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

Survey Methodology

Survey Methodology
48-2

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

Email at infostats@statcan.gc.ca

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

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

Standards of service to the public

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

Note of appreciation

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2022

•

Volume 48

•

Number 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 48, Number 2, December 2022

Contents

Waksberg Invited Paper Series

Special discussion paper

Regular papers

Waksberg Invited Paper Series

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

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2024 Waksberg Award.

This issue of *Survey Methodology* opens with the 22th paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Jean Opsomer (Chair), Jack Gambino, Maria Giovanna Ranalli and Elizabeth Stuart for having selected Roderick J. Little as the author of 2022 Waksberg Award paper.

2022 Waksberg Invited Paper

Author: Roderick J. Little

Roderick Little has published widely on methods for the analysis of data with missing values and model-based survey inference, and the application of statistics to diverse scientific areas, including medicine, demography, economics, psychiatry, aging and the environment. He is Richard D. Remington Distinguished University Professor of Biostatistics at the University of Michigan, where he also holds appointments in the Department of Statistics and the Institute for Social Research. He chaired the Biostatistics Department at Michigan for 11 years. From 2010 to 2012, Little was the inaugural Associate Director for Research and Methodology and Chief Scientist at the U.S. Bureau of the Census. Distinctions include the American Statistical Association's Founder's Award and Wilks Medal for research contributions, and the President's Invited Address and COPSS Fisher Lectureship at the Joint Statistical Meetings. Little is an elected member of the International Statistical Institute, a Fellow of the American Statistical Association and the American Academy of Arts and Sciences, and a member of the U.S. National Academy of Medicine.

Waksberg Award honorees and their invited papers since 2001

2023 Raymond **Chambers**, Manuscript in preparation expected for the December 2023 issue.

2022 Roderick **Little**,“Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference”. *Survey Methodology*, vol. 48, 2, 257-281.

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

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

2019 Chris **Skinner**.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference

Roderick J. Little[1]

Abstract

Conceptual arguments and examples are presented suggesting that the Bayesian approach to survey inference can address the many and varied challenges of survey analysis. Bayesian models that incorporate features of the complex design can yield inferences that are relevant for the specific data set obtained, but also have good repeated-sampling properties. Examples focus on the role of auxiliary variables and sampling weights, and methods for handling nonresponse. The article offers ten top reasons for favoring the Bayesian approach to survey inference.

Key Words: Calibrated Bayes inference; Design-based inference; Penalized splines; Post-stratification; Probability proportional to size sampling; Proxy pattern-mixture models; Response propensity; Super-population models; Survey weighting.

1. Introduction

Bayesian inference is in my view the best overarching inferential paradigm for statistical inference from surveys, whether from probability or non-probability samples. See for example Ericson (1969), Binder (1982), Rubin (1987), Ghosh and Meeden (1997), Little (2003ab, 2004, 2012, 2015), Sedransk (2008) and Fienberg (2011). However, design-based properties of Bayesian inferences are important, because "all models are wrong", and broad acceptance of results requires inferences that have good operating characteristics in repeated sampling. In particular, Bayesian models need to incorporate complex design features to yield inferences that are approximately calibrated, in the sense that credible intervals have close to nominal levels when treated as confidence intervals in repeated sampling (Rubin, 1984, 2019; Little, 2006). In large samples, flexible working models can avoid strong parametric assumptions that lead to potentially biased estimates.

To focus discussion, consider the problem of deriving a point estimate q of a finite population quantity $Q,$ and a 95% interval estimate $I_{0.95} = (l, u)$ that captures uncertainty in $q,$ the interval may have a frequentist interpretation as a 95% confidence interval, or a Bayesian interpretation as a 95% posterior credible interval for $Q.$ I think scientists who are not statisticians generally interpret the interval I in a Bayesian way, as a fixed interval capturing the uncertainty about $Q.$ However, I do not focus unduly on the difference in interpretation of $I_{0.95}$ under the two paradigms. The 95% nominal value is by convention and other levels could be chosen.

An appealing feature of finite population survey sampling is that it deals with real (though unknown) quantities. For "analytic" survey inference, where the focus is on parameters of idealized models of the population, such as regression coefficients in a multiple regression model, define the finite population

[1] Roderick J. Little, Department of Biostatistics, University of Michigan. E-mail: rlittle@umich.edu.

quantity Q as the estimate of the parameter of interest if the model was fitted to data for the whole population, according to some agreed fitting method such as least squares or maximum likelihood (ML). A useful feature of this construction is that Q is then a real quantity, rather than a feature of a simplified hypothetical model of the population (e.g., Little, 2004).

Under the Bayesian paradigm, inference for Q is based on its posterior predictive distribution given the data, for judicious choices of model and prior distribution for unknown parameters. Thus q may be the posterior predictive mean of $Q,$ and $I_{0.95}$ the 2.5th to 97.5 percentile of the posterior predictive distribution, or the limits of the range of values of Q with the highest posterior density, assuming the posterior predictive distribution is unimodal. A useful feature of the Bayesian approach is that "finite population corrections" are automatically incorporated in the posterior predictive distribution of finite population quantities – as the sample converges to the finite population, the posterior variance tends to zero.

The focus is on developing suitable models and prior distributions. Computation used to be a major challenge and is still a practical consideration, though less so now with the advent of Markov Chain Monte Carlo methods and rapid advances in Bayesian computation. Thus, the complaint that Bayes is conceptually appealing but simply too difficult to implement is harder to sustain than it was, say, thirty years ago.

The remainder of the paper is organized as follows. In Section 2, I introduce some notation and describe formally the models and prior distributions required by the Bayesian approach to survey data, with and without nonresponse. In Section 3, I describe generally desirable features of an inference about $Q,$ and discuss why I believe the Bayesian paradigm for suitably-chosen models can be more successful at achieving these features than the design-based approach. In Section 4 I present a variety of examples, intended to illustrate the points in Section 3. I conclude in Section 5 by proposing ten reasons to be Bayesian in the survey sampling setting.

2. Notation, and a seminal paper

In this section, I introduce some notation and a seminal paper that underlies much of the thinking in this paper. Let $Y = (y_1, \ldots, y_N)$ and $S = (S_1, \ldots, S_N)$ where $N < \infty$ is the number of units in the population, y_i is the set of survey variables and S_i is the selection indicator for the i^{th} unit, with value 1 when the i^{th} unit is selected and 0 otherwise. Let Z represent design information such as stratum or cluster indicators, and z_i the value of Z for unit $i.$ Consider inference about a finite population quantity $Q(Y, Z),$ for example the population total $Q(Y, Z) = \sum_{i=1}^{N} y_i,$ where $Y = (y_1, \ldots, y_N).$ A general model-based approach treats both S and Y as random variables, with joint distribution given Z:

$$f_{S,Y|Z}(S, Y | Z, \theta, \psi) = f_{Y|Z}(Y | Z, \theta) f_{S|Y,Z}(S | Z, Y, \psi), \tag{2.1}$$

where $f_{Y|Z}$ represents the density of survey variables Y indexed by unknown parameters $\theta,$ and $f_{S|Y,Z}$ represents the model for inclusion indexed by unknown parameters $\psi.$ For a probability sample with no nonresponse, the sampling distribution is known and does not depend on $Y,$ that is,

$$f_{S|Y,Z}\left(S\,|Z,Y,\psi\right) = f_{S|Z}\left(S\,|Z\right); \tag{2.2}$$

design-based methods base inferences on the distribution of statistics in repeated sampling from this distribution.

For a survey with unit nonresponse, inclusion occurs when a unit is selected, and then responds given selection. Accordingly, let $R_i = 1$ if selected unit i responds and $R_i = 0$ otherwise. The model-based approach models the joint distribution of $S,$ R and Y given Z as

$$f_{S,R,Y|Z}(S,R,Y\,|Z,\theta,\psi) = f_{Y|Z}\left(Y\,|Z,\theta\right) f_{S|Y,Z}\left(S\,|Z,Y,\psi\right) f_{R|S,Y,Z}\left(R\,|Z,Y,S,\phi\right), \tag{2.3}$$

adding to equation (2.1) a model for unit nonresponse with density $f_{R|S,Y,Z}.$ Item nonresponse can also be treated by modeling indicators for the patterns of item missingness (e.g., Little, 2003b).

Treating $S,$ R and Y as random variables is a key feature of Rubin (1978), which I regard as one of the landmark statistics papers in the history of statistics. The paper provides conditions under which the missingness and selection mechanisms are ignorable, that is, do not need to be modeled for likelihood-based inference, extending definitions of ignorability for missing data in Rubin (1976), while providing a framework for inference when selection and/or missingness is non-ignorable. The significance of the paper for survey sampling is easily missed, because its main focus is on the role of the treatment assignment mechanism in the context of inference about causal effects. The assignment mechanism is ignorable under random treatment assignment, as in randomized clinical trials. The paper thus lays a general framework for causal inferences comparing treatments, and it is for this feature that the paper is best known. However, the paper also provides a Bayesian justification for random sampling, as a means of avoiding the need for a model for selection.

In frequentist superpopulation modeling (e.g., Valliant, Dorfman and Royall, 2000), the parameters in models are treated as fixed; in Bayesian survey modeling, these parameters are assigned a prior distribution, and inferences for $Q(Y)$ are based on its posterior predictive distribution given the data. In large samples, the prior distribution plays a minor role, and the two approaches yield similar answers for comparable models; in particular the ML estimate of a parameter is essentially the mode of the posterior distribution under a uniform prior, and as such has a Bayesian interpretation. In small samples, uncertainty about the model parameters is propagated when they are integrated out of the posterior distribution. This approach to propagating error in parameters allows Bayesian inferences for judiciously chosen models and priors to be better calibrated than inferences from superpopulation modeling inferences, in a sense of having better frequentist properties in repeated sampling (Rubin, 1978). So, in my opinion "superpopulation modeling is super, but Bayes is better".

3. Design-based versus model-based inference

The survey sampling literature features many lively controversies (e.g., Smith, 1976, 1994; Kish, 1995; Brewer, 2013; Little, 2014) between "design-based" inference, where inference is based on the sampling distribution (2.2) and "model-based" inference, where inference is based on model distribution $f_{Y|Z}(Y|Z, \theta)$ if selection is ignorable, or on the full model distribution (2.1) or (2.3) if selection is nonignorable or there is nonresponse. I seek inferences that are both design-based and model-based, in that they are based on Bayesian models but have good design-based properties.

Rubin (1984) distinguished between *statistical inference for a particular data set*, and the *properties* of that inference – consistency, confidence coverage – in repeated sampling. To be broadly credible, the inference should also have good repeated-sampling properties. This goal of the "design-based" approach should also be a goal of Bayesian models for surveys – the model and prior should be chosen to yield inferences with good design-based properties. To achieve this, features of the complex probability sample design need to be part of the model – stratification and weighting incorporated via covariates, multistage sampling incorporated via hierarchical models. The inclusion of a prior distribution in Bayesian modeling, decried by some as yet another assumption to be added to the model, for me provides an additional tool over superpopulation modeling. It provides more flexibility than superpopulation modeling, which effectively restricts the choice to uniform priors.

In addition to having good frequentist properties, the inference based on q and I needs to be appropriate – Rubin (2019) uses the term "relevant" – for the realized data set. Let D denote the data that are the basis for the inference, and $\tilde{D}$ the particular realization of $D,$ the sample and respondent values actually obtained. Whether q and I are derived from a formal Bayesian model, an estimating equation, or some algorithmic procedure, they should provide good inference for the data $\tilde{D},$ not other data sets D that may have been obtained. Bayesian methods tend to have this property, because the posterior distribution conditions on $\tilde{D};$ but a confidence interval should also be approximately valid when viewed as a credible interval that conditions on $\tilde{D},$ if only because this is how a non-statistician tends to interpret it. Design-based confidence intervals can be lacking from this perspective, as illustrated in examples 2 and 4 below.

To summarize, a common goal of design-based and model-based inference is to arrive at a value of q that makes efficient use of the data, has some property like design consistency which implies that it is not too far from $Q,$ and an interval I that is as narrow as the information in the data allows, while including Q with a probability close to the nominal 95% value. Rubin (2019) associates these properties with a "sage" statistician in his Waksberg lecture.

Design-based methods are often rationalized as avoiding the need for a model, because properties like design consistency are not based on a model for the data. However, the performance of design-based methods often depends on an implicit model, and modifying the estimate based on a more realistic model can improve the inference, from a design-based or model-based perspective. This point is illustrated in examples 3-5 below.

The question of the appropriate reference set for repeated sampling properties like confidence intervals is fraught with difficulties, specifically on whether to condition on ancillary statistics or on statistics that are close to ancillary (e.g., Birnbaum, 1962; Berger and Wolpert, 1988; Ghosh, Reid and Fraser, 2010). These questions also arise when assessing the repeated-sampling properties of a Bayesian inference, but do not apply to the inference itself because the posterior distribution conditions on $\tilde{D}$.

The design-based approach to sample survey inference is too limited in scope, failing to address adequately many of the problems of sample surveys in practice. Limitations include the following:

1. Design-based inference is asymptotic, and does not provide valid inferences in small samples. Consider the following simple example.

   Example 1. Inference about a population mean from a simple random sample. Consider inference about a population mean of a variable Y from a simple random sample of size n from a population of size N. The standard design-based 95% confidence interval takes the form

   $$\bar{y} \pm 1.96 s\sqrt{(1/n-1/N)}, \tag{3.1}$$

   where $\bar{y}$ is the sample mean and s is the sample standard deviation. This interval is asymptotic and does not provide valid small sample inferences. In particular, if Y is continuous, better inference is usually obtained by replacing 1.96, the 97.5th percentile of the normal distribution, by the 97.5th percentile of a distribution that reflects uncertainty about estimating the variance, such as the t distribution with $n-1$ degrees of freedom. However, that procedure assumes a normal distribution for Y and hence is not design-based. If Y is binary with values 0 and 1, then $\bar{y}$ is the sample proportion, $s=\sqrt{\bar{y}(1-\bar{y})}$ and (3.1) is the asymptotic Wald interval, which performs very poorly in small samples, particularly when the true proportion is near to 0 or 1. The Bayesian credible interval for a Jeffreys or uniform prior has much better frequentist properties. See Dean and Pagano (2015) and Franco, Little, Louis and Slud (2019) for comparisons of Wald intervals with alternatives for complex designs. Design-based inference is often a poor option for small samples, and in particular for small area estimation, where a model for Y is invoked to "borrow strength" across areas.

2. Design-based inference does not handle survey unit or item nonresponse or response errors, because these problems require models to yield generally satisfactory results.

3. Design-based inference is not prescriptive, in a sense of prescribing the appropriate choice of inference method for the data at hand. The appropriate choice of estimator effectively requires an implicit model, as in “model-assisted” estimation (e.g., Särndal, Swensson and Wretman, 1992). For example, the regression or ratio estimator for incorporating auxiliary information, or the Horvitz-Thompson or Hájek estimator for incorporating survey weights, are all based on implicit models, and if that model is far from realistic these methods may be severely suboptimal – Basu’s (1971) elephants being an extreme and satirical example. Bayesian inference based on more flexible models tend to do better, as discussed in example 5 below.
4. Design-based inference does not address how to provide inferences for non-probability samples, which are increasingly prevalent given the expense and difficulty of obtaining true random samples.

