. However, the computation may be intolerable for an extremely large number of areas (say, one million), and some adjustments would be needed to the current procedures.

It is pertinent to discuss the computations for the simulation scenario with 12 areas. For posterior inference under the BFH model, we have used 1,000 random draws, and this runs in just a few seconds. On the other hand, it is more difficult to run a Gibbs sampler for deleting one at a time or random deletion benchmarking. However, we have provided an efficient Gibbs sampler as follows. We used a long run of 20,000 iterations, with a "burn in" of the first 10,000 iterations, choosing every tenth iterate thereafter. This was obtained by trial and error that is gauged by the autocorrelations, the Geweke test for stationarity and the effective sample sizes. For the 1,000 selected iterations, the autocorrelations are all negligible. For random deletion benchmarking, the p-values of the Geweke test for the three regression coefficients and δ^2 are, respectively, 0.651, 0.087, 0.828 and 0.699 (i.e., stationarity is not rejected), and the effective sample sizes are all 1,000. Also, the trace plots show no evidence of nonstationarity. Therefore, the Gibbs sampler is efficient, taking a few seconds despite the large number of runs.

The performance of benchmarking methods is assessed using a set of metrics that include posterior means (PM) and posterior standard deviations (PSD), and when it is convenient, posterior coefficients of variation (PCV), numerical standard errors (NSE) of the estimates and 95% highest posterior density intervals (95% HPD). Numerical results are presented in Tables 4.1-4.8.

A summarized version of the basic results is presented in Table 4.1, and serves for comparison of the average, standard error and coefficient of variation of the observed data with the PMs, PSDs, PCVs from the BFH model, benchmarking (deleting the last one, LO) model and random benchmarking (RD) model. The results in Table 4.1 apply to two simulation scenarios, where $S^2 = 163,$ small variation in the observed data, and where $S^2 = 1{,}630,$ relatively larger variation in the observed data. When $S^2 = 163,$ there are very little differences between the observed data and the posterior quantities from the BFH, LO and RD models. Given the small coefficients of variation for the survey estimates, it is difficult for any model to further reduce variability. Hence, the PCVs are comparable to the CVs of the survey estimates. On the other hand, three interesting points can be made for the scenario where $S^2 = 1{,}630.$ First, the PMs under the BFH model can be very different from those of LO and RD models and these latter two PMs are very close. Second, the PSDs are much smaller than the standard errors of the observed data; there are substantial gains in precision under the BFH model. However, the PSDs are about four to five times smaller than those for the observed data and the PSDS under the LO and RD model are about twice those of the BFH model. The

PCVs follow the same pattern. Third, LO and RD are very close in all three measures (PMs, PSDs, PCVs) with RD model having just slightly smaller PSDs. As expected, there is small difference between the LO model and the RD model. But one must also observe that benchmarking the BFH model is important because we can get answers that are different from the BFH model at least in terms of posterior standard deviations and coefficients of variation. Benchmarking is a jittering procedure, which helps to protect the model from misspecification, and therefore it must lead to increased variability in the small area estimates.

Table 4.1
Comparison of BFH model with no benchmarking, deleting the last one benchmarking and random benchmarking via posterior mean (PM), posterior standard deviation (PSD) and posterior coefficient of variation (PCV) for two values of S^2

		PM				PSD				PCV			
	A	OB	BFH	LO	RD	OB	BFH	LO	RD	OB	BFH	LO	RD
a. $S^2 = 163$; $a = 1{,}435$	1	135.6	134.0	133.8	133.5	6.03	5.62	5.47	5.41	0.044	0.042	0.041	0.041
	2	102.0	103.5	103.1	103.0	7.10	6.50	6.11	5.82	0.070	0.063	0.059	0.057
	3	117.7	121.0	120.7	120.5	7.31	6.72	6.55	6.25	0.062	0.056	0.054	0.052
	4	77.0	81.5	81.4	81.0	5.88	6.00	5.46	5.53	0.076	0.074	0.067	0.068
	5	126.9	127.8	127.5	127.5	5.63	5.25	5.25	5.06	0.044	0.041	0.041	0.040
	6	113.1	113.4	112.9	113.1	8.06	7.15	6.82	6.74	0.071	0.063	0.060	0.060
	7	137.2	133.7	133.5	133.9	6.74	6.38	5.93	6.02	0.049	0.048	0.044	0.045
	8	124.8	124.7	124.7	124.7	4.03	3.91	3.83	3.76	0.032	0.031	0.031	0.030
	9	118.3	116.5	115.8	116.6	7.54	6.79	6.29	6.65	0.064	0.058	0.054	0.057
	10	156.5	153.4	153.3	153.3	4.37	4.45	4.12	4.18	0.028	0.029	0.027	0.027
	11	109.5	110.3	110.3	110.2	4.88	4.64	4.70	4.70	0.045	0.042	0.043	0.043
	12	116.3	118.1	117.9	117.7	7.23	6.62	6.26	6.00	0.062	0.056	0.053	0.051
b. $S^2 = 1{,}630$; $a = 1{,}482$	1	129.1	129.8	127.2	126.5	19.07	4.64	10.71	10.45	0.148	0.036	0.084	0.083
	2	117.3	126.3	122.1	122.1	22.46	5.08	12.73	12.51	0.191	0.040	0.104	0.102
	3	120.0	145.5	137.3	136.9	23.11	5.93	12.91	12.68	0.193	0.041	0.094	0.093
	4	68.8	107.3	94.0	93.6	18.60	7.47	12.04	11.86	0.270	0.070	0.128	0.127
	5	142.4	146.4	142.3	142.2	17.80	4.52	11.98	11.15	0.125	0.031	0.084	0.078
	6	108.8	120.2	115.2	115.4	25.49	5.43	11.75	11.66	0.234	0.045	0.102	0.101
	7	136.8	116.2	118.2	119.0	21.31	5.37	11.32	11.90	0.156	0.046	0.096	0.100
	8	124.5	132.5	127.3	127.3	12.76	4.39	9.00	8.91	0.102	0.033	0.071	0.070
	9	144.2	127.5	128.0	129.5	23.86	5.33	12.74	14.00	0.165	0.042	0.100	0.108
	10	172.9	129.2	145.5	145.3	13.81	9.23	10.28	10.37	0.080	0.071	0.071	0.071
	11	109.1	114.7	110.6	110.2	15.42	4.31	10.53	10.43	0.141	0.038	0.095	0.095
	12	108.4	120.3	114.6	114.2	22.87	5.10	12.42	12.01	0.211	0.042	0.108	0.105

Note: OB: observed data; BFH: Bayesian Fay-Herriot model; LO: benchmarking (deleting the last one) model; RD: random benchmarking model; a is the target. For OB, the direct estimate, standard error and coefficient of variation are presented under PM, PSD and PCV, respectively. Under the DGSM benchmarking procedure, at $S^2 = 163$, the benchmarking values are 133.7, 103.3, 120.8, 81.3, 127.6, 113.2, 133.4, 124.5, 116.3, 153.1, 110.1, 117.9, and at $S^2 = 1{,}630$, the benchmarking values are 126.9, 123.5, 142.3, 105.0, 143.2, 117.5, 113.6, 129.5, 124.7, 126.4, 112.1, 117.6.