Point 4 does not rule out the use of design-based methods for deriving q and I, because we can always pretend that we have a probability sample, by assuming a model for the selection indicators S that describe inclusion in the sample, and estimating the unknown parameters in that model (e.g., Elliott and Valliant, 2017). Statisticians who use design-based methods for inference from random samples tend to favor this “quasi-randomization” approach. However, it shares the limitations of the design-based approach for probability samples, namely the inability to handle small samples, missing data or response errors; the Bayesian toolkit for these problems is much more extensive.

4. Examples

A variety of examples are offered, admittedly somewhat slanted towards my own work with colleagues. I avoid topics such as small area or time series estimation, because the need for modeling is well established there.

Example 1 continued. Normal random sample. A common critique of model-based methods is “if I base an inference on a model and the model is wrong, the inference must be wrong. Because (to paraphrase George Box), all models are wrong, therefore all model-based inference is wrong. So I prefer design-based inference, which does not require modeling assumptions”. The reasoning is plausible but it’s not that simple. The validity of model-based methods depends both on the design and on the degree and nature of model misspecification; the fact that design-based methods do not overtly depend on models does not mean they are necessarily superior to methods that do.

Suppose a simple random sample of size n is taken from a continuous distribution with unknown mean μ and standard deviation σ. Consider three interval estimates for the population mean:

(i) Interval A is the standard design-based 95% confidence interval, equation (3.1).

(ii) Interval B replaces the 97.5th normal percentile in equation (3.1), namely 1.96, by the 97.5th percentile of the t distribution with $n-1$ df.

(iii) Interval C is the 95% posterior credible interval based on a normal model with the Jeffreys' prior distribution $p(\mu, \sigma) \propto 1/\sigma$.

From a frequentist perspective, which of these intervals is the best? Interval A makes no distributional assumption on Y, and B makes a normal assumption – does that mean that A is superior? If n is large then the intervals are essentially the same, and if n is small then B is arguably better than A even if the data are not normal, because it is reflecting uncertainty about the variance.

In an informal survey, two thirds of a recent class of our well-trained Ph.D. students preferred B to C, on the grounds that it avoids the choice of prior distribution and hence makes fewer assumptions. But B and C are equally good or bad, because they are the same procedure! That this interval is an exact 95% confidence interval under normality, and is a 95% credible interval for the stated choice of prior, are just two properties of the procedure. Judging a method by its overt assumptions is an over-simplification.

Example 2. Simple random sample with a lower bound on the variance (Little, 2006). Suppose in the previous example that $n = 7$, $\bar{y} = 1, s = 1$. The standard t interval (ignoring finite population corrections) is

$$\bar{y} \pm t_{6,975}\, s/\sqrt{n} = 1 \pm 2.447/\sqrt{7} = 1 \pm 0.925 \qquad (4.1)$$

if in fact we know that $\sigma =$ 1.5, a better interval is

$$\bar{y} \pm z_{0.975}\, \sigma/\sqrt{n} = 1 \pm 1.96 \times 1.5/\sqrt{7} = 1 \pm 1.11, \qquad (4.2)$$

the wider interval reflecting the fact that σ is greater than the particular value of s for this sample. The t interval (4.1) has exact confidence coverage, but, given what we know about σ, it is the wrong inference for this specific data set: we should not pick it over (4.2) because it is narrower!

Now suppose we know that $\sigma >$ 1.5, because of some unaccounted source of additional variation. The Bayesian approach then incorporates this information into the prior distribution for σ, resulting in a credible interval that is wider than (4.2). What is the frequentist answer? The t confidence interval still has exactly nominal coverage in repeated sampling, but it is clearly too narrow as an inference for the observed dataset, because it fails to reflect what is known about σ. The interval (4.2) is anticonservative, despite the fact that it is wider than (4.1) for the realized data set – a property that can happen for confidence intervals but cannot happen for credible intervals under a specific model. Asymptotic frequentist methods are no help here, so what is the alternative to Bayes in this example?

A related example arises in one group random-effects analysis of variance, when the least squares estimate of the between-group variance is negative – a Bayesian analysis addresses this with a prior distribution on the between-group variance that does not allow negative values. Random-effects and mixed-effects models are important to handle clustering in surveys, and Bayesian methods are better than ML in this setting.

Example 3. Post-stratification on a categorical covariate. Prediction (as in modeling) is a more reliable general approach to inference than weighting (as in design-based inference). I illustrate this general statement with the simple example of post-stratification on a single variable $Z.$ A more complex example – Example 7 – is given later.

Consider an equal probability sample with a single categorical post-stratifying variable $Z,$ for which known population counts N_h are available for each post-stratum $h,$ $h = 1, 2, \ldots, H.$ Let $\bar{y}_h$ be the sample mean in post-stratum $h,$ based on sample size $n_h,$ and $n = \sum_{h=1}^{H} n_h,$ $N = \sum_{h=1}^{H} N_h.$ The standard estimate of the population mean is the post-stratified mean:

$$\bar{y}_{\text{PS}} = \sum_{h=1}^{H} P_h \, \bar{y}_h, \tag{4.3}$$

where $P_h = N_h / N$ and N is the population size. This can be viewed as a weighted mean

$$\bar{y}_{\text{PS}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_i \, y_i,$$

where $w_i = N_h / (N n_h)$ is the post-stratification weight for sampled units in poststratum $h.$ These weights can be very large in post-strata with small sample sizes $n_h,$ which, unlike stratified sampling based on $Z,$ are not under the control of the sampler. These large weights can lead to excessive variability in $\bar{y}_{\text{PS}}.$ In fact, strictly speaking, $\bar{y}_{\text{PS}}$ does not have a distribution in repeated sampling, because with positive probability the sample sizes in some post-strata may be zero. This remains true if the post-strata are modified to ensure that the post-strata sample counts are all positive for the observed sample, for example by pooling adjacent strata.

The standard design-based approach to excessive variability of $\bar{y}_{\text{PS}}$ is to modify the weights $\{w_i\},$ for example by trimming the large ones. However, from a prediction perspective, this is misguided. The problem is not the weights – the population proportions $\{P_h\}$ in each post-stratum are known, after all – the problem is that sparsity of sample in some post-strata renders the estimates $\{\bar{y}_h\}$ unreliable. It is the estimates in sparse post-strata that need to be modified, not the weights $\{w_i\}$ attached to sampled units. The principled way to modify $\{\bar{y}_h\}$ is to assume a model relating Y and $Z.$ The design-based approach, by avoiding such a model, leads to the wrong principle – modifying the weights rather than the predictions of non-sampled values.

A related point: a common design-based approximation (Kish, 1992) measures the proportionate increase in variance from weighting as $1 + \text{cv}(w_i),$ where $\text{cv}(w_i)$ is the coefficient of variation of the weights; trimming the weights reduces $\text{cv}(w_i)$ and hence this proportionate increase. However, this rule of thumb is only valid when Y and Z are unrelated, in which case post-stratification is useless. If Y and Z are related and the sample size is not too small, the variance of the post-stratified mean is *smaller*, not *larger*, than the variance of the unweighted sample mean (Holt and Smith, 1979; Little and Vartivarian, 2005). The rule of thumb fails because the relationship between Y and Z is not modeled.

What is the Bayesian approach to excessive variability of $\bar{y}_{\text{PS}}$? The latter is the posterior mean for the stratified normal model

$$\left(y_{hi}\,|\,\mu_h, \sigma^2\right) \overset{\text{iid}}{\sim} G\left(\mu_h, \sigma^2\right), \tag{4.4}$$

$$p\left(\mu_h, \sigma\right) \propto 1/\sigma, \tag{4.5}$$

where $G\left(a, b^2\right)$ denotes the normal (Gaussian) distribution with mean $a,$ variance $b^2,$ and equation (4.5) is the Jeffreys' prior distribution on the mean and standard deviation in each post-stratum. Given a sparse sample in some post-strata, the prior distribution needs to be modified to allow borrowing of strength from other strata. One approach is to assume the normal random-effects model

$$p\left(\mu_h\,|\,\mu, \tau, \sigma\right) \overset{\text{iid}}{\sim} G\left(\mu, \tau^2\right),\, p\left(\mu, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.6}$$

which treats the stratum means as random effects. The variances in each post-stratum might also be treated as distinct random effects and assigned a prior distribution, rather than pooled. The posterior mean of $\bar{Y}$ for the prior distribution (4.6) moves the weight w_i of sampled units in post-stratum h towards one, with a degree of shrinkage that depends on the relative size of estimates of σ and τ (Lazzeroni and Little, 1998).

The prior distribution (4.6) makes the non-trivial assumption that the post-stratum means are exchangeable. It can be relaxed by restricting the random effects model to a subset of post-strata with small sample counts; or the constant mean μ in (4.6) might be replaced by a regression on known post-stratum characteristics $C_h,$ as in:

$$p\left(\mu_h\,|\,\beta_0, \beta_1, \tau, \sigma, c_h\right) \overset{\text{ind}}{\sim} G\left(\beta_0 + \beta_1 c_h, \tau^2\right),\, p\left(\beta_0, \beta_1, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.7}$$

which limits the exchangeability assumption to the errors in the regression of μ_h on $c_h.$ For extensive generalizations of this basic example, see Gelman and Little (1997), Elliott and Little (2000), Elliott (2007), Gelman (2007) and Si, Trangucci, Gabry and Gelman (2020).

Example 4. Regression estimator of the mean, given a population auxiliary variable. If the auxiliary variable Z in the previous example is continuous, a common way to incorporate it in the inference is via the regression estimate of the mean:

$$q = \bar{y}_{\text{REG}} = \bar{y} + \hat{\beta}_1\left(\bar{Z} - \bar{z}\right),$$

where $\hat{\beta}_1$ is the least squares estimate of the slope of Y on Z in the sample, and $\bar{z}$ and $\bar{Z}$ are respectively the sample and population mean of $Z.$ In a simulation study of five real populations, Royall and Cumberland (1981, 1985) assess inferences centered at $q,$ with (a) the standard design-based standard error based on simple random sampling, namely:

$$I_{0.95D} = \bar{y}_{\text{reg}} \pm 1.96\,\hat{\text{se}}_D\left(\hat{\beta}\right),\ \hat{\text{se}}_D\left(\hat{\beta}\right) = \sqrt{(1-f)\,s_{Y.Z}/n},$$

where $s^2_{Y.Z}$ is the sample residual variance and $f = n/N$ is the sampling fraction; and (b) the prediction standard error based on the normal linear regression model with constant variance, namely:

$$\begin{aligned} I_{0.95M} &= \bar{y}_{\text{reg}} \pm t_{0.975,\,n-2} 1.96\,\hat{\text{se}}_M\left(\hat{\beta}\right), \\ \hat{\text{se}}_M\left(\hat{\beta}\right) &= \hat{\text{se}}_D\left(\hat{\beta}\right)\sqrt{\left(1+\left(\bar{z}-\bar{Z}\right)^2\right)\Big/(1-f)\left(\textstyle\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2/n\right)}. \end{aligned}$$

The design-based confidence intervals $I_{0.95D}$ exhibit very poor conditional confidence coverage when the observed $\bar{z}$ deviates substantially from $\bar{Z}.$ The model-based confidence interval $I_{0.95M}$ takes into account this lack of balance with respect to $\bar{z},$ but is vulnerable to model misspecification, specifically lack of linearity in the relationship between Y and Z or non-constant residual variance. Robust estimates of standard error, based on the sandwich estimator or the jackknife, yield intervals with better conditional coverage properties, although still sometimes deviating from nominal coverage levels. An alternative approach is Bayesian inference based on a more flexible model relating Y to $Z,$ such as the penalized spline model:

$$\begin{aligned} &\left(y_i \mid z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\text{spline}\left(z_i, \beta\right), \sigma^2 z_i^{\alpha}\right), \\ &\text{spline}\left(z_i, \beta\right) = \beta_0 + \sum_{j=1}^{p} \beta_j z_i^j + \sum_{\ell=1}^{m} \beta_{\ell+p}\left(z_i - \kappa_\ell\right)_+^p, \\ &\left(\beta_{l+p} \mid \tau\right) \overset{\text{iid}}{\sim} N\left(0, \tau^2\right), l = 1, \ldots, m;\ p\left(\beta_0, \ldots, \beta_p, \alpha, \sigma, \tau\right) \propto 1/\sigma, 0 < \alpha < 2, \end{aligned} \tag{4.8}$$

where the constants $\kappa_1 < \ldots < \kappa_m$ are selected fixed knots, and $(u)_+^p = u^p$ if $u > 0$ and 0, otherwise (see, for example Ruppert, Wand and Carroll, 2003). The parameter α allows for a variety of common forms of heteroskedasticity. The Bayesian standard errors then reflect imbalance in distribution of Z in the sample and population, and the flexibility of the model limits bias from model misspecification.

Suppose that the target quantity is not the population mean of $Y,$ but the least squares slope of Y on Z in the population. A robust approach is to impute the non-sampled values of Y using the model (4.8), and then estimate the slope of Y on Z as the least squares slope estimated on the filled-in population data. Uncertainty can be propagated by multiple imputation (Rubin, 1987), a method founded on Bayesian ideas. In this context, Little (2004) distinguishes between the "target model" that determines the target population quantity of interest, here the linear regression of Y on $X,$ and the "working model" (4.8) that is the basis for inference, and is used to predict survey variables for the non-sampled and nonresponding units in the population. Distinguishing between these two models provides for a robust form of Bayesian survey inference.

Szpiro, Rice and Lumley (2010) apply a similar idea in a superpopulation regression setting, and Little (2019) argues that this is more straightforward than changing the interpretation of the estimand, the approach adopted by Buja, Berk, Brown, George, Pitkin, Zhan and Zhang (2019). Szpiro et al. (2010) show that the approach provides a Bayesian interpretation of the sandwich estimator of variance in regression, which is asymptotically equivalent to sample reuse estimates of variance like the bootstrap or jackknife, which are commonly applied in sample survey settings.

Example 5. Inference for samples with unequal probabilities of selection. For designs with unequal selection probabilities, classic papers critiquing the modeling approach to surveys (Kish and Frankel, 1974; Hansen, Madow and Tepping, 1983) do not include the selection probabilities in the model, yielding inferences that are vulnerable to model misspecification. The selection probabilities play an important role in robust model-based inference, but as model covariates rather than as sampling weights.

Consider, for example, inference about a population mean $\bar{Y}$. If y_i is the value of a survey variable Y and π_i is the selection probability for unit i, the usual design-based estimator of the mean of Y weights sampled units (say units $i = 1, \ldots, n$) by the inverse of π_i resulting in the Horvitz-Thompson estimate (Horvitz and Thompson, 1952)

$$\bar{y}_{\mathrm{HT}} = N^{-1} \sum_{i=1}^{n} y_i / \pi_i, \tag{4.9}$$

if the population size N is known, or the Hájek (1971) estimate

$$\bar{y}_{\mathrm{HK}} = \left(\sum_{i=1}^{n} y_i / \pi_i \right) \Bigg/ \left(\sum_{i=1}^{n} 1 / \pi_i \right), \tag{4.10}$$

if N is estimated. Weighting is a “one size fits all” approach, with sampled units receiving the same weight irrespective of the relationship between π_i and y_i. This is potentially inefficient – if the relationship is weak, weighting simply reduces the precision of the estimate without a compensating reduction in bias.

The modeling approach incorporates the selection probabilities by regressing y_i on π_i. The strength of relationship between y_i and π_i then moderates how the selection probability affects the estimator – if the relationship is weak, the regression coefficient of π_i is small and the sampling weight has little influence. This results in more efficient estimates.

A linear regression of y_i on π_i is vulnerable to bias if the linearity is misspecified, but the impact of misspecification can be reduced by choosing a model that results in a design-consistent estimate. Many models satisfy this requirement; see for example Firth and Bennett (1998).

In stratified sampling with stratifying variable Z, the natural regression model includes covariates that are dummy variables for the strata. The resulting estimate of the population mean of a continuous survey variable Y is the stratified mean, which has the same form as equation (4.3) but with Z forming strata rather than post-strata. Weighting by the inverse of the selection probability in each stratum and dummy variable regression both yield the estimator (4.3) in this situation. The Bayesian approaches to reducing variance described in example 3 tend to have less pay-off in the case of stratified sampling than for post-stratification, because the sampler has control over the sample sizes in each stratum.

In probability proportional to size (PPS) sampling, the covariate Z is the size of the unit and $\pi_i = \min\left(cz_i, 1\right)$. Now Z is a continuous variable, and weighting and regression may yield different answers. The approaches can be unified by considering models that yield the design-weighted estimates when used to predict the non-sampled units. In particular, ignoring finite population corrections, the HT estimate (4.9) is the posterior mean for the "Horvitz Thompson model":

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta z_i, \sigma^2 z_i^2\right),\ p\left(\beta, \sigma\right) \propto 1/\sigma, \tag{4.11}$$

and the Hájek estimate (4.10) is the posterior mean for the "Hájek model":

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta, \sigma^2 z_i\right),\ p\left(\beta, \sigma\right) \propto 1/\sigma. \tag{4.12}$$

These underlying models describe situations where the corresponding design-based estimates are optimal. However, they involve strong parametric assumptions. A robust Bayesian modeling approach embeds these models within a larger model, such as the penalized spline model (4.8), as proposed by Zheng and Little (2003). Zheng and Little (2005) and Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk and Thompson (2017) provide simulation studies suggesting that this modeling approach can yield substantial gains over HT or Hájek estimation, both in terms of efficiency and closer to nominal (frequentist) confidence coverage in moderate samples – remember, design-based results are asymptotic. The model (4.8) is readily expanded to include other auxiliary variables measured for all the population units, and the flexibility of small-sample inferences increased by including proper prior distributions for the model parameters.

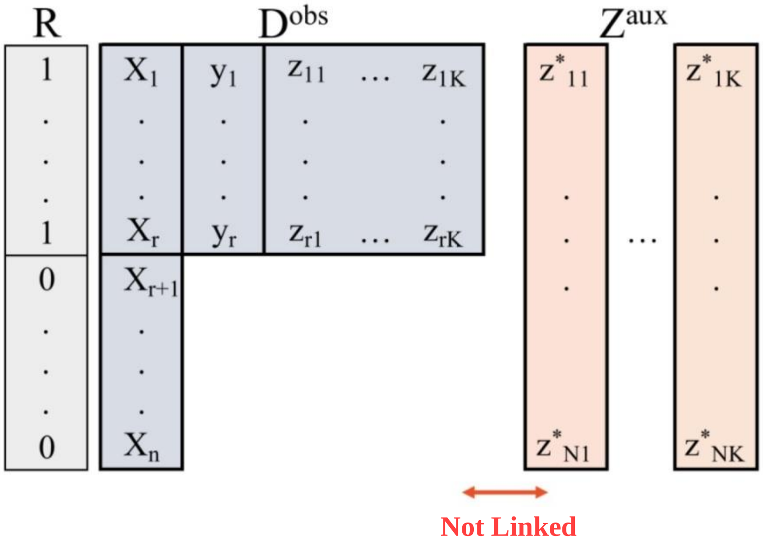
Leon-Novelo and Savitsky (2019) consider Bayesian models for the joint distribution of y_i and π_i, calling models that fix π_i a "plug-in" approach. However, assigning a distribution to π_i seems to me both unnatural and unnecessary in this setting; if π_i is recorded for all population units, it can be conditioned in the model, as in standard regression approaches that treat covariates as fixed. Even if values of π_i for non-sampled units are not provided to analysts, their values can be predicted via Bayes Theorem, as in Zangeneh and Little (2015).

Example 6. Penalized spline of response propensity models. Consider unit nonresponse on a survey variable Y, when a set of variables say $X_1, \ldots, X_p$ are observed for respondents and nonrespondents in the sample. The response propensity for unit i, $\theta_i = \Pr\left(R_i = 1 \,|\, x_{i1}, \ldots, x_{ip}, \psi\right)$, where ψ represents unknown parameters, plays an important role in both weighting and prediction approaches to survey nonresponse. In nonresponse weighting the sampling weight is multiplied by the nonresponse weight, as in

$$\begin{aligned} w_i &= 1/\Pr(\text{unit } i \text{ is selected and responds}) \\ &= (1/\Pr(i \text{ selected})) \times (1/\Pr(i \text{ responds} \mid \text{selected})) \\ & \quad \text{sampling weight} \times \text{response weight.} \end{aligned}$$

Unlike the sampling weight, the response weight is unknown, and the definition needs to be clear on what the probability is conditioned. Little (2022) argues that it should condition on auxiliary and survey variables, but not on other variables that might affect it. As with sampling weights, weighting by the inverse of the estimated response propensity is a “one size fits all” approach that does not take into account the strength of relationship between R and Y. Penalized Spline of Propensity Prediction (PSPP, Zhang and Little, 2009) regresses Y on a penalized spline of the estimated response propensity, with other variables that are observed for respondents and nonrespondents entering parametrically. The spline models a flexible relationship between the response propensity and Y, providing robustness to model misspecification. The balancing property of the propensity score also provides a double robustness property, in that the parametric form of the regression on other predictors can be misspecified without bias, provided the penalized spline captures the relationship between the propensity and Y. This method performs favorably with weighting methods in simulations (Zhang and Little, 2009; Yang and Little, 2015), and the fully Bayes version with prior distributions on the parameters propagates error in estimating the propensities.

Figure 4.1 Data structure for example 7.



Example 7. Unit nonresponse with poststratification. Another example where Bayesian and design-based approaches to inference differ concerns unit nonresponse with post-stratification. Consider the setting of Figure 4.1, where Y is a survey variable subject to nonresponse, X is a variable observed for all units in the sample, and $Z^{\text{aux}} = (Z_1, \ldots, Z_K)$ consists of a set of categorical auxiliary variables. The joint distribution of Z^{aux} is observed for survey respondents, and the marginal distributions of Z_k, $k = 1, \ldots, K,$ are also observed for the population or sample from external sources. A distinctive feature is that the units in the auxiliary data are not linked with the units in the survey. This scenario occurs frequently in settings where post-stratification is used for nonresponse adjustment. Define R_i as the response indicator for (y_i, z_i) for sample unit $i,$ and suppose that

$$\Pr\left(R_i = 1 \middle| x_i, z_i, y_i, \psi\right) = \Pr\left(R_i = 1 \middle| x_i, z_i, \psi\right), \tag{4.13}$$

so that this probability does not depend on y_i. Note that if this probability depends on $z_i,$ the response mechanism is missing not at random (MNAR) according to Rubin's (1976) definition, because values of z_i are missing for survey nonrespondents. Zangeneh and Little (2022) consider Bayes and ML estimation for data with this pattern. Considering for simplicity models that are i.i.d. over the units $i,$ the joint distribution of $X,$ $Z,$ Y and R is factored as

$$\begin{aligned} f_{X,Z,Y,R}\left(x_i, z_i, y_i, r_i \,|\, \theta, \phi\right) &= f_{Y|X,Z,R}\left(y_i \,|\, x_i, z_i, \theta, R_i = r_i\right) f_{X,Z,R}\left(x_i, z_i, r_i \,|\, \phi\right) \\ &\overset{\text{Under Eq. (4.13)}}{=} f_{Y|X,Z,R}\left(y_i \,|\, x_i, z_i, \theta\right) f_{X,Z,R}\left(x_i, z_i, r_i \,|\, \phi\right), \end{aligned}$$

where θ and ϕ are distinct parameters (Little and Rubin, 2019, Chapter 6). The parameters θ in this factorization can thus be estimated from the respondent survey data with $R_i = 1,$ and parameters of the joint distribution of X and Z are estimated by combining the respondent and auxiliary data on those variables.

Two special cases considered by Zangeneh and Little (2022) are as follows:

(1) No covariates X and a single post-stratifier Z. The missingness assumption in equation (4.13) then reduces to $\Pr\left(R_i = 1 \middle| Z_i, Y_i, \psi\right) = \Pr\left(R_i = 1 \middle| Z_i, \psi\right)$. The parameters of the conditional distribution of Y given Z are estimated from the survey respondents, and the parameters of the marginal distribution of Z are estimated from the auxiliary data on Z. In particular, if Z is multinomial with $\Pr\left(z_i = j \,|\, \phi\right) = \phi_j$ the ML estimate of ϕ_j is simply the proportion of the auxiliary counts in post-stratum j. If, in addition, Y given $Z = j$ is assumed normal with mean μ_j and variance $\sigma_j^2,$ the resulting ML estimate of the population mean of Y is simply the post-stratified mean given in equation (4.3). This simple example is interesting theoretically, because the response mechanism is MNAR but ignorable for likelihood inference; MAR is a sufficient but not always necessary condition for ignorability.

(2) X is a single categorical variable and Z is a single categorical post-stratifier. Now the joint distribution of $X,$ $Z,$ and R is not identified under a saturated model, and requires additional

assumptions to identify the model. In particular, a constrained MNAR "RAKE" model assumes that the marginal distributions of Z_1 and Z_2 are different for respondents and nonrespondents, but the odds ratios of Z_1 and Z_2 are the same for respondents and nonrespondents. This yields a just-identified model. Raking the table of respondent counts of Z_1 and Z_2 to the auxiliary margins of Z_1 and Z_2 yields ML estimates of ϕ under this RAKE model. (Little and Wu, 1991; Little, 1993). The post-stratified estimator of the mean of Y is then

$$\bar{Y}_{\text{rake}} = \sum_{j=1}^{J}\sum_{k=1}^{K} P_{jk}\left(\hat{\phi}\right)\bar{Y}_{jk}\left(\hat{\theta}\right),$$

where $P_{jk}\left(\hat{\phi}\right)$ is the proportion of the population with $X = j, Z = k$ from raking the respondent counts to the margins, and $\bar{Y}_{jk}\left(\hat{\theta}\right)$ is the estimated mean of Y given $X = j, Z = k$ from the model for Y given X and Z.

Some comments on this approach are as follows:

(1) Note again this is a MNAR model, and it is weaker than the MAR model that assumes that response depends on X but not on Z. As a result, the method tends to be less biased than alternative methods that assume MAR for consistent estimates.

(2) As was the case in example 4, the ML approach to this model does not involve modifying the weights applied to the data in cell $X = j, Z = k$ using arbitrary distance functions – the ML estimates are fully efficient under the assumed model.

(3) Sparse data in the cells can be addressed by adding proper prior distributions for the parameters and applying Bayesian methods. For example, for inference about the parameters θ of the distribution of Y given $X = j, Z = k,$ one might assume a flat prior on the main effects of X and Z but a normal prior on the interactions, shrinking them towards zero. This achieves shrinkage of the cell mean $\bar{y}_{jk}$ towards the fitted means from an additive model. In frequentist terminology, this is a mixed ANOVA model with fixed main effects and random interactions.

(4) The idea of raking to the X and Z margins is familiar, but note that if X is a stratifier and Z is a post-stratifier, the standard approach is to perform one iteration of raking, first matching the X margin and then matching the Z margin: under the assumed model, raking iteratively is the correct procedure. Raking is ML here, but Bayesian forms of raking also can be used to propagate parameter uncertainty.

Example 8. Proxy pattern-mixture analysis. Proxy pattern-mixture analysis is a method for assessing nonresponse bias for the mean of a survey variable Y subject to nonresponse, when there is a set of covariates observed for nonrespondents and respondents. Historically the amount of missing data, as measured by the response rate, has been the most often-used metric for evaluating survey quality. However, response rates ignore the information contained in auxiliary covariates observed for nonrespondents. Methods based on the estimated probability of nonresponse such as the R indicator (Schouten, Cobben and Bethlehem, 2009) and q2 measure (Särndal and Lundström, 2010) do not take into account the strength of association of the survey variable of interest and the probability of response, which arguably should be factored into the assessment of bias. These measures assume MAR, but if the auxiliary variables are available for nonresponse adjustment, it is deviations from MAR that lead to bias.