Under the basic simulation scenario, we compare the deletion benchmarking methods to one of the methods in DGSM that provides benchmarked posterior estimates without deletion. To match the notation in DGSM, the benchmarking equation must be rewritten as

$$\sum_{i=1}^{\ell} \omega_i \theta_i = \frac{a}{\ell} = t,$$

where $\omega_i = 1/\ell$, $\sum_{i=1}^{\ell} \omega_i = 1$. Let $\hat{\theta}_i^{(B)}$ denote the posterior means from the BFH model. Now, define $\bar{\hat{\theta}}_B = \sum_{i=1}^{\ell} \omega_i \hat{\theta}_i^{(B)}$,

$$\phi_i = \frac{\omega_i}{\hat{\theta}_i^{(B)}},\ r_i = \frac{\omega_i}{\phi_i},\ i = 1, \ldots, \ell,$$

and $S^* = \sum_{i=1}^{\ell} \omega_i^2 / \phi_i$. Note that among the several specifications in DGSM, we have selected ϕ_i at random (no preference). Then, the benchmarked Bayes estimators of DGSM are

$$\hat{\theta}_i^{(BM)} = \hat{\theta}_i^{(B)} + \left(t - \bar{\hat{\theta}}_B\right) r_i / S^*,\ i = 1, \ldots, \ell.$$

Empirical results using the estimator $\hat{\theta}_i^{(BM)}$ are presented in the note to Table 4.1. The largest difference between the benchmarked estimates under different benchmarking methods is for area 10 (OB: 172.9; BFH: 129.2; LO: 145.5; RD: 145.3; DGSM: 126.4). In general, the PMs from LO and RD are closer to OB (observed data). Otherwise, these estimates compare reasonably well with the LO benchmarking and RD deletion although there are some small differences; DGSM does not provide posterior standard deviations and credible intervals.

More detailed results for $S^2 = 163$, are presented in Tables 4.2-4.8 and in Figures 4.1-4.4. Our interest is mainly to compare deletion of a single area (e.g., LO) and RD.

Using the results in Table 4.2, we conclude that the PMs from the BFH model (without benchmarking) are slightly different from the direct estimates, and as expected, larger than the smaller direct estimates and smaller than the larger ones. Except for two areas, as expected, the PSDs are smaller than the direct standard deviations. For example, the smallest direct estimate (76.997) has the largest shrinkage with a larger standard deviation (5.881 vs. 5.995); the results are consistent with the standard shrinkage that occurs in small area estimation. We note that the PCVs are all small and the NSEs are reasonably small, too.

Table 4.2
Comparison of the direct estimator with posterior inference from the Bayesian Fay-Herriot model for the area parameters

Area	n	$\hat{\theta}$	s	PM	PSD	PCV	NSE	95% HPD
1	5	135.575	6.031	133.985	5.617	0.042	0.057	(123.422, 145.402)
2	7	101.980	7.101	103.461	6.498	0.063	0.065	(90.598, 116.134)
3	24	117.655	7.309	121.006	6.716	0.056	0.066	(107.730, 134.124)
4	23	76.997	5.881	81.473	5.995	0.074	0.058	(69.046, 92.578)
5	21	126.917	5.629	127.832	5.248	0.041	0.052	(117.850, 138.406)
6	9	113.132	8.061	113.393	7.147	0.063	0.068	(99.441, 127.451)
7	5	137.236	6.739	133.661	6.378	0.048	0.064	(121.771, 146.662)
8	20	124.839	4.034	124.732	3.906	0.031	0.039	(117.233, 132.309)
9	16	118.306	7.544	116.479	6.785	0.058	0.071	(103.225, 130.003)
10	9	156.503	4.368	153.355	4.449	0.029	0.045	(144.785, 162.031)
11	23	109.546	4.877	110.348	4.637	0.042	0.047	(101.179, 119.294)
12	9	116.314	7.232	118.098	6.623	0.056	0.068	(105.135, 131.186)

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean, PSD is the posterior standard deviation and HPD is highest posterior density interval. NSE is the numerical standard errors of the posterior means. The benchmarking value is 1,435 and the sum of the posterior mean is 1,437.823 (not benchmarked).

The estimates from the BFH model with deleting the last area and with random deletion under a uniform prior (equal weights) are presented in Tables 4.3 and 4.4. The posterior weights barely differ from 0.083

with the largest one (0.097) of the last area and smallest one (0.056) of the 8^{th} area. Both random deletion and deleting the last one provide improved precision, as the PSDs of the benchmarked estimates are all smaller than the observed standard errors, for both benchmarking methods. The NSEs are larger than for no benchmarking, but this barely matters as these are errors of the PMs (the characteristic of the PM has three digits).

Table 4.3
Comparison of the direct estimator with posterior inference from the Bayesian Fay-Herriot model for the area parameters under random deletion benchmarking

Area	n	$\hat{\theta}$	s	PM	PSD	PCV	NSE	95% HPD
1	5	135.575	6.031	133.516	5.431	0.041	0.171	(123.414, 143.541)
2	7	101.980	7.101	102.903	5.793	0.056	0.199	(92.378, 114.250)
3	24	117.655	7.309	120.671	6.237	0.052	0.194	(107.744, 132.190)
4	23	76.997	5.881	81.170	5.597	0.069	0.202	(69.781, 91.177)
5	21	126.917	5.629	127.652	5.036	0.039	0.170	(118.293, 137.228)
6	9	113.132	8.061	112.805	6.707	0.059	0.223	(100.926, 126.074)
7	5	137.236	6.739	133.908	6.007	0.045	0.177	(122.135, 145.344)
8	20	124.839	4.034	124.703	3.757	0.030	0.120	(117.962, 132.304)
9	16	118.306	7.544	116.451	6.650	0.057	0.249	(103.400, 129.316)
10	9	156.503	4.368	153.222	4.216	0.028	0.134	(144.392, 160.854)
11	23	109.546	4.877	110.221	4.694	0.043	0.150	(101.038, 119.570)
12	9	116.314	7.232	117.780	5.997	0.051	0.208	(104.619, 128.158)

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean, PSD is the posterior standard deviation and HPD is highest posterior density interval. NSE is the numerical standard errors of the posterior means. The benchmarking value is 1,435. Under a uniform prior (equal weights) the posterior probabilities that the areas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are deleted are respectively 0.090, 0.084, 0.095, 0.077, 0.066, 0.093, 0.097, 0.056, 0.098, 0.068, 0.079, 0.097.

Table 4.4
Comparison of the direct estimator with posterior inference from the Bayesian Fay-Herriot model for the area parameters under deleting the last area

Area	n	$\hat{\theta}$	s	PM	PSD	PCV	NSE	95% HPD
1	5	135.575	6.031	133.772	5.519	0.041	0.151	(122.213, 143.991)
2	7	101.980	7.101	103.026	6.319	0.061	0.171	(89.424, 113.857)
3	24	117.655	7.309	120.470	6.458	0.054	0.209	(108.783, 134.261)
4	23	76.997	5.881	81.391	5.906	0.073	0.171	(69.636, 92.634)
5	21	126.917	5.629	127.883	5.158	0.040	0.142	(117.282, 137.305)
6	9	113.132	8.061	112.895	6.270	0.056	0.216	(100.664, 124.320)
7	5	137.236	6.739	133.298	5.948	0.045	0.178	(121.831, 144.727)
8	20	124.839	4.034	124.664	3.810	0.031	0.124	(117.321, 131.941)
9	16	118.306	7.544	116.542	6.531	0.056	0.203	(104.238, 129.622)
10	9	156.503	4.368	153.229	4.353	0.028	0.132	(144.443, 161.593)
11	23	109.546	4.877	109.997	4.563	0.041	0.168	(101.428, 118.953)
12	9	116.314	7.232	117.835	6.344	0.054	0.215	(106.421, 131.483)

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean, PSD is the posterior standard deviation and HPD is highest posterior density interval. NSE is the numerical standard errors of the posterior means. The benchmarking value is 1,435.

The three methods (BFH, RD, LO) are compared using the results in Table 4.5. The PMs are comparable, so that benchmarking (RD, LO) does not distort (shrink) the estimates much beyond the shrinkage under the BFH model. Also, the PSDs under LO and RD are almost always smaller than those under the BFH model. For eight of the twelve areas, RD has smaller PSDs than LO; in these areas, RD shows roughly 1% decrease in PSD over LO and roughly 4% over the PSDs from BFH.

To investigate how sensitive the PSDs are to different benchmarking targets, we present results using three choices of targets in Table 4.6. The PSDs change only slightly over different targets and are still better than the standard errors of the direct estimates.