Andridge and Little (2011) propose proxy pattern-mixture analysis (PPMA), a method based on a pattern-mixture model for nonresponse that combines in a simple and intuitive way the key features of nonresponse adjustment. Let Y_i denote the value of a continuous survey outcome and $Z_i = \left(Z_{i1}, Z_{i2}, \ldots, Z_{ip}\right)$ denote the values of p covariates for unit i in the sample. Only r of the n sampled units respond, so observed data consist of $\left(Y_i, Z_i\right)$ for $i = 1, \ldots, r$ and Z_i for $i = r+1, \ldots, n.$ Let R denote the response indicator, such that for unit i $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing. To reduce dimensionality, we replace Z by a single proxy variable X that has the highest correlation with Y in the respondent sample. This proxy variable can be estimated by regressing Y on Z using the respondent data, including important predictors of $Y,$ as well as interactions and nonlinear terms where appropriate. Specifically, we assume the regression model $E\left(Y \mid Z, R=1\right) = \alpha_0 + \alpha Z,$ and let $X = \alpha Z.$ The joint distribution of $X,$ Y and R using the following proxy pattern-mixture model, similar in form to that discussed in Little (1994):

$$\left(X, Y \mid R = r\right) \sim G_2\left((\mu_x^{(r)}, \mu_y^{(r)}), \Sigma^{(r)}\right)$$

$$R \sim \text{Bernoulli}(\pi)$$

$$\Sigma^{(r)} = \begin{pmatrix} \sigma_{xx}^{(r)} & \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} \\ \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} & \sigma_{yy}^{(r)} \end{pmatrix}$$

where N_2 denotes the bivariate normal distribution. Our interest is bias in the marginal mean of $Y,$ which can be written as $\mu_y = \pi\mu_y^{(1)} + \left(1-\pi\right)\mu_y^{(0)}.$ Andridge and Little (2011) assume that

$$\Pr\left(R = 1 \middle| Y, X\right) = f\left(X^* + \lambda Y\right),$$

for some unspecified function f and known constant $\lambda.$ Here $X^* = X\sqrt{\sigma_{yy}^{(1)} / \sigma_{xx}^{(1)}}$ is the proxy X scaled to have the same variance as $Y,$ which aids the interpretation of λ by putting X and Y on the same scale. This mechanism is MAR when $\lambda = 0,$ and deviates increasingly from MAR as λ increases. With this assumption, the parameters are just identified and the ML estimate of the mean of $Y,$ averaging over patterns, is

$$\hat{\mu}_y = \bar{y}_1 + \left(\frac{n-r}{n}\right)\left(\frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}\right)\left(\bar{x}_0 - \bar{x}_1\right),$$

where $\overline{x}_1, \overline{y}_1$ are the respondent means of X and Y and $\overline{x}_0$ is the mean of X for nonrespondents. The index of bias is then the adjustment of the sample mean $\overline{y}_1$ implied by this estimate, namely

$$\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\overline{x}_0-\overline{x}_1\right), \tag{4.14}$$

where $\hat{\rho}$ is the respondent sample correlation. This adjustment incorporates three key factors that affect the potential bias in a simple and intuitive manner – the nonresponse rate $(n-r)/n$, the correlation $\hat{\rho}$ between X and Y, and the deviation of $\overline{x}_0$ from $\overline{x}_1$. In particular, the index increases with the nonresponse rate and the deviation of the mean of X for respondents and nonrespondents.

There is no information about the parameter λ in the data – this is generally the case for methods that model deviations from MAR. Following Little (1994), Andridge and Little (2011) propose a sensitivity analysis, where estimates are generated for a range of values of λ between 0 and infinity, specifically 0, 1 and infinity; the choice of 0 corresponds to MAR, the intermediate choice of 1 implies the bias of Y is the same as the bias of the proxy variable X^*, and the choice of infinity is the most extreme deviation from MAR; estimates for this case have the highest variance. As λ varies between 0 and infinity, the middle factor $(\lambda+\hat{\rho})/(\lambda\hat{\rho}+1)$ varies between $\hat{\rho}$ (when $\hat{\mu}_y$ is the standard regression estimator of the mean) and $1/\hat{\rho}$ (when $\hat{\mu}_y$ is the inverse regression estimator proposed by Brown (1990). The sensitivity of the estimate to the choice of λ is small when $\hat{\rho}$ is close to 1, that is we have a strong proxy variable, and large when $\hat{\rho}$ is close to 0, that is we have a weak proxy variable. So having auxiliary variables that are good predictors of the survey outcomes is crucial.

Bayesian versions of the proxy pattern-mixture model are readily developed, allowing for propagation of error in the model parameters. Also, the model can be applied to create multiple imputations of nonrespondent values, allowing the incorporation of complex sample design elements into the index.

More recently, the model has been used to develop indices of departure from random sampling, by applying it to model selection rather than nonresponse (Little, West, Boonstra and Hu, 2020; Boonstra, Little, West, Andridge and Alvaredo-Leiton, 2021). Refinements in this work are (a) to replace the parameter λ by $\phi = \lambda/(1+\lambda)$, a better parametrization because ϕ ranges from 0 (MAR) to 1, and (b) (with some additional assumptions) to allow missingness also to depend on auxiliary variables orthogonal to X. The method has also been extended to handle binary outcomes (Andridge, West, Little, Boonstra and Alvarado-Leiton, 2019) and indices of potential bias in regression coefficients (West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton, 2021).

5. Conclusion: Ten reasons to be Bayesian for survey inference

My examples are intended to give some idea of the richness of Bayesian modeling possible for survey data, but they are far from exhaustive. Bayes is also useful in areas I have not touched on, including multistage sampling, time series, latent class and factor analysis models, measurement error, the combination of data from multiple sources, the creation of synthetic data for disclosure avoidance, and so on. I conclude by summarizing my reasons for advocating the Bayesian approach to survey inference:

1. The design-based approach is asymptotic, and too limited to handle the varied problems of inference from surveys, whether probability or non-probability based.
2. The Bayesian approach is both unified and flexible enough to handle the various problems encountered in surveys, and it includes superpopulation modeling as a form of large-sample inference.
3. Carefully-chosen Bayesian models can yield credible intervals that have good design-based properties in repeated sampling. In particular, weighting, stratification and post-stratification can be modeled via covariates, and clustering incorporated via Bayesian hierarchical models. Flexible models that incorporate design features render hybrid approaches such as model-assisted estimation (e.g., Särndal, Swensson and Wretman, 1992) unnecessary.
4. Early critiques of the modeling approach concern models that do not incorporate design-features, and hence are vulnerable to model misspecification. Such models can and should be avoided.
5. The Bayesian calculus of integrating out nuisance parameters provides inferences that have good frequentist properties in small as well as large samples.
6. The approach of being design-based for some problems and model-based for others leads to logical inconsistencies (see, for example, Little, 2012, Section 4.3); the Bayesian approach yields inferences that are unified and logically consistent.
7. The specification of prior distributions in the Bayesian approach is a strength, not a weakness, because it provides additional modeling flexibility. For some problems, weak "objective" priors yield results that parallel standard frequentist solutions. For other problems, stronger "subjective" priors provide useful answers for models that are not identified, as in MNAR nonresponse.
8. Computational challenges in the Bayesian approach have been greatly reduced by recent methodological advances and expanded computing power.
9. Modeling puts survey research in the mainstream of statistical modeling for other types of data. The particular features of survey research – complex sampling design, and the focus on finite population quantities, are well handled by the Bayesian paradigm.
10. The Bayesian approach does not negate the utility of probability sampling for design, which is enormously valuable for achieving robust inferences that limit the need for debatable assumptions concerning representativeness of the sample.

Acknowledgements

I thank the Waksberg Award committee for providing me the opportunity to present this work, and Yajuan Si and a referee for helpful comments.

References

Andridge, R.H., and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P.S. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 68, 5, 1465-1483.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Berger, J.O., and Wolpert, R.L. (1988). The likelihood principle. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 6, 1-199.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44,3, 388-393.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57, 269-326.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvaredo-Leiton, F. (2021). A simulation study of diagnostics for bias in non-probability samples. *Journal of Official Statistics*, 37, 3, 751-769.

Brewer, K. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39, 2, 249-262. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11883-eng.pdf.

Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 46, 143-155.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Zhan, L. and Zhang, K. (2019). Models as approximations 1: Consequences illustrated with linear regression. *Statistical Science*, 34, 4, 580-583.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R.J., Sedransk, J. and Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32, 2, 227-248.

Dean, N., and Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3, 484-503.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 23-34. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007001/article/9849-eng.pdf.

Elliott, M.R., and Little, R.J. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 2, 249-264.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B*, 31, 195-233.

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26, 2, 212-226.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

Franco, C., Little, R.J., Louis, T.A. and Slud, E.V. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 3, 334-364.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 2, 153-188.

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

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman and Hall.

Ghosh, M., Reid, N. and Fraser, D.A.S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20, 1309-1332.

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

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

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142, 1, 33-46.

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

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830. Reproduced as Chapter 1 of *Leslie Kish: Selected Papers,* (2003, Eds., G. Kalton and S. Heeringa), New York: John Wiley & Sons, Inc.

Kish, L., and Frankel, M.R. (1974). Inferences from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Lazzeroni, L.C., and Little, R.J. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

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

Little, R.J. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J. (1994). A class of pattern-mixture models for normal missing data. *Biometrika*, 81, 3, 471-483.

Little, R.J. (2003a). The Bayesian approach to sample survey inference. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 49-57.

Little, R.J. (2003b). Bayesian methods for unit and item nonresponse. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Little, R.J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60, 3, 213-223.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics,* 28, 3, 309-372.

Little, R.J. (2014). Survey sampling: Past controversies, current orthodoxies, and future paradigms. In *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, (Eds., X. Lin, D.L. Banks, C. Genest, G. Molenberghs, D.W. Scott and J.-L. Wang), CRC Press.

Little, R.J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the International Association of Survey Statisticians*,31, 4, 555-563.

Little, R.J. (2019). Comment on "Models as approximations 1: Consequences illustrated with linear regression" by A. Buja et al. *Statistical Science*, 34, 4, 580-583.

Little, R.J. (2022). A note about the definition of propensity weights. To appear in *Journal of Survey Statistics and Methodology*.

Little, R.J., and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Little, R.J., 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.

Little, R.J., and Wu, M.M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, 87-95.

Little, R.J., West, B.T., Boonstra, P.S. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 5, 932-964.

Royall, R.M., and Cumberland, W.G. (1981). The finite population linear regression estimator and estimator of its variance-an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Royall, R.M., and Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 1, 34-58.

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

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (2019). Conditional calibration and the sage statistician. *Survey Methodology*, 45, 2, 187-198. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00010-eng.pdf.

Ruppert, D., Wand, M.P and Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

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.

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

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.

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

Si, Y, Trangucci, R., Gabry, J.S. and Gelman, A. (2020). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology*, 46, 2, 181-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020002/article/00003-eng.pdf.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62, 5-34.

Szpiro, A.A., Rice, K.M. and Lumley, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Annals of Applied Statistics*, 4, 4, 2099-2113.

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

West, B.T., Little, R.J., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and Alvarado-Leiton, F. (2021). Measures of selection bias in regression coefficients estimated from non-probability samples. *The Annals of Applied Statistics*, 15(3), 1556-1581.

Yang, Y., and Little, R.J. (2015). A comparison of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation,* 85, 16, 3383-3403.

Zangeneh, S.Z., and Little, R.J. (2015). Bayesian inference for the finite population total in heteroscedastic probability proportional to size samples. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Zangeneh, S.Z., and Little, R.J. (2022). Likelihood based estimation of the finite population mean with post-stratification information under nonignorable nonresponse. To appear in *International Statistical Review*.

Zhang, G., and Little, R.J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 3, 911-918.

Zheng, H., and Little, R.J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics,* 19, 2, 99-117.

Zheng, H., and Little, R.J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 283-311
Statistics Canada, Catalogue No. 12-001-X

Statistical inference with non-probability survey samples

Changbao Wu[1]

Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

Key Words: Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

1. Introduction

The field of survey sampling distinguishes itself from other areas of statistics with a number of unique features. The target population consists of finite number of well defined units, and the population parameters can be determined without error, at least conceptually, by conducting a census. Operational constraints and administrative convenience for data collection often make it necessary to consider stratification, clustering and unequal probability selection. Since the seminal paper of Neyman (1934), probability sampling methods have become one of the primary data collection tools for official statistics and researchers in health sciences, social and economic studies, business and marketing, agricultural and natural resource inventories, and other areas. Probability survey samples have also been used for analytic studies involving models and model parameters; see, for instance, Binder (1983), Godambe and Thompson (1986), Thompson (1997), Rao and Molina (2015), among others. Probability survey samples and design-based inference have been a successful story as part of statistical sciences in the past 80 years.

In recent years, however, "*there has been a wind of change and other data sources are being increasingly explored*" (Beaumont, 2020). The success of probability survey samples led to more ambitious study designs, long and complicated questionnaires and increased burden on respondents. The response rates have been declining and the cost of data collection has been soaring over the years. With the advances of new technology and the explosion of information over the Internet, there is also a strong desire to access real-time statistics. Statistics Canada has launched the so-called modernization initiatives, "*moving beyond a survey-first approach with new methods and integrating data from a variety of existing sources*".

Non-probability survey samples are one of those data sources which have gained increased popularity in recent years. Non-probability samples are not something new to the field of survey sampling. They have been used since the early days of conducting surveys. Quota surveys, for instance, lead to

[1] Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1.E-mail: cbwu@uwaterloo.ca.

non-probability samples, and the method is widely used and can be successful under certain conditions; see Section 5 for further discussions. Non-probability survey samples had not gained true momentum in the past in survey practice due to the lack of a mature theoretical framework for analyzing the data. Nevertheless, they are an available data source that is cheaper and quicker to obtain and have become prevalent for online research. Commercial survey firms create and maintain a long list of individuals, called the *opt-in panels*, who agreed to be contacted to participate in surveys either as volunteers or with incentives. The precise mechanisms for individuals being included in the panel are typically unknown, resulting in panel-based non-probability survey samples.

The main issue with non-probability survey samples is that they are biased samples and do not represent the target population. One might argue that, other than iid samples, most samples are biased, and even probability survey samples are biased. The reason that we do not worry about the biased nature of probability survey samples is the known inclusion probabilities from the survey design, which lead to valid estimation methods through suitable weighting procedures. The real main issue with non-probability survey samples thus is the unknown sample inclusion or participation mechanisms. It will become clear from discussions in Section 4 that the biased nature of non-probability samples cannot be corrected by using the sample itself. It requires additional auxiliary information on the target population.

This paper provides a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. Section 2 describes the general setting, commonly used assumptions, and inferential frameworks for statistical procedures discussed in the paper. Section 3 presents model-based prediction approach to non-probability survey samples. Section 4 discusses estimation of propensity scores and constructions of propensity score based estimators. Section 5 shows the connections between inverse probability weighted estimators and quota surveys with extensions to poststratification. Section 6 focuses on techniques as well as issues with variance estimation. In Section 7, we address the important question on how to check and verify the required assumptions in practice. Some concluding remarks are given in Section 8.

2. Assumptions and inferential frameworks

Suppose that the target population $U = \{1, 2, \ldots, N\}$ consists of N labelled units. Associated with unit i are values $\mathbf{x}_i$ and y_i for the auxiliary variables $\mathbf{x}$ and the study variable y. The discussions focus on a single y but the dataset most likely contains multiple study variables. Let $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ be the population mean which is the parameter of interest. Let $\{(y_i, \mathbf{x}_i), i \in S_A\}$ be the dataset for the non-probability survey sample S_A with n_A participating units. For most practical scenarios, the simple sample mean $\bar{y}_A = n_A^{-1}\sum_{i\in S_A} y_i$ is a biased estimator of μ_y and hence is invalid.

2.1 Assumptions

Let $R_i = I(i \in S_A)$ be the indicator variable for unit i being included in the non-probability sample S_A. Note that the variable R_i is defined for all i in the target population. Let

$$\pi_i^A = P\left(i \in S_A \mid \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right), \quad i = 1, 2, \ldots, N.$$

We call the π_i^A the propensity scores, a term borrowed from the missing data literature (Rosenbaum and Rubin, 1983). Some authors use the term participation probabilities; see, for instance, Beaumont (2020) and Rao (2021), among others. The propensity scores π_i^A characterize the sample inclusion and participation mechanisms. They are unknown and require suitable model assumptions for the development of valid estimation methods. The following three basic assumptions were used by Chen, Li, and Wu (2020), which were adapted from the missing data literature.

A1 The sample inclusion and participation indicator R_i and the study variable y_i are independent given the set of covariates $\mathbf{x}_i$, i.e., $\left(R_i \perp y_i\right) \mid \mathbf{x}_i$.

A2 All the units in the target population have non-zero propensity scores, i.e., $\pi_i^A > 0$, $i = 1, 2, \ldots, N$.

A3 The indicator variables $R_1, R_2, \ldots, R_N$ are independent given the set of auxiliary variables $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$.

Assumption A1 is similar to the missing at random (MAR) assumption for missing data analysis. Under A1, we have $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \mid \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$. Assumption A2 can be problematic in practice; see Section 7 for further discussions. Assumption A3 typical holds when participants are approached one at a time but can be questionable when clustered selections are used. It is shown in Section 4 that estimation of $\pi_i^A = \pi\left(\mathbf{x}_i\right)$ under assumption A1 requires auxiliary information from the target population. The ideal scenario is that the complete auxiliary information $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$ is available. The more practical scenario is that auxiliary information can be obtained from an existing probability survey.

A4 There exists a probability survey sample S_B of size n_B with information on the auxiliary variables $\mathbf{x}$ (but not on y) available in the dataset $\left\{(\mathbf{x}_i, d_i^B), i \in S_B\right\}$, where d_i^B are the design weights for the probability sample S_B.

The S_B is called the reference probability survey sample. The most crucial part of assumption A4 is that the set of auxiliary variables $\mathbf{x}$ is observed in both the non-probability sample S_A and the probability sample S_B. A reference probability survey sample is often available in practice but the common set of auxiliary variables may not contain all the components to satisfy assumption A1.

2.2 Inferential frameworks

There are three possible sources of variation under the general setting of two samples S_A and S_B: (i) The model q for the propensity scores on the sample inclusion and participation in the non-probability survey sample S_A; (ii) The model ξ for the outcome regression $\left(y \mid \mathbf{x}\right)$ or imputation; and (iii) The probability sampling design p for the reference probability survey sample S_B. For the three approaches to inference to be discussed in Sections 3 and 4, the reference probability sample S_B is always involved. Each of the three approaches requires a joint randomization framework involving p and one of (q, ξ).

(a) Model-based prediction approach: The ξp framework under the joint randomization of the outcome regression model ξ and the probability sampling design p.

(b) Inverse probability weighting using estimated propensity scores: The qp framework under the joint randomization of the propensity score model q and the probability sampling design p.

(c) Doubly robust inference: The qp framework or the ξp framework, with no specification of which one.

The inferential framework is the foundation for theoretical development. Consistency of point estimators needs to be established under the suitable joint randomization. Theoretical variances typically involve two components, one from each source of variation, and correct derivations of the two components are the key to the construction of consistent variance estimators under the designated inferential framework.

3. Model-based prediction approach

Model-based prediction methods for finite population parameters require two critical ingredients: the amount of auxiliary information that is available at the estimation stage and the reliability of the assumed model for inference. In the absence of any auxiliary information, the common mean model $E_\xi(y_i) = \mu_0$, $V_\xi(y_i) = \sigma^2$, $i = 1, \ldots, N$ may be viewed as reasonable but the model-based prediction estimator $\hat{\mu}_y = \bar{y}_A = n_A^{-1} \sum_{i \in S_A} y_i$, although unbiased under the model since $E_\xi(\bar{y}_A - \mu_y) = 0$, is generally not an acceptable estimator of μ_y. The variance σ^2 for the common mean model is typically large and it renders the estimator $\hat{\mu}_y = \bar{y}_A$ with a prediction variance that is too large to be practically useful.

3.1 Semiparametric outcome regression models

Without loss of generality, we assume that $\mathbf{x}$ contains 1 as its first component corresponding to the intercept of a regression model. Under the setting described in Section 2, we consider the following semiparametric model for the finite population, denoted as ξ:

$$E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}), \quad \text{and} \quad V_\xi(y_i \mid \mathbf{x}_i) = v(\mathbf{x}_i)\sigma^2, \quad i = 1, 2, \ldots, N, \tag{3.1}$$

where the mean function $m(\cdot,\cdot)$ and the variance function $v(\cdot)$ have known forms, and the y_i's are also assumed to be conditionally independent given the $\mathbf{x}_i$'s. Let $\boldsymbol{\beta}_0$ and σ_0^2 be the true values of the model parameters $\boldsymbol{\beta}$ and σ^2 under the assumed model. The first major implication of assumption A1 is that $E_\xi(y_i \mid \mathbf{x}_i, R_i = 1) = E_\xi(y_i \mid \mathbf{x}_i)$ and $V_\xi(y_i \mid \mathbf{x}_i, R_i = 1) = V_\xi(y_i \mid \mathbf{x}_i)$. The model (3.1) which is assumed for the finite population also holds for the units in the non-probability survey sample S_A. The quasi maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is obtained using the dataset $\{(y_i, \mathbf{x}_i), i \in S_A\}$ from the non-probability survey sample as the solution to the quasi score equations (McCullagh and Nelder, 1989) given by

$$\mathrm{S}(\boldsymbol{\beta}) = \sum_{i \in S_A} \frac{\partial m(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \{v(\mathbf{x}_i)\}^{-1} \{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\} = \mathbf{0}. \tag{3.2}$$

The semiparametric model (3.1) can be extended to replace $v(\mathbf{x}_i)$ by a general variance function $v(\mu_i)$ where $\mu_i = m(\mathbf{x}_i, \boldsymbol{\beta})$. The quasi maximum likelihood estimation theory covers linear or nonlinear regression models with the weighted least square estimators, the logistic regression model and other generalized linear models. Let $m_i = m(\mathbf{x}_i, \boldsymbol{\beta}_0)$ and $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, $i = 1, 2, \ldots, N$.

3.2 Two general forms of prediction estimators

There are two commonly used model-based prediction estimators for μ_y in the presence of complete auxiliary information $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$; see Chapter 5 of Wu and Thompson (2020). Note that $E_\xi(\mu_y) = N^{-1} \sum_{i=1}^N m_i$. The two prediction estimators are constructed as

$$\hat{\mu}_{y1} = \frac{1}{N} \sum_{i=1}^N \hat{m}_i \quad \text{and} \quad \hat{\mu}_{y2} = \frac{1}{N} \left\{ \sum_{i \in S_A} y_i - \sum_{i \in S_A} \hat{m}_i + \sum_{i=1}^N \hat{m}_i \right\}. \tag{3.3}$$

The estimator $\hat{\mu}_{y2}$ is built based on $\mu_y = N^{-1} \left\{ \sum_{i \in S_A} y_i + \sum_{i \notin S_A} y_i \right\}$ and uses $\sum_{i \notin S_A} \hat{m}_i = \sum_{i=1}^N \hat{m}_i - \sum_{i \in S_A} \hat{m}_i$ to predict the unobserved term $\sum_{i \notin S_A} y_i$. Under a linear regression model where $m(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$, the two estimators given in (3.3) reduce to

$$\hat{\mu}_{y1} = \mu_{\mathbf{x}}'\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mu}_{y2} = \frac{n_A}{N} \left(\bar{y}_A - \bar{\mathbf{x}}_A' \hat{\boldsymbol{\beta}} \right) + \mu_{\mathbf{x}}' \hat{\boldsymbol{\beta}}, \tag{3.4}$$

where $\mu_{\mathbf{x}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$ is the vector of the population means of the $\mathbf{x}$ variables and $\bar{\mathbf{x}}_A = n_A^{-1} \sum_{i \in S_A} \mathbf{x}_i$ is the vector of the simple sample means of $\mathbf{x}$ from the non-probability sample S_A. If the linear regression model contains an intercept and $\hat{\boldsymbol{\beta}}$ is the ordinary least square estimator, we have $\hat{\mu}_{y2} = \hat{\mu}_{y1} = \mu_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$ since $\bar{y}_A - \bar{\mathbf{x}}_A'\hat{\boldsymbol{\beta}} = 0$ due to the zero sum of fitted residuals. The prediction estimators in (3.4) under a linear model only require the population means $\mu_{\mathbf{x}}$ in addition to the non-probability sample S_A. Under the setting described in Section 2 with auxiliary information on $\mathbf{x}$ provided through a reference probability sample S_B, we simply replace $\sum_{i=1}^N \hat{m}_i$ by $\sum_{i \in S_B} d_i^B \hat{m}_i$ for the estimators in (3.3) and substitute $\mu_{\mathbf{x}}$ by $\hat{\mu}_{\mathbf{x}} = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B \mathbf{x}_i$ for the estimators in (3.4), where $\hat{N}_B = \sum_{i \in S_B} d_i^B$. The population size N appearing in (3.3) or (3.4) should also be replaced by $\hat{N}_B$ even if it is known.

3.3 Mass imputation

Model-based prediction estimators of μ_y using a non-probability survey sample on $(y, \mathbf{x})$ and a reference probability survey sample on $\mathbf{x}$ have traditionally been presented as the *mass imputation estimator*. The study variable y is not observed for any units in the reference survey sample S_B and hence can be viewed as missing for all $i \in S_B$. Let y_i^* be an imputed value for y_i, $i \in S_B$. The mass imputation estimator of μ_y is then constructed as

$$\hat{\mu}_{y\text{MI}} = \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B y_i^*, \tag{3.5}$$

where $\hat{N}_B$ is defined as before and the subscript "MI" indicates "Mass Imputation" (not "Multiple Imputation"). Under the deterministic regression imputation where $y_i^* = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$, the estimator $\hat{\mu}_{y\text{MI}}$ reduces to the model-based prediction estimator $\hat{\mu}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}$ as discussed in Section 3.2.

The mass imputation approach to analyzing non-probability survey samples has the same spirit as model-based prediction methods but it opens the door for using more flexible models and imputation techniques that have been developed in the existing literature on missing data problems. The approach was first examined by Rivers (2007) through the so-called *sample matching* method. For each $i \in S_B$, the "missing" y_i is imputed as $y_i^* = y_j$ for some $j \in S_A$, where j is a matching donor from S_A selected through the nearest neighbor method as measured by the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The underlying model ξ for the nearest neighbor imputation method is nonparametric, i.e., $E_\xi\left(y_i \mid \mathbf{x}_i\right) = m\left(\mathbf{x}_i\right)$ for some unknown function $m(\cdot)$. The matching value y_j can be viewed as the predicted value of the missing y_i under the model. Theoretical properties of estimators based on nearest neighbor imputation were discussed by Chen and Shao (2000, 2001) for missing survey data problems.

The semiparametric model (3.1) can be used for deterministic regression mass imputation. Under assumption A1, a consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is first obtained from the non-probability sample dataset $\{(y_i, \mathbf{x}_i), i \in S_A\}$, and the estimator $\hat{\boldsymbol{\beta}}$ is then used to compute the imputed values $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ for $i \in S_B$. In other words, the assumption A1 implies the so-called *model transportability* by Kim, Park, Chen and Wu (2021): the model which is built for the non-probability sample can be used for prediction with the reference probability sample. The resulting mass imputation estimator $\hat{\mu}_{y\text{MI}}$ is identical to one of the model-based prediction estimators presented in Section 3.2. Asymptotic properties and variance estimation for the estimator $\hat{\mu}_{y\text{MI}}$ using the semiparametric model (3.1) were discussed by Kim et al. (2021).

Under the mass imputation approach, the only role played by the observed y_i for $i \in S_A$ is to estimate the model parameters $\boldsymbol{\beta}$. The estimator $\hat{\mu}_{y\text{MI}}$ is constructed using the fitted model and auxiliary information from the reference probability sample S_B. It seems that we did not fully use the information on the observed y_i given that μ_y is the main parameter of interest. This led to the research question described in Chapter 17 of Wu and Thompson (2020) on "*reverse sample matching*". The proposed estimator is constructed as $\hat{\mu}_{y\text{A}} = (\hat{N}^*)^{-1} \sum_{i \in S_A} d_i^* y_i$ using all the observed y_i in the non-probability sample, where $\hat{N}^* = \sum_{i \in S_A} d_i^*$. The d_i^* is a matched survey weight from S_B such that $d_i^* = d_j^B$ with $j \in S_B$ being the nearest neighbor of $i \in S_A$ as measured by $\| \mathbf{x}_i - \mathbf{x}_j \|$. Theoretical properties of the reverse matched estimator $\hat{\mu}_{yA}$ using the nearest neighbor $j \in S_B$ to match d_i^* with d_j^B have not been formally investigated in the existing literature.

Wang, Graubard, Katki and Li (2020) proposed a kernel weighting approach to reverse sample matching using $d_i^* \propto \sum_{j \in S_B} K_{ij} d_j$, where K_{ij} is a kernel distance between $\hat{p}_i$ and $\hat{p}_j$; see the adjusted logistic propensity (ALP) weighting method discussed at the end of Section 4.1.1 on the calculation of $\hat{p}_i$. They showed that the estimator $\hat{\mu}_{y\mathrm{A}}$ is consistent under certain regularity conditions. In a recent working paper posted on arXiv by Liu and Valliant (2021), the authors discussed issues with the bias and the variance of the reverse matched estimator under different randomization frameworks involving one, two or all three of the sources (p, q, ξ). The authors also proposed a calibration step over the matched weights, which seems to be a promising idea. Further research on this topic is needed.

The mass imputation approach to analyzing non-probability survey samples leads to an interesting research question that is currently under investigation by a doctoral student at University of Waterloo: Is it theoretically feasible and practically useful to create a mass-imputed dataset $\{(y_i^*, \mathbf{x}_i, d_i^B), i \in S_B\}$ based on the reference probability survey sample that can be used for general statistical inferences? The answer clearly depends on the types of inferential problems to be conducted over the imputed dataset. A minimum requirement is that the conditional distribution of the study variable y given the covariates $\mathbf{x}$ is preserved for the mass-imputed dataset. The nearest neighbor imputation method and the random regression imputation method can be useful for this purpose. Fractional imputation is another possibility, especially for binary or ordinal study variables. Multiple imputation is also potentially useful in this direction to create multiple mass-imputed datasets. The subscript “MI” in this case might need to be changed to “MI^2”, meaning “Mass Imputation with Multiple Imputation”.

4. Propensity scores based approach

The propensity scores $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i)$ for the non-probability survey sample S_A are theoretically defined for all the units in the target population. Estimation of the propensity scores for units in S_A, which plays the most crucial role for propensity scores based methods, requires an assumed model on the propensity scores and auxiliary information at the population level. In this section, we first discuss estimation procedures for the propensity scores under the setting and assumptions described in Section 2, and then provide an overview of estimation methods proposed in the recent literature on the finite population mean μ_y involving the estimated propensity scores.