As part of designing a complex set of simulations, we consider using unequal probabilities (weights) in the random deletion benchmarking, and present results in Table 4.7. Uniform weights (EW) are compared to weights inversely proportional (IW) to the sample sizes and to weights directly proportional (DW) to the samples sizes. Again, small differences are present among the three PMs and among the three PSDs. The PSDs are still smaller than those of the direct estimates.

Using the results in Table 4.8, we study how extreme sample sizes in the last county (to be deleted) affect posterior inference. For this, we set the sample size of the last county to be outside the simulation range (5-25), at 2 and 50. First, consider the case in which the sample size of the last county is 2. Consistent with previous findings, there are minor differences of the PMs over no benchmarking, deleting the last one and random deletion for all counties. The PSDs for LO and RD are smaller than those of BFH with nine of these PSDs for RD smaller than LO. However, for the last county, we observe relatively large posterior standard deviations (10.00, 8.771, 8.525), roughly 15% decrease in PSD of RD over no benchmarking. Next, consider the case in which the sample size of the last county is 50. The patterns are similar, except the PSDs for the last county are comparable to the others under BFH, LO and RD and again there is an approximately 10% decrease (6.282, 5.958, 5.702) in PSD of RD over no benchmarking. It appears that deliberately putting the county with the most extreme sample size (small or large) as the last county can affect the benchmarking procedure. In contrast, minor changes are observed when the areas with extreme sample size are not systematically deleted. When the sample size is 2, the new PMs and PSDs are the following, BFH: 124.307, 9.993; LO: 123.371, 9.000 RD: 123.540, 8.887. When the sample size is 50, the new PMs and PSDs are the following, BFH: 118.167, 6.284; LO: 117.802, 6.094; RD: 117.716, 5.948.

Table 4.5
A summary of the comparison of inference from the direct estimator, the Bayesian Fay-Herriot (BFH) model, random deletion (RD) benchmarking and deleting the last one (LO)

				BFH		RD		LO	
Area	n	$\hat{\theta}$	s	PM	PSD	PM	PSD	PM	PSD
1	5	135.575	6.031	133.985	5.617	133.516	5.431	133.772	5.519
2	7	101.980	7.101	103.461	6.498	102.903	5.793	103.026	6.319
3	24	117.655	7.309	121.006	6.716	120.671	6.237	120.470	6.458
4	23	76.997	5.881	81.473	5.995	81.170	5.597	81.391	5.906
5	21	126.917	5.629	127.832	5.248	127.652	5.036	127.883	5.158
6	9	113.132	8.061	113.393	7.147	112.805	6.707	112.895	6.270
7	5	137.236	6.739	133.661	6.378	133.908	6.007	133.298	5.948
8	20	124.839	4.034	124.732	3.906	124.703	3.757	124.664	3.810
9	16	118.306	7.544	116.479	6.785	116.451	6.650	116.542	6.531
10	9	156.503	4.368	153.355	4.449	153.222	4.216	153.229	4.353
11	23	109.546	4.877	110.348	4.637	110.221	4.694	109.997	4.563
12	9	116.314	7.232	118.098	6.623	117.780	5.997	117.835	6.344

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean and PSD is the posterior standard deviation. The benchmarking value is 1,435. Under a uniform prior, the posterior probabilities that the areas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are deleted are respectively 0.090, 0.084, 0.095, 0.077, 0.066, 0.093, 0.097, 0.056, 0.098, 0.068, 0.079, 0.097.

Table 4.6
Comparison of posterior inference of the area parameters under random deletion benchmarking with different targets (a = 1,435)

				a		1.5a		0.5a	
Area	n	$\hat{\theta}$	s	PM	PSD	PM	PSD	PM	PSD
1	5	135.575	6.031	133.516	5.431	189.249	5.385	77.769	5.561
2	7	101.980	7.101	102.903	5.793	175.963	5.794	29.847	5.899
3	24	117.655	7.309	120.671	6.237	197.219	6.099	44.145	6.461
4	23	76.997	5.881	81.170	5.597	134.628	5.871	27.771	5.460
5	21	126.917	5.629	127.652	5.036	177.209	5.165	78.125	5.053
6	9	113.132	8.061	112.805	6.707	201.949	7.145	23.614	6.995
7	5	137.236	6.739	133.908	6.007	200.989	6.018	66.781	6.024
8	20	124.839	4.034	124.703	3.757	151.951	3.952	97.484	3.924
9	16	118.306	7.544	116.451	6.650	196.849	6.990	35.990	6.607
10	9	156.503	4.368	153.222	4.216	184.720	4.019	121.708	4.706
11	23	109.546	4.877	110.221	4.694	148.724	4.966	71.752	4.760
12	9	116.314	7.232	117.780	5.997	193.050	5.954	42.514	6.081

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean and PSD is the posterior standard deviation. The benchmarking value is 1,435. Under a uniform prior, the posterior probabilities that the areas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are deleted are respectively 0.090, 0.084, 0.095, 0.077, 0.066, 0.093, 0.097, 0.056, 0.098, 0.068, 0.079, 0.097. When the benchmarking value is increased by 50%, these probabilities are 0.090, 0.084, 0.095, 0.079, 0.064, 0.093, 0.097, 0.056, 0.098, 0.068, 0.079, 0.097. When the benchmarking value is decreased by 50%, these probabilities are 0.090, 0.084, 0.095, 0.077, 0.066, 0.093, 0.097, 0.057, 0.097, 0.068, 0.079, 0.097.

Table 4.7
Comparison of posterior inference of the area parameters under random deletion benchmarking with equal weights (EW), weights inversely proportional sample sizes (IW) and weights directly proportional to sample sizes (DW)

				EW		IW		DW	
Area	n	$\hat{\theta}$	s	PM	PSD	PM	PSD	PM	PSD
1	5	135.575	6.031	133.516	5.431	133.508	5.518	133.436	5.404
2	7	101.980	7.101	102.903	5.793	103.042	5.737	103.049	5.809
3	24	117.655	7.309	120.671	6.237	120.529	6.176	120.634	6.247
4	23	76.997	5.881	81.170	5.597	81.167	5.571	81.111	5.567
5	21	126.917	5.629	127.652	5.036	127.669	5.079	127.541	5.055
6	9	113.132	8.061	112.805	6.707	112.762	6.704	113.074	6.716
7	5	137.236	6.739	133.908	6.007	133.965	5.968	133.798	6.027
8	20	124.839	4.034	124.703	3.757	124.829	3.734	124.719	3.757
9	16	118.306	7.544	116.451	6.650	116.300	6.707	116.502	6.640
10	9	156.503	4.368	153.222	4.216	153.238	4.198	153.204	4.220
11	23	109.546	4.877	110.221	4.694	110.190	4.697	110.208	4.690
12	9	116.314	7.232	117.780	5.997	117.802	6.010	117.726	5.989

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean and PSD is the posterior standard deviation. The benchmarking value is 1,435. Under a uniform prior, the posterior probabilities that the areas 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are deleted are respectively 0.090, 0.084, 0.095, 0.077, 0.066, 0.093, 0.097, 0.056, 0.098, 0.068, 0.079, 0.097. When the benchmarking is done using weights inversely proportional to sample sizes, these probabilities are 0.167, 0.127, 0.039, 0.037, 0.026, 0.105, 0.184, 0.030, 0.061, 0.078, 0.039, 0.107. When the benchmarking is done using weights directly proportional to sample sizes, these probabilities are 0.032, 0.048, 0.168, 0.124, 0.103, 0.061, 0.036, 0.083, 0.112, 0.044, 0.123, 0.066.