4.1 Estimation of propensity scores

Under assumption A1, the propensity scores $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i) = \pi(\mathbf{x}_i)$ are a function of the auxiliary variables $\mathbf{x}_i$ but the functional form can be complicated and is completely unknown. Three popular parametric forms $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ in dealing with a binary response can be considered: (i) the inverse logit function $\pi_i^A = 1 - \{1 + \exp(\mathbf{x}_i'\boldsymbol{\alpha})\}^{-1}$; (ii) the inverse probit function $\pi_i^A = \Phi(\mathbf{x}_i'\boldsymbol{\alpha})$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$; and (iii) the inverse complementary log-log function

$\pi_i^A = 1 - \exp\{-\exp(\mathbf{x}_i'\boldsymbol{\alpha})\}$. Nonparametric techniques without assuming an explicit functional form for $\pi(\mathbf{x})$ are attractive alternatives for the estimation of propensity scores.

4.1.1 The pseudo maximum likelihood method

Let $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ be a specified parametric form with unknown model parameters $\boldsymbol{\alpha}$. Under the ideal situation where the complete auxiliary information $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is available and with the independence assumption A3, the full log-likelihood function on $\boldsymbol{\alpha}$ can be written as (Chen et al., 2020)

$$\ell(\boldsymbol{\alpha}) = \log\left\{\prod_{i=1}^N \left(\pi_i^A\right)^{R_i}\left(1-\pi_i^A\right)^{1-R_i}\right\} = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i=1}^N \log\left(1-\pi_i^A\right). \qquad (4.1)$$

The maximum likelihood estimator of $\boldsymbol{\alpha}$ is the maximizer of $\ell(\boldsymbol{\alpha})$. Under the current setting where the population auxiliary information is supplied by the reference probability sample S_B, we replace $\ell(\boldsymbol{\alpha})$ by the pseudo log-likelihood function (Chen et al., 2020)

$$\ell^*(\boldsymbol{\alpha}) = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i\in S_B} d_i^B \log\left(1-\pi_i^A\right). \qquad (4.2)$$

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\alpha}}$ is the maximizer of $\ell^*(\boldsymbol{\alpha})$ and can be obtained as the solution to the pseudo score equations given by $\mathbf{U}(\boldsymbol{\alpha}) = \partial\ell^*(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = \mathbf{0}$. If the inverse logit function is assumed for π_i^A, the pseudo score functions are given by

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{i\in S_A}\mathbf{x}_i - \sum_{i\in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha})\,\mathbf{x}_i. \qquad (4.3)$$

In general, the pseudo score functions $\mathbf{U}(\boldsymbol{\alpha})$ at the true values of the model parameters $\boldsymbol{\alpha}_0$ are unbiased under the joint qp randomization in the sense that $E_{qp}\{\mathbf{U}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$, which implies that the estimator $\hat{\boldsymbol{\alpha}}$ is qp-consistent for $\boldsymbol{\alpha}_0$ (Tsiatis, 2006).

Valliant and Dever (2011) made an earlier attempt to estimate the propensity scores by pooling the non-probability sample S_A with the reference probability sample S_B. Let $S_{AB} = S_A \cup S_B$ be the pooled sample without removing any potential duplicated units. Let $R_i^* = 1$ if $i \in S_A$ and $R_i^* = 0$ if $i \in S_B$. Valliant and Dever (2011) proposed to fit a survey weighted logistic regression model to the pooled dataset $\{(R_i^*, \mathbf{x}_i, d_i), i \in S_{AB}\}$, where the weights are defined as $d_i = 1$ if $i \in S_A$ and $d_i = d_i^B\left(1 - n_A/\hat{N}_B\right)$ if $i \in S_B$. The key motivation behind the creation of the weights d_i is that the total weight $\sum_{i\in S_{AB}} d_i = \sum_{i\in S_B} d_i^B = \hat{N}_B$ for the pooled sample matches the estimated population size, and the hope is that the survey weighted logistic regression model would lead to valid estimates for the propensity scores. It was shown by Chen et al. (2020) that the pooled sample approach of Valliant and Dever (2011) does not lead to consistent estimators for the parameters of the propensity scores model unless the non-probability sample S_A is a simple random sample from the target population.

The method of Valliant and Dever (2011) reveals a fundamental difficulty with approaches based on the pooled sample S_{AB}. If the units in the non-probability sample S_A are treated as exchangeable in the pooled sample S_{AB}, which was reflected by the equal weights $d_i = 1$ used in the method of Valliant and Dever (2011), the resulting estimates for the propensity scores will be invalid unless S_A is a simple random sample. This observation has implications to the validity of nonparametric methods or regression tree-based methods to be discussed in Section 4.1.3.

In a recent paper, Wang, Valliant and Li (2021) proposed an adjusted logistic propensity (ALP) weighting method. The method involves two steps for computing the estimated propensity scores. The initial estimates, denoted as $\hat{p}_i$ for $i \in S_A$, are obtained by fitting the survey weighted logistic regression model to the pooled sample S_{AB} similar to Valliant and Dever (2011), with the weights defined as $d_i = 1$ if $i \in S_A$ and $d_i = d_i^B$ if $i \in S_B$. The final estimated propensity scores are computed as $\hat{\pi}_i^A = \hat{p}_i / (1 - \hat{p}_i)$. The key theoretical argument is the equation $\pi_i^A = p_i / (1 - p_i)$ where $\pi_i^A = P(i \in S_A \mid U)$, $p_i = P(i \in S_A^* \mid S_A^* \cup U)$, and S_A^* is a copy of S_A but is viewed as a different set. However, there are conceptual issues with the arguments since the probabilities $\pi_i^A = P(i \in S_A \mid U)$ are defined under the assumed propensity scores model with the given finite population U, and the assumed model does not lead to a meaningful interpretation of the probabilities $p_i = P(i \in S_A^* \mid S_A^* \cup U)$. The latter require a different probability space and are conditional on the given S_A. As a matter of fact, one can easily argue that under the assumed propensity scores model and conditional on the given S_A, we have $p_i = 1$ if $i \in S_A$ and $p_i = 0$ otherwise.

4.1.2 Estimating equations based methods

The pseudo score equations $\mathbf{U}(\boldsymbol{\alpha}) = \mathbf{0}$ derived from the pseudo likelihood function $\ell^*(\boldsymbol{\alpha})$ may be replaced by a system of general estimating equations. Let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ be a user-specified vector of functions with the same dimension of $\boldsymbol{\alpha}$. Let

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha}) \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}). \tag{4.4}$$

It follows that $E_{qp}\{\mathbf{G}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ for any chosen $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. In principle, an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ can be obtained by solving $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ with the chosen parametric form $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ and the chosen functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$, and the estimator $\hat{\boldsymbol{\alpha}}$ is consistent.

The estimator $\hat{\boldsymbol{\alpha}}$ using arbitrary user-specified functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ is typically less efficient than the one based on the pseudo score functions, due to the optimality of the maximum likelihood estimator (Godambe, 1960). Some limited empirical results also show that the solution to $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ can be unstable for certain choices of $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. Nevertheless, the estimating equations based methods provide a useful tool for the estimation of the propensity scores under more restricted scenarios. For instance, if we let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha})$, the estimating functions given in (4.4) reduce to

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \sum_{i \in S_B} d_i^B \mathbf{x}_i. \tag{4.5}$$

The form of $\mathbf{G}(\boldsymbol{\alpha})$ in (4.5) looks like a "distorted" version of the pseudo score functions given in (4.3) under a logistic regression model for the propensity scores. The most practically important difference between the two versions, however, is the fact that the $\mathbf{G}(\boldsymbol{\alpha})$ given in (4.5) only requires the estimated population totals for the auxiliary variables $\mathbf{x}$. There are scenarios where the population totals of the auxiliary variables $\mathbf{x}$ can be accessed or estimated from an existing source but values of $\mathbf{x}$ at the unit level for the entire population or even a probability sample are not available. The use of estimating functions $\mathbf{G}(\boldsymbol{\alpha})$ given (4.5) makes it possible to obtain valid estimates of the propensity scores for units in the non-probability sample. Section 6.3 describes an example where the estimating equations based approach leads to a valid variance estimator for the doubly robust estimator of the population mean.

4.1.3 Nonparametric methods and regression-tree based methods

The propensity scores $\pi_i^A = P(R_i = 1 | \mathbf{x}_i)$ are the mean function $E_q(R_i | \mathbf{x}_i) = \pi(\mathbf{x}_i)$ for the binary response R_i. Nonparametric methods for estimating $\pi(\mathbf{x})$ can be an attractive alternative. The major challenge is to develop estimation procedures which provide valid estimates of the propensity scores. As noted in Section 4.1.1, estimation methods based on the pooled sample $S_{AB} = S_A \cup S_B$ may lead to invalid estimates. Strategies similar to the one used by Chen et al. (2020) can be theoretically justified under the joint qp framework, where the estimation procedures are first derived using data from the entire finite population and unknown population quantities are then replaced by estimates obtained from the reference probability sample.

We consider the kernel regression estimator of $\pi_i^A = \pi(\mathbf{x}_i)$. Suppose that the dataset $\{(R_i, \mathbf{x}_i), i = 1, 2, \ldots, N\}$ is available for the finite population. Let $K_h(t) = K(t/h)$ be a chosen kernel with a bandwidth h. The Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964) of $\pi(\mathbf{x})$ is given by

$$\tilde{\pi}(\mathbf{x}) = \frac{\sum_{j=1}^N K_h(\mathbf{x} - \mathbf{x}_j) R_j}{\sum_{j=1}^N K_h(\mathbf{x} - \mathbf{x}_j)}. \tag{4.6}$$

A kernel estimator in the form of $\tilde{\pi}(\mathbf{x})$ given in (4.6) usually has no practical values since we do not have complete auxiliary information for the finite population. It turns out that for the estimation of propensity scores the numerator in (4.6) only requires observations from the non-probability sample due to the binary variable R_j, and the denominator is a population total and can be estimated by using the reference probability sample. The nonparametric kernel regression estimator of the propensity scores is given by (Yuan, Li and Wu, 2022)

$$\hat{\pi}_i^A = \hat{\pi}(\mathbf{x}_i) = \frac{\sum_{j \in S_A} K_h(\mathbf{x}_i - \mathbf{x}_j)}{\sum_{j \in S_B} d_j^B K_h(\mathbf{x}_i - \mathbf{x}_j)}, \quad i \in S_A. \tag{4.7}$$

The estimator $\hat{\pi}_i^A$ given in (4.7) is consistent under the joint qp framework and the q-model for the propensity scores is very flexible due to the nonparametric assumption on $\pi(\mathbf{x})$. The estimated propensity scores are easy to compute when the dimension of $\mathbf{x}$ is not too high. Issues with high dimensional $\mathbf{x}$ and the choices of the kernel $K_h(\cdot)$ and the bandwidth h remain as in general applications of kernel-based estimation methods. Simulation results reported by Yuan et al. (2022) show that the kernel estimation method provides robust results for the propensity scores using the normal kernel and popular choices for the bandwidth.

Chu and Beaumont (2019) considered regression-tree based methods for estimating the propensity scores. Their proposed TrIPW method is a variant of the CART algorithm (Breiman, Friedman, Olshen and Stone, 1984) and uses data from the combined sample of the non-probability sample and the reference probability sample. The method aims to construct a classification tree with the terminal nodes of the final tree treated as homogeneous groups in terms of the propensity scores. The estimator of μ_y is constructed based on the final tree and post-stratification. Section 5 contains further details on poststratified estimators.

Statistical learning techniques such as classification and regression trees and random forests have been developed primarily for the purpose of prediction. Their use for estimating the propensity scores of non-probability samples requires further research. It is not a desirable approach to naively apply the methods over the pooled sample S_{AB} without theoretical justifications on the consistency of the final estimators. Further research towards this direction should be encouraged.

4.2 Inverse probability weighting

Let $\hat{\pi}_i^A$ be an estimate of $\pi_i^A = P\left(i \in S_A \mid \mathbf{x}_i\right)$ under a chosen method for the estimation of the propensity scores. Two versions of the inverse probability weighted (IPW) estimator of μ_y are constructed as

$$\hat{\mu}_{\text{IPW1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A} \quad \text{and} \quad \hat{\mu}_{\text{IPW2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A}, \tag{4.8}$$

where N is the population size and $\hat{N}^A = \sum_{i\in S_A}(\hat{\pi}_i^A)^{-1}$ is the estimated population size. The estimator $\hat{\mu}_{\text{IPW1}}$ is a version of the Horvitz-Thompson estimator and $\hat{\mu}_{\text{IPW2}}$ corresponds to the Hájek estimator as discussed in design-based estimation theory. There are ample evidences from both theoretical justifications and practical observations that the Hájek estimator $\hat{\mu}_{\text{IPW2}}$ performs better than the Horvitz-Thompson estimator and should be used in practice even if the population size N is known.

The validity of the IPW estimators $\hat{\mu}_{\text{IPW1}}$ and $\hat{\mu}_{\text{IPW2}}$ depends on the validity of the estimated propensity scores. Under the assumptions A1 and A2 and the parametric model $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$, the consistency of $\hat{\mu}_{\text{IPW1}}$ follows a standard two-step argument. Let $\tilde{\mu}_{\text{IPW}} = N^{-1}\sum_{i\in S_A} y_i / \pi_i^A$, which is not a computable estimator but an analytic tool useful for asymptotic purposes. It follows that $E_q\left(\tilde{\mu}_{\text{IPW}}\right) = \mu_y$ and the order $V_q\left(\tilde{\mu}_{\text{IPW}}\right) = O\left(n_A^{-1}\right)$ holds under the condition that $n_A \pi_i^A / N$ is bounded away from zero. As a consequence, we have $\tilde{\mu}_{\text{IPW}} \to \mu_y$ in probability as $n_A \to \infty$. Under the correctly specified model $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$ for the propensity scores, the typical root-n order $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p(n_A^{-1/2})$ holds for commonly encountered scenarios. We can show by treating $\hat{\mu}_{\text{IPW1}}$ as a function of $\hat{\boldsymbol{\alpha}}$ and using a Taylor series expansion that $\hat{\mu}_{\text{IPW1}} = \tilde{\mu}_{\text{IPW}} + O_p(n_A^{-1/2})$ under some mild finite moment conditions. The consistency of $\hat{\mu}_{\text{IPW2}}$ can be established using standard arguments for a ratio estimator (Section 5.3, Wu and Thompson, 2020) where $N^{-1}\sum_{i\in S_A}(\pi_i^A)^{-1} = 1 + o_p(1)$.

4.3 Doubly robust estimation

The dependence of the IPW estimator on the validity of the assumed propensity score model is viewed as a weakness of the method. The issue is not unique to the IPW estimators and is faced by many other approaches involving an assumed statistical model. Robust estimation procedures which provide certain degrees of protection against model misspecifications have been pursued by researchers, and the so-called doubly robust estimators have been a successful story since the work of Robins, Rotnitzky, and Zhao (1994).

The doubly robust (DR) estimator of μ_y is constructed using both the propensity score model q and the outcome regression model ξ. The DR estimator with the given propensity scores $\pi_i^A,\ i \in S_A$ and the mean responses $m_i = E_\xi(y_i \mid \mathbf{x}_i),\ i = 1, 2, \ldots, N$ has the following general form,

$$\tilde{\mu}_{\text{DR}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i - m_i}{\pi_i^A} + \frac{1}{N}\sum_{i=1}^{N} m_i. \tag{4.9}$$

The second term on the right hand side of (4.9) is the model-based prediction of μ_y. The first term is a propensity score based adjustment using the errors $\varepsilon_i = y_i - m_i$ from the outcome regression model. The magnitude of the adjustment term is negatively correlated to the "goodness-of-fit" of the outcome regression model. It can be shown that $\tilde{\mu}_{\text{DR}}$ is an exactly unbiased estimator of μ_y if one of the two models q and ξ is correctly specified and hence it is doubly robust. The estimator $\tilde{\mu}_{\text{DR}}$ has an identical structure to the generalized difference estimator of Wu and Sitter (2001). It is important to note that the double robustness property of $\tilde{\mu}_{\text{DR}}$ does not require the knowledge of which of the two models being correctly specified. It is also apparent that the estimator $\tilde{\mu}_{\text{DR}}$ given in (4.9) is not computable in practical applications.

Let $\hat{\pi}_i^A$ and $\hat{m}_i$ be respectively the estimators of π_i^A and m_i under the assumed models q and ξ. Under the two-sample setting described in Section 2, the two DR estimators of μ_y proposed by Chen et al. (2020) are given by

$$\hat{\mu}_{\text{DR1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \tag{4.10}$$

and

$$\hat{\mu}_{\text{DR}2} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i^B \hat{m}_i, \tag{4.11}$$

where d_i^B are the design weights for the probability sample S_B, $\hat{N}^A = \sum_{i \in S_A} \left(\hat{\pi}_i^A\right)^{-1}$ and $\hat{N}^B = \sum_{i \in S_B} d_i^B$. The estimator $\hat{\mu}_{\text{DR}2}$ using the estimated population size has better performance in terms of bias and mean squared error and should be used in practice.

The probability survey design p is an integral part of the theoretical framework for assessing the two estimators $\hat{\mu}_{\text{DR}1}$ and $\hat{\mu}_{\text{DR}2}$. It is assumed that S_A and S_B are selected independently, which implies that $E_p\left(\sum_{i \in S_B} d_i^B \hat{m}_i\right) = \sum_{i=1}^N \hat{m}_i$. Consistency of the estimators $\hat{\mu}_{\text{DR}1}$ and $\hat{\mu}_{\text{DR}2}$ can be established under either the qp or the ξp framework. It should be noted that even if the non-probability sample S_A is a simple random sample with $\pi_i^A = n_A / N$, the doubly robust estimator in the form of (4.9) does not reduce to the model-based prediction estimator $\hat{\mu}_{y2}$ given in (3.3).

4.4 The pseudo empirical likelihood approach

The pseudo empirical likelihood (PEL) methods for probability survey samples have been under development over the past two decades. Two early papers on the topic are Chen and Sitter (1999) on point estimation incorporating auxiliary information and Wu and Rao (2006) on PEL ratio confidence intervals. The PEL approaches are further used for multiple frame surveys (Rao and Wu, 2010a) and Bayesian inferences with survey data (Rao and Wu, 2010b; Zhao, Ghosh, Rao and Wu, 2020b). Using the PEL methods for general inferential problems with complex surveys has been studied in two recent papers (Zhao and Wu, 2019; Zhao, Rao and Wu, 2020a).

Chen, Li, Rao and Wu (2022) showed that the PEL provides an attractive alternative approach to inference with non-probability survey samples. Let $\hat{\pi}_i^A$, $i \in S_A$ be the estimated propensity scores under an assumed parametric or non-parametric model, q. The PEL function for the non-probability survey sample S_A is defined as

$$\ell_{\text{PEL}}(\mathbf{p}) = n_A \sum_{i \in S_A} \tilde{d}_i^A \log\left(p_i\right), \tag{4.12}$$

where $\mathbf{p} = (p_1, \ldots, p_{n_A})$ is a discrete probability measure over the n_A selected units in S_A, $\tilde{d}_i^A = (\hat{\pi}_i^A)^{-1} / \hat{N}^A$ and $\hat{N}^A = \sum_{j \in S_A} (\hat{\pi}_j^A)^{-1}$ which is defined earlier in Section 4. Without using any additional information, maximizing $\ell_{\text{PEL}}(\mathbf{p})$ under the normalization constraint

$$\sum_{i \in S_A} p_i = 1 \tag{4.13}$$

leads to $\hat{p}_i = \tilde{d}_i^A$, $i \in S_A$. The maximum PEL estimator of μ_y is given by $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$, which is identical to the IPW estimator $\hat{\mu}_{\text{IPW}2}$ given in (4.8).

The PEL approach to non-probability survey samples provides flexibilities in combining information through additional constraints and constructing confidence intervals and conducting hypothesis tests using the PEL ratio statistic. The maximum PEL estimator $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$ is doubly robust if $\left(\hat{p}_1, \ldots, \hat{p}_{n_A}\right)$ is the maximizer of $\ell_{\text{PEL}}(\mathbf{p})$ under both the normalization constraint and the model-calibration constraint given by

$$\sum_{i \in S_A} p_i \hat{m}_i = \bar{m}^B, \tag{4.14}$$

where $\bar{m}^B = \left(\hat{N}^B\right)^{-1} \sum_{i \in S_B} d_i^B \hat{m}_i$ is computed using the fitted values $\hat{m}_i,\, i \in S_B$ from an assumed outcome regression model, ξ. The equation (4.14) is a modified version of the original model-calibration constraint of Wu and Sitter (2001) using the probability sample S_B. Chen et al. (2022) contain further details on the asymptotic distributions of the PEL ratio statistic and simulation studies on the performances of PEL ratio confidence intervals on a finite population proportion.

5. Quota surveys and poststratification

Quota surveys are one of the oldest non-probability survey sampling methods which are still used in practice in present days. For a pre-specified overall sample size n_A, quotas of sample sizes are set for subpopulations which are defined by demographic variables and social-economic status indicators or other characteristic variables suitable for units of the target population. Data collection processes continue until quotas for each of the subpopulations are filled. Units from the population are typically approached using whatever convenient ways available and there are little or no controls on how units are selected for the final sample other than the pre-specified quotas.

The theory of the IPW estimators for non-probability survey samples provides an opportunity to examine scenarios where quota surveys may succeed or fail. For the convenience of notation without loss of generality, let S_A be the quota survey sample and $\mathbf{x}$ be the set of categorical variables used for defining the subpopulations and setting the quotas. The overall sample can be partitioned into $S_A = S_{A1} \cup \cdots \cup S_{AK}$ corresponding to the cross-classification of sampled units using the combinations of levels of the $\mathbf{x}$ variables. For instance, if $\mathbf{x} = (x_1, x_2)'$ with x_1 having two levels and x_2 having three levels, we have a total of $K = 2 \times 3 = 6$ subpopulations defined by $\mathbf{x}$. Let n_k be the pre-specified size of S_{Ak} and N_k be the size of the corresponding subpopulation. Under the assumption A1, the propensity scores $\pi_i^A = \pi(\mathbf{x}_i)$ become a constant for units in the same subpopulation and are given by $\pi_i^A = n_k / N_k$ for the k^{th} subpopulation. The IPW estimator $\hat{\mu}_{\text{IPW2}}$ given in (4.8) reduces to

$$\hat{\mu}_{\text{IPW2}} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i \in S_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k, \tag{5.1}$$

where $\bar{y}_k = n_k^{-1} \sum_{i \in S_{Ak}} y_i,\, \hat{W}_k = \hat{N}_k / \hat{N}^A$, $\hat{N}_k$ is the size of the k^{th} subpopulation obtained or estimated from external sources, and $\hat{N}^A = \sum_{k=1}^{K} \hat{N}_k$. Under the current setting with the availability of a reference probability sample S_B, we form the same partition as cross-classified by levels of $\mathbf{x}$ and obtain $S_B = S_{B1} \cup \cdots \cup S_{BK}$. We can then use $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$.

The estimator given in (5.1) is the standard poststratified estimator of μ_y. It requires the information on the "stratum weights" $\hat{W}_k,\ k=1,\ldots,K,$ which is not available from the sample data itself. Quota surveys, combined with the use of the poststratified estimator, can be successful in producing valid population estimates for the study variable y if the following conditions hold:

(i) The categorical variables $\mathbf{x}$ used in defining the subpopulations and setting the quotas provide characterizations of the participation behavior of the units for voluntary surveys.

(ii) The inclusion of units in the survey is relatively random within each subpopulation and no specific groups are intentionally excluded from the survey.

(iii) The information on the stratum weights corresponding to the cross-classifications in setting the quotas can be reliably obtained from external sources.

(iv) The hardcore nonrespondents in the population who never take any voluntary surveys possess similar features to respondents in terms of the study variable y.

The IPW estimators $\hat{\mu}_{\text{IPW1}}$ and $\hat{\mu}_{\text{IPW2}}$ given in (4.8) may be sensitive to small values of estimated propensity scores. The poststratified estimator in the form of (5.1) serves as a robust alternative under general scenarios where the dimension of $\mathbf{x}$ is not low and/or some components of $\mathbf{x}$ are continuous. The K strata are formed based on homogeneous groups in terms of the propensity scores. Suppose that $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}),\ i \in S_A$ are computed based on a parametric model, q. Suppose also that $n_A = m_A K$ with the chosen K where m_A is an integer. Let $\hat{\pi}_{(1)}^A \leq \cdots \leq \hat{\pi}_{(n_A)}^A$ be the estimated propensity scores in ascending order. Let S_{A1} be the set of the first m_A units in the sequence, S_{A2} be the second m_A units in the sequence, and so on. The poststratified estimator of μ_y is computed as $\hat{\mu}_{\text{PST}} = \sum_{k=1}^K \hat{W}_k \bar{y}_k,$ which has the same form of the estimator given in (5.1). The estimates of the stratum weights, $\hat{W}_k,\ k=1,2,\ldots,K$ can be obtained by using the reference probability sample S_B as follows. Let $b_k = \max\left\{\hat{\pi}_i^A\colon i \in S_{Ak}\right\},$ $k=1,2,\ldots,K-1.$ Let $b_0 = 0$ and $b_K = 1.$

(a) Compute $\hat{\pi}_i = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}),\ i \in S_B.$

(b) Define $S_{Bk} = \left\{i \mid i \in S_B, b_{k-1} < \hat{\pi}_i \leq b_k\right\},\ k=1,2,\ldots,K.$

(c) Calculate $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B,\ k=1,2,\ldots,K.$

It is apparent that $S_B = S_{B1} \cup \cdots \cup S_{BK}$ and $\sum_{k=1}^K \hat{N}_k = \hat{N}^B = \sum_{i \in S_B} d_i^B.$ The estimated stratum weights are given by $\hat{W}_k = \hat{N}_k \big/ \hat{N}^B.$

The choice of K needs to reflect the balance between homogeneity of the units within each post-stratum (in terms of the propensity scores) and the stability of the poststratified estimator (in terms of the stratum sample sizes). When the sample size n_A is small or moderate, a small number such as $K=5$ should be used. For scenarios where n_A is large, a larger K should be used such that units within the same poststratified sample S_{Ak} have similar estimated propensity scores. A practical guidance for the choice of K is to ensure that $m_A \geq 30$ for the poststratified samples. For those who are old enough, do you remember the good old days when "the sample size is large" means "$n \geq 30$"?

6. Variance estimation

Variance estimation under the two sample S_A and S_B setup involves at least two different sources of variation. The probability sampling design for the reference sample S_B remains one of the sources regardless of the approaches used for non-probability survey samples. Estimation of the variance component due to the use of S_B requires either suitable variance approximation formulas or replication weights as part of the dataset from the reference probability sample. Our discussion in this section assumes that a design-based variance estimator for the survey weighted point estimator based on S_B is available.

6.1 Variance estimation for mass imputation estimators

Variance estimation for the model-based prediction estimator $\hat{\mu}_y$ involves first deriving the asymptotic variance formula for $\operatorname{Var}\left(\hat{\mu}_y - \mu_y\right)$ under the assumed outcome regression model or the imputation model ξ and the probability sampling design $p,$ and then using plug-in estimators for various unknown population quantities.

The mass imputation estimator $\hat{\mu}_{y\mathrm{MI}} = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B y_i^*$ given in (3.5) is a special type of model-based prediction estimator, where the model ξ refers to the one used for imputation and is not necessarily the same as the outcome regression model. The imputation method plays a key role in deriving the asymptotic variance formula, and the variance estimator needs to be constructed accordingly. Noting that $\hat{\mu}_{y\mathrm{MI}}$ is a Hájek type estimator due to the use of the estimated population size $\hat{N}_B,$ derivations of the asymptotic variance formula start with putting the true value N in first and then dealing with $\hat{\mu}_{y\mathrm{MI}}$ as a ratio estimator. Kim et al. (2021) considered variance estimation for $\hat{\mu}_y = N^{-1} \sum_{i \in S_B} d_i^B y_i^*,$ where $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ is the imputed value for y_i based on the semiparametric model (3.1). The asymptotic variance formula is developed in two steps. First, a linearized version of $\hat{\mu}_y$ is obtained by using a Taylor series expansion at $\boldsymbol{\beta}^*,$ where $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right).$ Second, two variance components are derived for $\operatorname{Var}\left(\hat{\mu}_y - \mu_y\right)$ based on the linearized version using the semiparametric model (3.1) and the sampling design for $S_B.$ The process is tedious, which is the case for most model-based variance estimation methods. A bootstrap variance estimator turns out to be more attractive for practical applications. See Kim et al. (2021) for further details.