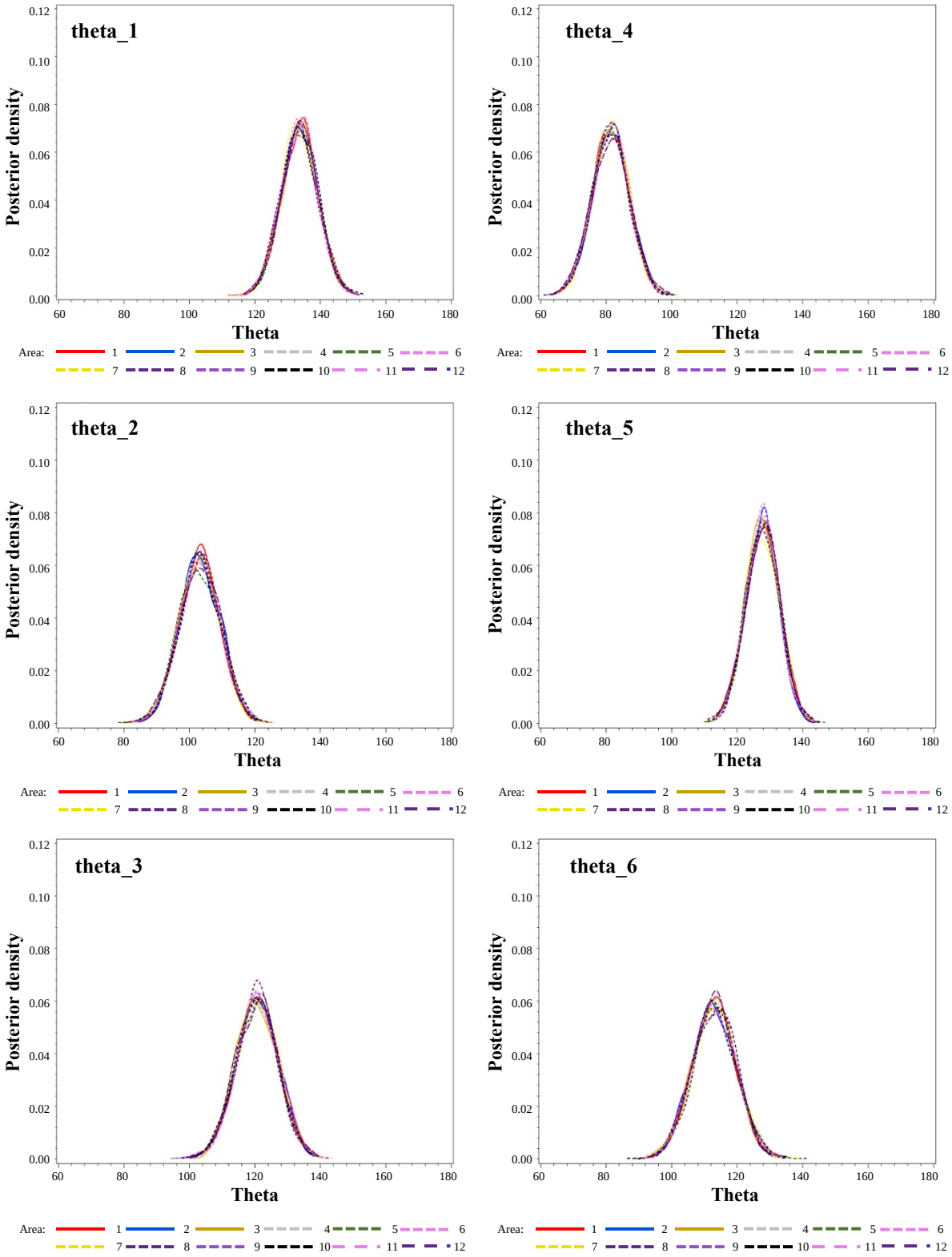
For comparison, different posterior densities are presented in Figures 4.1-4.4. In Figures 4.1 and 4.2, we present posterior densities of all twelve area parameters when each area, in turn, is deleted. We observe that the posterior densities are slightly different around the modes, but nothing remarkable. In Figures 4.3 and 4.4, we present posterior densities of all twelve area parameters under the FH model (unconstrained), random deletion benchmarking and deleting the last one. There are some differences among the three densities, but again these are not alarmingly different.

Finally, empirical results are presented for a simulation scenario with 99 areas, reflecting the 99 counties in Iowa. The data are generated as previously described, and the BFH model without benchmarking, with random deletion benchmarking, and with deleting the last one benchmarking is fit using 20,000 iterations for the Gibbs sampler. For each model fit, the first 10,000 iterations are used as a burn-in and every tenth iteration is kept thereafter. The BFH model fitting takes 15 seconds, while the deletion benchmarking models takes slightly less than three minutes each. For the random deletion benchmarking model parameters, the regression coefficients $\boldsymbol{\beta}$ and the variance σ^2, the p-values of the Geweke test are, respectively, 0.822, 0.128, 0.752 and 0.219, and the effective sample sizes are all 1,000 for the 1,000 selected iterations (i.e., an efficient Gibbs sampler). Note that the target is 12,162.93 and the sum of the PMs from the BFH model is 12,168.49, a difference of 5.56. In Figure 4.5, we present a plot of the coefficients of variation under random deletion benchmarking, deleting the last one benchmarking and BFH model versus the direct estimates by area. The differences among these models are not remarkable. Most of the points with direct CVs larger than about 0.04 fall below the 45^o straight line. However, some points (diamond) under the BFH model are above the 45^o line, four of them are noticeable, possibly shrinking too much. We conclude that it is sensible to perform the random deletion benchmarking.

Table 4.8
A summary of the comparison of inference from the direct estimator, the Bayesian Fay-Herriot (BFH) model, deleting the last one (LO) and random deletion (RD) benchmarking when the last county is extreme

					BFH		LO		RD	
	Area	n	$\hat{\theta}$	s	PM	PSD	PM	PSD	PM	PSD
a. The last county size is 2.	1	5	135.575	6.031	134.116	5.607	133.772	5.473	133.510	5.409
	2	7	101.980	7.101	103.205	6.482	102.818	6.118	102.745	5.837
	3	24	117.655	7.309	121.110	6.730	120.911	6.577	120.666	6.260
	4	23	76.997	5.881	81.586	6.021	81.741	5.544	81.196	5.631
	5	21	126.917	5.629	127.901	5.252	127.552	5.264	127.619	5.041
	6	9	113.132	8.061	113.454	7.147	112.889	6.818	113.074	6.815
	7	5	137.236	6.739	133.938	6.339	133.479	5.968	133.947	5.994
	8	20	124.839	4.034	124.753	3.906	124.699	3.824	124.738	3.735
	9	16	118.306	7.544	116.199	6.806	115.329	6.327	116.065	6.785
	10	9	156.503	4.368	153.419	4.434	153.148	4.174	153.240	4.213
	11	23	109.546	4.877	110.512	4.645	110.473	4.696	110.324	4.686
	12	2	121.881	12.75	124.243	10.00	123.755	8.771	123.444	8.525
b. The last county size is 50.	1	5	135.575	6.031	133.984	5.618	133.745	5.461	133.452	5.385
	2	7	101.980	7.101	103.462	6.499	103.136	6.086	103.044	5.780
	3	24	117.655	7.309	121.006	6.716	120.832	6.536	120.698	6.232
	4	23	76.997	5.881	81.473	5.995	81.596	5.512	81.162	5.728
	5	21	126.917	5.629	127.832	5.248	127.519	5.238	127.661	5.001
	6	9	113.132	8.061	113.393	7.146	112.929	6.777	112.899	6.675
	7	5	137.236	6.739	133.659	6.380	133.351	5.947	133.851	5.941
	8	20	124.839	4.034	124.732	3.906	124.713	3.821	124.726	3.825
	9	16	118.306	7.544	116.480	6.785	115.766	6.269	116.319	6.601
	10	9	156.503	4.368	153.355	4.449	153.225	4.173	153.306	4.230
	11	23	109.546	4.877	110.347	4.637	110.378	4.692	110.155	4.689
	12	50	116.538	6.791	118.117	6.282	118.035	5.958	117.952	5.702

Note: n is the area sample size, $\hat{\theta}$ is the direct estimator and s its standard error. PM is the posterior mean and PSD is the posterior standard deviation. When the sample size of the last county is 50 (2), the benchmarking value is 1,435 (1,441). The uniform prior is used in the random benchmarking.



Figures 4.1 Comparison of the posterior densities for θ_1 to θ_6 when each area is deleted at a time (e.g., the first area is deleted in the first panel etc.).

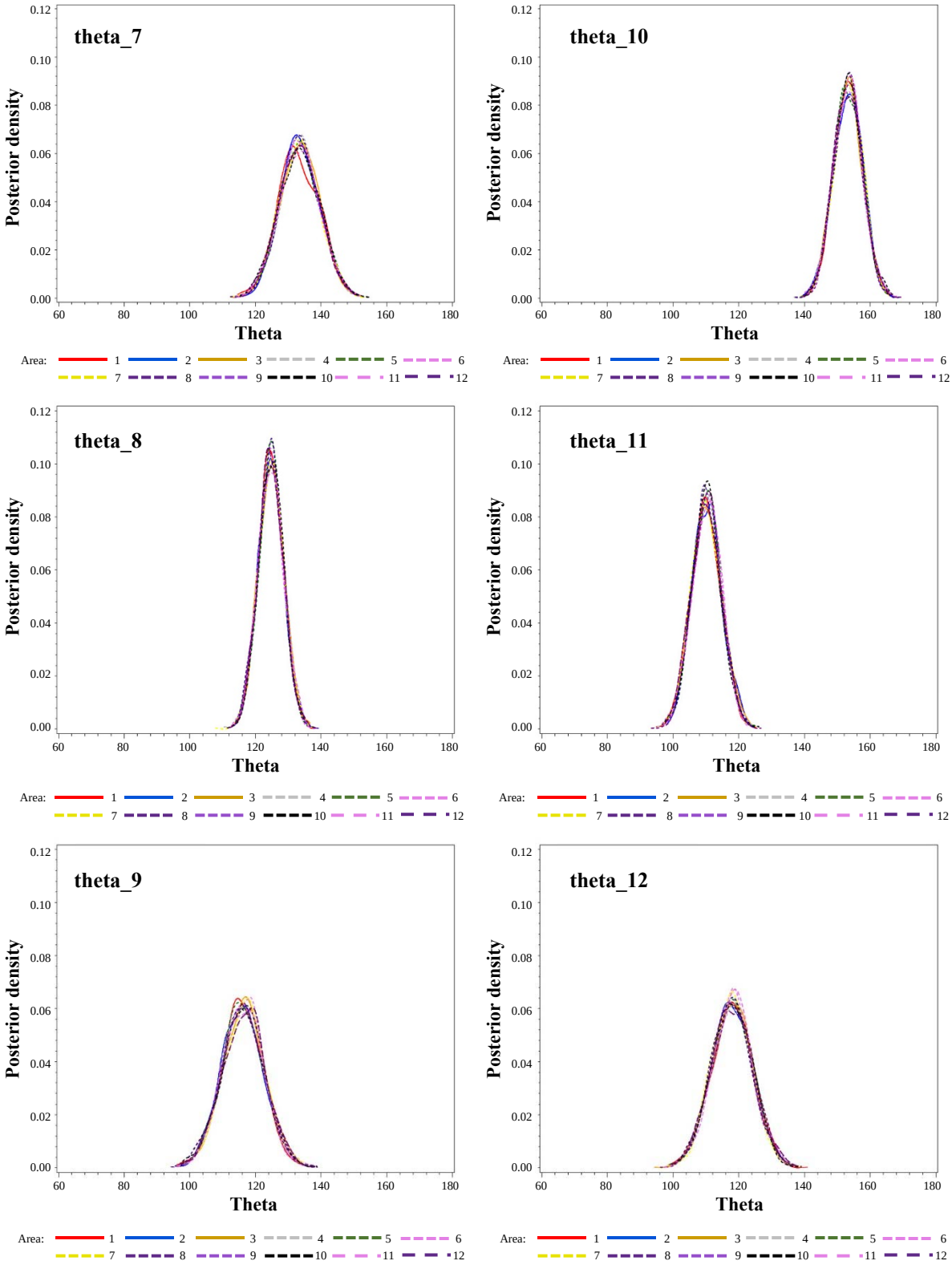


Figure 4.2 Comparison of the posterior densities for θ_7 to θ_{12} when each area is deleted at a time.

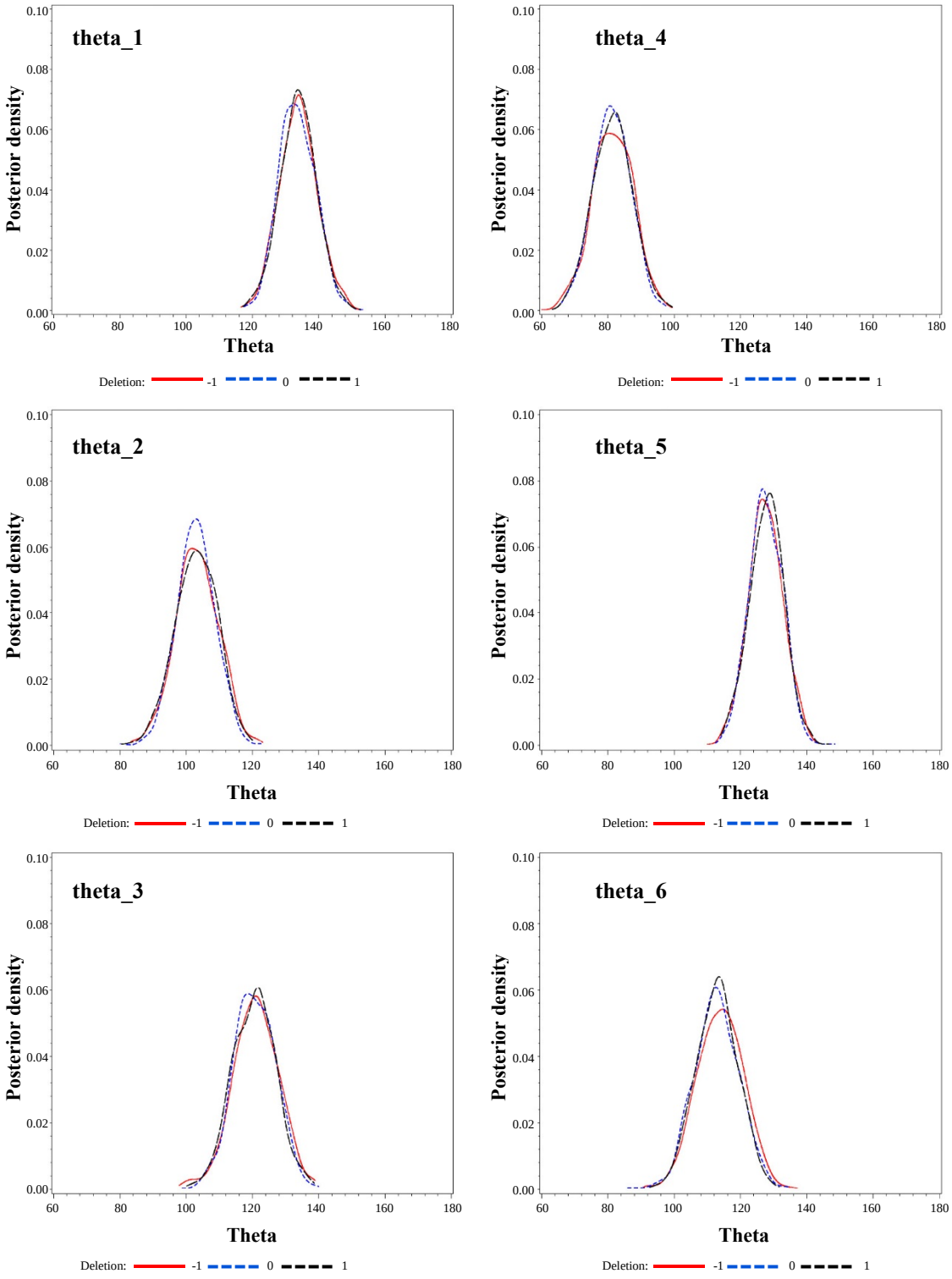


Figure 4.3 Comparison of the posterior densities for θ_1 to θ_6 under the Fay-Herriot model (-1), random deletion benchmarking (0) and area-12 deletion.

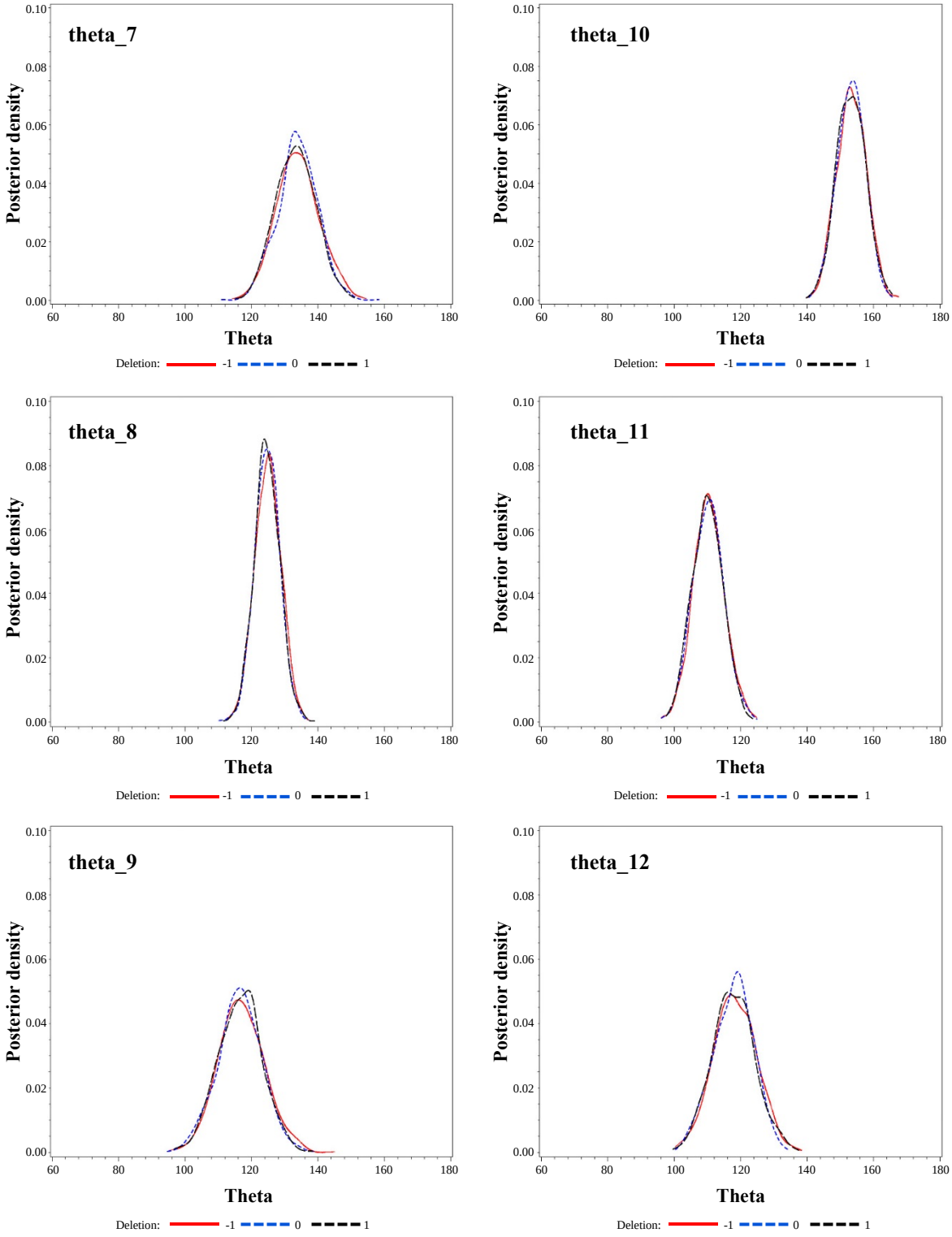


Figure 4.4 Comparison of the posterior densities of θ_7 to θ_{12} under the Fay-Herriot model (-1), random deletion benchmarking (0) and area-12 deletion.

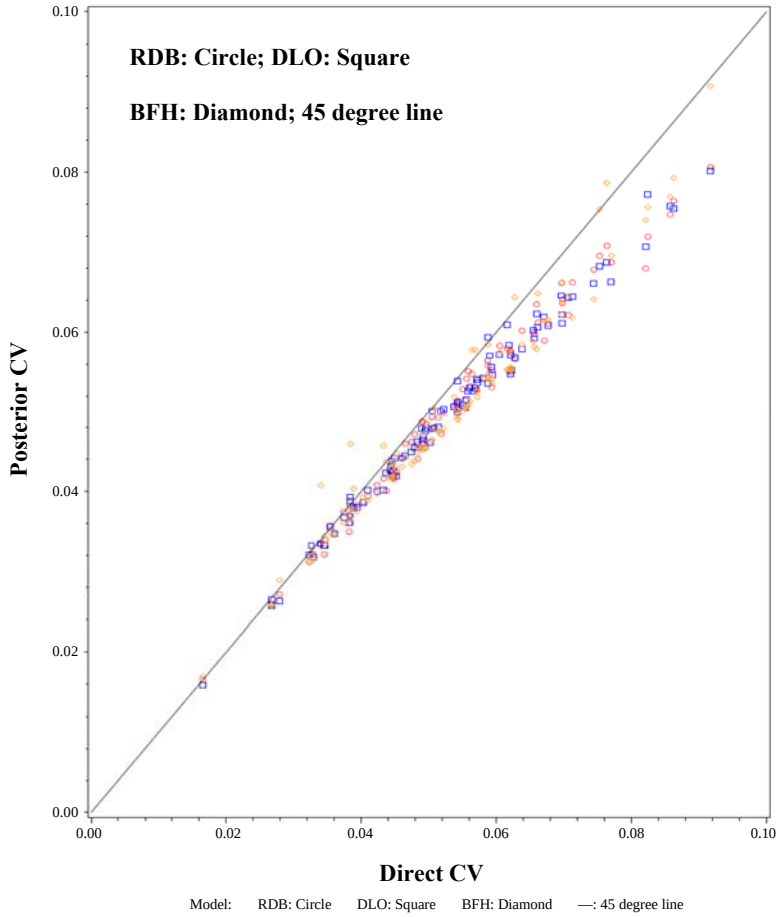


Figure 4.5 Plot of the coefficients of variation under the random deletion benchmarking, deleting the last one and the Bayesian Fay-Herriot model for 99 areas.

5 Concluding remarks

The Bayesian Fay-Herriot (BFH) model is discussed in detail. We show that the BFH can be fit using random samples rather than a Markov chain Monte Carlo sampler. Since random samples required no monitoring, this method is beneficial because there is little time at NASS between receiving the county-level survey summary data and presenting the final estimates. In support to the BFH model, we show that the posterior density under the BFH model is proper, providing a baseline for benchmarking. The effects of benchmarking are studied in a simulation study, comparing the BFH model without benchmarking to the BFH model with two benchmarking methods.

In this study, we assume that the benchmarking constraint is of the form $\sum_{i=1}^{\ell} \theta_i = a$. A straightforward generalization of the benchmarking methods may be developed for the constraint of the form $\sum_{i=1}^{\ell} w_i \theta_i = a$, where the w_i are weights. For example, this latter situation occurs for benchmarking yield, ratio of production and harvested acres.

Our major contribution is the extension of BFH model to accommodate benchmarking. Previous approaches delete the last area, giving rise to the question “Does it matter which area is deleted?”. In this paper, we develop and illustrate a method that gives each area a chance to be deleted. We show how to fit this extended BFH model using the Gibbs sampler. Because of the complexity of the joint posterior density, a sampling based method, without Markov chains, cannot be used. Using empirical studies, we show that the differences in the posterior means over no benchmarking, deleting the last county and random deletion are very small.

The effects of changing the benchmarking target are studied in a sensitivity analysis. As expected, changing the benchmarking target leads to different estimates, but, unexpectedly, the changes in the posterior standard deviations are small. Small changes in the estimates are noted for the benchmarking methods using different probabilities of deletion.

It is expected that the posterior standard deviations from deleting the last one benchmarking and random benchmarking be larger than those from the BFH model because of the jittering effect from benchmarking. However, in the empirical studies we present, deleting the last one benchmarking and random benchmarking have about the same posterior standard deviations with a small reduction when random benchmarking is used. The key strength of the random benchmarking approach is that there is no preferential treatment for any area/county.

Disclaimer and acknowledgements

The Findings and Conclusions in This Preliminary Publication Have Not Been Formally Disseminated by the U.S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy. This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agriculture Statistics Service.

Dr. Nandram’s work was supported by a grant from the Simons Foundation (353953, Balgobin Nandram). The authors thank the Associate Editor and the referees for their comments and suggestions. The work of Erciulescu was completed as a Research Associate at the National Institute of Statistical Sciences (NISS) working on NASS projects.

Appendix A

Exemplification of the sensitivity of deletion

Let $y_i \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, \sigma_i^2)$, $i = 1, 2$, such that $y_1 + y_2 = a$, and $\lambda = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. Then, if we start by deleting y_2, the joint density of (y_1, y_2) is

$$f(y_1, y_2 \,|\, \phi = 0) = \delta_{y2}(a - y_1)\,\text{Normal}\left\{\lambda_{\mu_1} + (1-\lambda)(a - \mu_2), (1-\lambda)\sigma_2^2\right\},$$

where $\delta_a(b) = 1$ if $a = b$ and $\delta_a(b) = 0$ if $a \neq b$. However, if we start by deleting y_1, the joint density of (y_1, y_2) is

$$g(y_1, y_2 \,|\, \phi = 0) = \delta_{y1}(a - y_2)\,\text{Normal}\{\lambda(a - \mu_1) + (1-\lambda)\mu_2, (1-\lambda)\sigma_2^2\}.$$

It will matter in the estimation procedure which variable is deleted because the two joint distributions are different. Note that the two distributions are the same if and only if

$$\lambda\mu_1 + (1-\lambda)(a - \mu_2) = \lambda(a - \mu_1) + (1-\lambda)\mu_2,$$

which gives

$$\lambda(\mu_1 - a/2) = (1-\lambda)(\mu_2 - a/2).$$

Even if we assume that $\sigma_1^2 = \sigma_2^2$, the two distributions are different. However, under this assumption, $\lambda = 1/2$, and the condition for the two distributions to be the same is that $\mu_1 = \mu_2$. That is, overall the condition for the two joint distributions to be the same is that $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, thereby making y_1 and y_2 exchangeable. However, this is a very restricted situation.

One way out of this diffculty is to actually delete both y_1 and y_2 in the following way. Let $z = 1$ if y_1 is deleted and let $z = 0$ if y_2 is deleted. Then,

$$p(y_1, y_2, z \,|\, \phi = 0) = [pg(y_1, y_2 \,|\, \phi = 0)]^z\,[(1-p)\,f(y_1, y_2 \,|\, \phi = 0)]^{1-z},$$

where we have taken $z \sim \text{Bernoulli}(p)$ and, because z is not really identifiable, we will take $p = 1/2$ (i.e., we randomly delete one or the other). However, note that

$$z \,|\, y_1, y_2, \phi = 0 \sim \text{Bernoulli}\left\{\frac{pg(y_1, y_2 \,|\, \phi = 0)}{[pg(y_1, y_2 \,|\, \phi = 0)] + [(1-p)\,f(y_1, y_2 \,|\, \phi = 0)]}\right\}.$$

Appendix B

Fitting the Bayesian Fay-Herriot model

The Bayesian Fay-Herriot (BFH) model is given in (2.1) and the joint posterior density under the BFH model is given in (2.3), which for convenience we state here,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}) \propto \frac{1}{(1+\sigma^2)^2}\left(\frac{1}{\sigma^2}\right)^{\ell/2}\prod_{i=1}^{\ell}\left\{\exp\left[-\frac{1}{2}\left\{\frac{1}{s_i^2}\left(\hat{\theta}_i - \theta_i\right)^2 + \frac{1}{\sigma^2}\left(\theta_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right]\right\}. \quad \text{(B.1)}$$

We show how to fit the joint posterior density of the parameters using random samples (not even a Gibbs sampler), thereby avoiding any monitoring. We will use the multiplication rule to write

$$\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \hat{\boldsymbol{\theta}}) = \pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma^2, \hat{\boldsymbol{\theta}})\,\pi_2(\boldsymbol{\beta} \,|\, \sigma^2, \hat{\boldsymbol{\theta}})\,\pi_3(\sigma^2 \,|\, \hat{\boldsymbol{\theta}}),$$

where $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma^2, \hat{\boldsymbol{\theta}})$ and $\pi_2(\boldsymbol{\beta} \,|\, \sigma^2, \hat{\boldsymbol{\theta}})$ have standard forms and $\pi_3(\sigma^2 \,|\, \hat{\boldsymbol{\theta}})$ is nonstandard but it is density of a single parameter.

Momentarily, we will drop the term, $\frac{1}{(1+\sigma^2)^2}\left(\frac{1}{\sigma^2}\right)^{\ell/2}$, because it only affects the posterior density of σ^2. That is,

$$\pi_1\left(\boldsymbol{\theta}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\hat{\boldsymbol{\theta}}\right) \propto \prod_{i=1}^{\ell}\left\{\exp\left[-\frac{1}{2}\left\{\frac{1}{s_i^2}\left(\hat{\theta}_i-\theta_i\right)^2+\frac{1}{\sigma^2}\left(\theta_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right]\right\}.$$

Standard calculations reduce the argument (without $-1/2$) of the exponential term to

$$\frac{1}{(1-\lambda_i)\,\sigma^2}\left\{\theta_i-\left(\lambda_i\hat{\theta}_i+(1-\lambda_i)\,\mathbf{x}_i'\boldsymbol{\beta}\right)\right\}^2+\frac{\lambda_i}{\sigma^2}\left(\hat{\theta}_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2,\ \ \lambda_i=\frac{\sigma^2}{s_i^2+\sigma^2},\ i=1,\ldots,\ell.$$

Hence, for $\pi_1\left(\boldsymbol{\theta}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\hat{\boldsymbol{\theta}}\right)$,

$$\theta_i\,|\,\boldsymbol{\beta},\,\sigma^2,\,\hat{\boldsymbol{\theta}} \overset{\text{ind}}{\sim} \text{Normal}\left\{\lambda_i\hat{\theta}_i+(1-\lambda_i)\,\mathbf{x}_i'\boldsymbol{\beta},\,(1-\lambda_i)\,\sigma^2\right\},\ i=1,\ldots,\ell. \tag{B.2}$$

Momentarily, we will drop the term, $\prod_{i=1}^{\ell}[(1-\lambda_i)\,\sigma^2]^{1/2}$. Then, integrating out the θ_i, we get

$$\pi_2\left(\boldsymbol{\beta}\,|\,\sigma^2,\,\hat{\boldsymbol{\theta}}\right) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{\ell}\frac{\lambda_i}{\sigma^2}\left(\hat{\theta}_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}.$$

Hence, the exponent (without $-1/2$) can be written as,

$$\sum_{i=1}^{\ell}\frac{\lambda_i}{\sigma^2}\left(\hat{\theta}_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\hat{\Sigma}^{-1}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right),$$

where

$$\hat{\boldsymbol{\beta}}=\hat{\Sigma}\sum_{i=1}^{\ell}\frac{\hat{\theta}_i\mathbf{x}_i}{s_i^2+\sigma^2} \quad \text{and} \quad \hat{\Sigma}^{-1}=\sum_{i=1}^{\ell}\frac{\mathbf{x}_i\mathbf{x}_i'}{s_i^2+\sigma^2}.$$

It is worth noting that $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$ are well defined for all σ^2 provided that the design matrix, X, where $X'=(\mathbf{x}_1,\ldots,\mathbf{x}_\ell)$ is full rank. Then,

$$\pi_2\left(\boldsymbol{\beta}\,|\,\sigma^2,\,\hat{\boldsymbol{\theta}}\right) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{\ell}\frac{\lambda_i}{\sigma^2}\left(\hat{\theta}_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2-\frac{1}{2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\hat{\Sigma}^{-1}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\}.$$

That is,

$$\boldsymbol{\beta}\,|\,\sigma^2,\,\hat{\boldsymbol{\theta}} \sim \text{Normal}\left(\hat{\boldsymbol{\beta}},\,\hat{\Sigma}\right). \tag{B.3}$$

Now, integrating out $\boldsymbol{\beta}$ and incorporating the terms in σ^2, which were dropped, we have

$$\pi_3\left(\sigma^2\,|\,\hat{\boldsymbol{\theta}}\right) \propto Q(\sigma^2)\frac{1}{\left(1+\sigma^2\right)^2}, \tag{B.4}$$

where

$$Q(\sigma^2)=\left|\hat{\Sigma}\right|^{1/2}\prod_{i=1}^{\ell}\frac{1}{\left(s_i^2+\sigma^2\right)^{1/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{\ell}\frac{1}{s_i^2+\sigma^2}\left(\hat{\theta}_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2\right\}.$$

To obtain a random sample from (B.1), we sample σ^2 from (B.4), $\boldsymbol{\beta}$ from (B.3) and the θ_i independently from (B.2). The conditional posterior density in (B.4) is nonstandard, and to draw a sample from it, we use a grid method (e.g., Nandram and Yin, 2016). First, we transform σ^2 to $\phi = \sigma^2/(1+\sigma^2)$ so that $0 < \phi < 1$. Then, we divide (0, 1) into 100 grids. Actually, we have located the range of ϕ in (0, 1) and we have divided this interval into 100 grids. This gives us a probability mass function that we sample. Jittering is used in the selected grid to get deviates, which are different with probability one; see Nandram and Yin (2016) for more details.

Appendix C

Proof of Theorem 2

It is convenient to make the following transformations,

$$\theta_i = \theta_i,\ i = 1, \ldots, \ell - 1,\ \phi = \sum_{i=1}^{\ell} \theta_i - a.$$

Here, ϕ is a dummy variable, which holds the benchmarking constraint, and it ensures a non-singular transformation. The Jacobian is unity and the inverse transformation is

$$\theta_i = \theta_i,\ i = 1, \ldots, \ell - 1,\ \theta_\ell = \phi + a - \sum_{i=1}^{\ell-1} \theta_i.$$

The transformed density is

$$\tilde{\pi}(\theta_1, \ldots, \theta_{\ell-1}, \phi).$$

Then, the density that holds the benchmarking constraint exactly is

$$\tilde{\pi}(\theta_1, \ldots, \theta_{\ell-1} \,|\, \phi = 0),\ \theta_\ell = a - \sum_{i=1}^{\ell-1} \theta_i.$$

Therefore,

$$\pi(\theta_1, \ldots, \theta_{\ell-1} \,|\, \phi = 0) \propto \exp\left\{-\frac{1}{2\delta^2}\left[\sum_{i=1}^{\ell-1}\left(\theta_i - \mathbf{u}_i'\boldsymbol{\beta}\right)^2 + \left\{\sum_{i=1}^{\ell-1}\theta_i - \left(a - \mathbf{u}_i'\boldsymbol{\beta}\right)\right\}^2\right]\right\}.$$

Dropping terms that do not involve $\boldsymbol{\theta}_{(\ell)} = (\theta_1, \ldots, \theta_{\ell-1})'$, it is easy to show that the exponent is

$$\frac{1}{2\delta^2}\left\{\boldsymbol{\theta}_{(\ell)}'(I+J)\boldsymbol{\theta}_{(\ell)} - 2\left[(\mathbf{u}_1 - \mathbf{u}_\ell)'\boldsymbol{\beta}, \ldots, (\mathbf{u}_{\ell-1} - \mathbf{u}_\ell)'\boldsymbol{\beta} + a\mathbf{j}'\right]\boldsymbol{\theta}_{(\ell)}\right\}.$$

Then, using the properties of a multivariate normal density, we have

$$\boldsymbol{\theta}_{(\ell)} \,|\, \phi = 0 \ \sim\ \text{Normal}\left((I+J)^{-1}\left(a\mathbf{j}' + (\mathbf{u}_1 - \mathbf{u}_\ell)'\boldsymbol{\beta}, \ldots, (\mathbf{u}_{\ell-1} - \mathbf{u}_\ell)'\boldsymbol{\beta}\right)', \delta^2(I+J)^{-1}\right).$$

Finally, using the Sherman-Morrison formula, $(I + J)^{-1} = I - \frac{1}{\ell} J$, we have

$$\boldsymbol{\theta}_{(\ell)} \,|\, \phi = 0 \;\sim\; \text{Normal}\left\{\left(I - \frac{1}{\ell} J\right)\left(a\mathbf{j}' + (\mathbf{u}_1 - \mathbf{u}_\ell)' \boldsymbol{\beta}, \ldots, (\mathbf{u}_{\ell-1} - \mathbf{u}_\ell)' \boldsymbol{\beta}\right)', \delta^2\left(I - \frac{1}{\ell} J\right)\right\}.$$

It is worth noting that the matrix determinant lemma gives $\det(I + J) = \ell$ and so $\det\left(\delta^2 (I - \frac{1}{\ell} J)\right) = \frac{1}{\ell}(\delta^2)^{\ell-1}$.

References

Battese, G., Harter, R. and Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28-36. doi:10.2307/2288915.

Bell, W.R., Datta, G.S. and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100(1), 189-202.

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

Datta, G.S., Ghosh, M., Steorts, R. and Maples, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20(3), 574-588.

Erciulescu, A.L., Cruze, N.B. and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of Royal Statistical Society, Series A*, 182, 283-303. doi:10.1111/rssa.12390.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366), 269-277.

Ghosh, M., and Steorts, R. (2013). Two stage Bayesian benchmarking as applied to small area estimation. *Test*, 22(4), 670-687.

Janicki, R., and Vesper, A. (2017). Benchmarking techniques for reconciling Bayesian small area models at distinct geographic levels. *Statistical Methods and Applications*, 26, 4, 557-581.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1), 1-96.

Nandram, B., and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37, 2, 137-152. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11603-eng.pdf.

Nandram, B., and Toto, M.C.S. (2010). Bayesian predictive inference for benchmarking crop production for Iowa counties. *Journal of the Indian Society of Agricultural Statistics*, 64 (2), 191-207.

Nandram, B., and Yin, J. (2016). A nonparametric Bayesian prediction interval for a finite population mean. *Journal of Statistical Computation and Simulation*, 86 (16), 3141-3157.

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

Nandram, B., Toto, M.C.S. and Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81 (11), 1593-1608.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Pfeffermann, D., Sikov, A. and Tiller, R. (2014). Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test*, 23(4), 631-666.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101 (476), 1387-1397.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, Wiley Series in Survey Methodology.

Toto, M.C.S., and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference*, 140, 2963-2979.

Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 1, 29-36. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10619-eng.pdf.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. and Dick, J.P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census under coverage estimation. *Statistics in Transition*, 6, 631-640.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 34, No. 4, December 2018

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 35, No. 1, March 2019

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 46, No. 3, September/septembre 2018

Original Articles

CONTENTS TABLE DES MATIÈRES

Volume 46, No. 4, December/décembre 2018

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.