6.2 Variance estimation for IPW estimators

The commonly used IPW estimator $\hat{\mu}_{\mathrm{IPW2}}$ given in (4.8) is valid under the assumed model q for the propensity scores. An explicit asymptotic variance formula for $\hat{\mu}_{\mathrm{IPW2}}$ can be derived under the joint qp-framework when the propensity scores are estimated using the pseudo maximum likelihood method or an estimating equation based method as discussed in Section 4.1. The theoretical tool is the sandwich-type variance formula for point estimators defined as the solution to a combined system of estimating equations for both μ_y and $\boldsymbol{\alpha}_0.$

Consider the parametric form $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ for the propensity scores, where the model parameters $\boldsymbol{\alpha}$ are estimated through the estimating equations (4.4) with user-specified functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. The first major step in deriving the asymptotic variance formula for $\hat{\mu}_{\text{IPW2}}$ is to write down the system of joint estimating equations for both μ_y and $\boldsymbol{\alpha}_0$. Let $\boldsymbol{\eta} = \left(\mu, \boldsymbol{\alpha}'\right)'$ be the vector of the combined parameters. The estimator $\hat{\boldsymbol{\eta}} = \left(\hat{\mu}_{\text{IPW2}}, \boldsymbol{\alpha}'\right)'$ is the solution to the system of joint estimating equations $\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \mathbf{0}$, where

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \begin{pmatrix} N^{-1}\sum_{i=1}^{N} R_i\left(y_i - \mu\right)/\pi_i^A \\ N^{-1}\sum_{i=1}^{N} R_i \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) - N^{-1}\sum_{i\in S_B} d_i^B \pi_i^A \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \end{pmatrix}. \tag{6.1}$$

The factor N^{-1} is redundant but useful in facilitating asymptotic orders. The estimating functions defined by (6.1) are unbiased under the joint qp-framework, i.e., $E_{qp}\left\{\boldsymbol{\Phi}\left(\boldsymbol{\eta}_0\right)\right\} = \mathbf{0}$, where $\boldsymbol{\eta}_0 = \left(\mu_y, \boldsymbol{\alpha}_0'\right)'$. There are two major consequences from the unbiasedness of the estimating equations system. First, consistency of the estimator $\hat{\boldsymbol{\eta}}$ can be argued using the theory of general estimating functions similar to those presented in Section 3.2 of Tsiatis (2006). Second, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\eta}}$, denoted as $\text{AV}\left(\hat{\boldsymbol{\eta}}\right)$, has the standard sandwich form and is given by

$$\text{AV}\left(\hat{\boldsymbol{\eta}}\right) = \left[E\left\{\boldsymbol{\phi}_n\left(\boldsymbol{\eta}_0\right)\right\}\right]^{-1} \text{Var}\left\{\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}_0\right)\right\}\left[E\left\{\boldsymbol{\phi}_n\left(\boldsymbol{\eta}_0\right)\right\}'\right]^{-1},$$

where $\boldsymbol{\phi}_n(\boldsymbol{\eta}) = \partial\boldsymbol{\Phi}_n(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, which depends on the forms of $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ and $\mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha})$. The term $\text{Var}\left\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\right\}$ consists of two components, one due to the propensity score model q and the other from the probability sampling design for S_B. More specifically, we have $\text{Var}\left\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\right\} = V_q(\mathbf{A}_1) + V_p(\mathbf{A}_2)$, where $V_q(\cdot)$ denotes the variance under the propensity score model q and $V_p(\cdot)$ represents the design-based variance under the probability sampling design p, and

$$\mathbf{A}_1 = \frac{1}{N}\sum_{i=1}^{N} R_i \begin{pmatrix} (y_i - \mu)/\pi_i^A \\ \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \end{pmatrix}, \quad \mathbf{A}_2 = \frac{1}{N}\sum_{i\in S_B} d_i^B \begin{pmatrix} 0 \\ \pi_i^A \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \end{pmatrix}.$$

The analytic expression for $V_q(\mathbf{A}_1)$ follows immediately from $V_q(R_i) = \pi_i^A(1 - \pi_i^A)$ and the independence among $R_1, \ldots, R_N$. The design-based variance component $V_p(\mathbf{A}_2)$ requires additional information on the survey design for S_B or a suitable variance approximation formula with the given design.

The asymptotic variance formula for the IPW estimator $\hat{\mu}_{\text{IPW2}}$ is the first diagonal element of the matrix $\text{AV}(\hat{\boldsymbol{\eta}})$. The final variance estimator for $\hat{\mu}_{\text{IPW2}}$ can then be obtained by replacing various population quantities with sample-based moment estimators. Chen et al. (2020) presented the variance estimator with explicit expressions when $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ are modelled by the logistic regression and the $\hat{\boldsymbol{\alpha}}$ is obtained by the pseudo maximum likelihood method.

6.3 Variance estimation for doubly robust estimators

It turns out that variance estimation for the doubly robust estimator is a challenging problem. While double robustness is a desirable property for point estimation, it creates a dilemma for variance estimation.

The estimator $\hat{\mu}_{\mathrm{DR2}}$ given in (4.11) is consistent if either the propensity score model q or the outcome regression model ξ is correctly specified. There is no need to know which model is correctly specified, which is the most crucial part behind double robustness. This ambiguous feature, however, becomes a problem for variance estimation. The asymptotic variance formula under the model q is usually different from the one under the model $\xi,$ and consequently, it is difficult to construct a consistent variance estimator with unknown scenarios on model specifications.

There have been several strategies proposed in the literature on variance estimation for the doubly robust estimators. A naive approach is to use the variance estimator derived under the assumed propensity score model q and take the risk that such a variance estimator might have non-negligible biases under the outcome regression model. One good news is that, under the propensity score model, the estimation of the parameters $\boldsymbol{\beta}$ for the outcome regression model has no impact asymptotically on the variance of doubly robust estimators. This can be seen by using $\hat{\mu}_{\mathrm{DR1}}$ of (4.10) as an example. Let $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right),$ where $\hat{\boldsymbol{\beta}}$ is obtained based on the working model (3.1) which is not necessarily correct. Let $\boldsymbol{\beta}^*$ be the probability limit of $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$ regardless of the true outcome regression model (White, 1982). Let $m_i^* = m\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$ and $\mathbf{a}(\mathbf{x}, \boldsymbol{\beta}) = \partial m(\mathbf{x}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}.$ It can be seen that

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i - \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} = \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* - \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} + \left\{\mathbf{B}\left(\boldsymbol{\beta}^*\right)\right\}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p\left(n_A^{-1/2}\right),$$

where

$$\mathbf{B}\left(\boldsymbol{\beta}^*\right) = \frac{1}{N}\sum_{i\in S_B} d_i^B \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right) - \frac{1}{N}\sum_{i\in S_A} \frac{\mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)}{\hat{\pi}_i^A}. \tag{6.2}$$

Since the two terms on the right hand side of (6.2) are both consistent estimators of $N^{-1}\sum_{i=1}^N \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right),$ we conclude that $\mathbf{B}\left(\boldsymbol{\beta}^*\right) = o_p(1)$ and

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i - \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} = \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* - \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} + o_p\left(n_A^{-1/2}\right).$$

It follows that

$$\hat{\mu}_{\mathrm{DR1}} = \frac{1}{N}\sum_{i\in S_A} \frac{y_i - m_i^*}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* + o_p\left(n_A^{-1/2}\right).$$

The same arguments apply to $\hat{\mu}_{\mathrm{DR2}}.$ We can treat $\hat{\boldsymbol{\beta}}$ as if it is fixed in deriving the asymptotic variance for $\hat{\mu}_{\mathrm{DR1}}$ and $\hat{\mu}_{\mathrm{DR2}}$ under the assumed propensity score model. The techniques described in Section 6.2 can be directly used where the first estimating function in (6.1) is replaced by the one for defining $\hat{\mu}_{\mathrm{DR1}}$ or $\hat{\mu}_{\mathrm{DR2}}.$ See Theorem 2 of Chen et al. (2020) for further details. The variance estimator derived under the propensity score model, however, is generally biased under the outcome regression model.

Chen et al. (2020) also described a technique using the original idea presented in Kim and Haziza (2014) for the construction of the so-called doubly robust variance estimator. The technique is a delicate one with some theoretical attractiveness but has various issues for practical applications. We use $\hat{\mu}_{\text{DR1}}$ as an example to illustrate the steps for the construction of the doubly robust variance estimator. Let

$$\hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} R_i \frac{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} + \frac{1}{N} \sum_{i \in S_B} d_i^B m(\mathbf{x}_i, \boldsymbol{\beta}).$$

It follows that $\hat{\mu}_{\text{DR1}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ if $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are from the original estimation methods. The first step is to modify the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are obtained as solutions to

$$\frac{\partial \hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} = \mathbf{0} \quad \text{and} \quad \frac{\partial \hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{6.3}$$

Under the logistic regression model q where $\operatorname{logit}\{\pi(\mathbf{x}_i, \boldsymbol{\alpha})\} = \mathbf{x}_i'\boldsymbol{\alpha}$ and the linear regression model ξ where $m(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, the equation system (6.3) becomes

$$\frac{1}{N} \sum_{i=1}^{N} R_i \left\{ \frac{1}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - 1 \right\} (y_i - \mathbf{x}_i'\boldsymbol{\beta})\, \mathbf{x}_i = \mathbf{0}, \tag{6.4}$$

$$\frac{1}{N} \sum_{i=1}^{N} \frac{R_i \mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \frac{1}{N} \sum_{i \in S_B} d_i^B \mathbf{x}_i = \mathbf{0}. \tag{6.5}$$

The estimating equations in (6.5) are unbiased under the joint qp-framework. They are identical to (4.5) discussed in Section 4.1.2. The estimating equations in (6.4) are also unbiased under the outcome regression model, but they are different from the quasi score equations given in (3.2). The estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ obtained as solutions to (6.4) and (6.5) are less stable than those from standard methods. In addition, the equations system (6.4) and (6.5) will not have a solution if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not of the same dimension, since the number of equations in (6.4) is decided by the dimension of $\boldsymbol{\alpha}$ and the number of equations in (6.5) is the same as the dimension of $\boldsymbol{\beta}$. The final estimator $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ also suffers from efficiency losses when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by solving (6.4) and (6.5).

The reason behind the use of the equations system (6.3) is purely technical. It can be shown through a first order Taylor series expansion that the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ obtained from (6.3) have no impact asymptotically on the variance of $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. This technical maneuver enables that simple explicit expressions for the variance $V_{qp}(\hat{\mu}_{\text{DR}})$ under the qp framework and for the prediction variance $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ under the ξp framework can easily be obtained. Construction of the doubly robust variance estimator for $\hat{\mu}_{\text{DR}}$ starts with the plug-in estimator for $V_{qp}(\hat{\mu}_{\text{DR}})$ under the propensity scores model q. A bias-correction term is then added to obtain a valid estimator for $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ under the outcome regression model ξ. The happy ending of the story is that the bias-correction term has the analytic form $N^{-2} \sum_{i=1}^{N} (R_i / \pi_i^A - 1)\, \sigma_i^2$ where $\sigma_i^2 = E_{\xi}(y_i \mid \mathbf{x}_i)$, which is negligible under the propensity score model q. The bias-corrected variance estimator is valid under either the propensity score model or the outcome regression model.

A doubly robust variance estimator for the commonly used $\hat{\mu}_{\text{DR}2}$ is not available in the literature. A practical solution is to use bootstrap methods. Chen et al. (2022) demonstrated that standard with-replacement bootstrap procedures applied separately to S_A and S_B provide doubly robust confidence intervals using the pseudo empirical likelihood approach to non-probability survey samples when the reference sample is selected by single stage unequal probability sampling designs. Complications will arise when the probability sample S_B uses stratified multi-stage sampling methods, a known challenge for variance estimation with complex surveys. Construction of doubly robust variance estimators for the doubly robust estimator $\hat{\mu}_{\text{DR}2}$ under general settings deserves efforts in future research.

7. Assumptions revisited

Our discussions on estimation procedures for non-probability survey samples are under the assumptions A1-A4 and the focuses are on the validity and efficiency of estimators for the finite population mean under three inferential frameworks. The theoretical results on model-based prediction, inverse probability weighting and doubly robust estimation have been rigorously established under those assumptions. It seems that researchers are triumphant in dealing with the emerging area of non-probability data sources. However, as pointed out by the 2021 ASA President Robert Santos in his opinion article entitled "Using Our Superpowers to Contribute to the Public Good" (Amstat News, May 2021), "*Our superpowers are only as good as their underlying assumptions, assumptions that are all too often embraced with aplomb, yet cannot be proven*." How to check assumptions A1-A4 in practical applications of the methods is a question that can never be fully answered, and yet there are steps to follow to boost the confidence in using the theoretical results. It is also important to understand the potential consequences when certain assumptions become seriously questionable.

7.1 Assumption A1

Assumption A1 states that $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \mid \mathbf{x}_i\right)$. It is the most crucial assumption for the validity of the pseudo maximum likelihood estimator of Chen et al. (2020) and the nonparametric kernel smoothing estimator presented in Section 4.1.3 for the propensity scores, although all other assumptions are also involved. It is equivalent to the missing at random (MAR) assumption in the missing data literature. It is well understood that the MAR assumption cannot be tested using the sample data itself. The same statement holds for assumption A1 with non-probability survey samples.

In a nutshell, assumption A1 indicates that the auxiliary variables $\mathbf{x}$ included in the non-probability sample fully characterize the participation behaviour or the sample inclusion mechanism for units in the population. Sufficient attention should be given at the study design stage before data collection, if such a stage exists, to investigate potential factors and features of units which might be related to participation and sample inclusion. For human populations, the factors and features may include demographical variables, social and economic indicators, and geographical variables.

Assumption A1 leads to the conclusion that the conditional distribution of y given $\mathbf{x}$ for units in the non-probability sample is the same as the conditional distribution of y given $\mathbf{x}$ for units in the target population. It implies that the auxiliary variables $\mathbf{x}$ should include relevant predictors for the study variable y. With the given datasets S_A and S_B, sensitivity analysis through comparisons of marginal distributions and conditional models can be helpful in building confidence on assumption A1. For variables which are available in both S_A and S_B, one can compare the empirical distribution functions (or moments) from S_A to the survey weighted empirical distribution functions (or moments) from S_B. Marked differences between the two indicate that S_A is a non-probability sample with unequal propensity scores. One possible sensitivity analysis on assumption A1 is to select a variable z which has certain similarities to y, and a set of auxiliary variables $\mathbf{u}$ with both z and $\mathbf{u}$ available from S_A and S_B. We fit a conditional model $z \mid \mathbf{u}$ using data from S_A and a survey weighted conditional model $z \mid \mathbf{u}$ using data from S_B. If $\mathbf{u}$ includes all the key auxiliary variables for assumption A1, we should see the two versions of fitted models to be similar to each other. Drastic differences between the two fitted models are a strong sign that either the z is itself an important auxiliary variable for assumption A1 or the assumption is questionable.

7.2 Assumption A2

A casual look at assumption A2 may have people believe that it should easily be satisfied in practice, since a similar assumption is widely used in missing data analysis and causal inference. It turns out that the assumption can be highly problematic, and for scenarios where the assumption fails to hold, the target population is different from the one assumed for the estimation methods. It is similar to the frame undercoverage and nonresponse problems which are discussed extensively in probability sampling.

Assumption A2 states that $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$ for all i. It is equivalent to stating that every unit in the target population has a non-zero probability to be included in the non-probability sample. If the sample was taken by a probability sampling method, this would be the scenario where the sampling frame is complete and there are no hardcore nonrespondents. For most non-probability samples, the concept of "*sampling frame*" is often irrelevant or simply a convenient list, and the selection and inclusion of units for the sample may not have a structured process. In her presentation at the 2021 CANSSI-NISS Workshop, Mary Thompson pointed out that "*the statement that the sample inclusion indicator R is a random variable is itself an assumption*" for non-probability survey samples.

Let U be the set of N units for the target population. Let $U_0 = \left\{i \mid i \in U \text{ and } \pi_i^A > 0\right\}$. It is apparent that $U_0 \subset U$ and $U_0 \neq U$ when assumption A2 is violated. There are two typical scenarios in practice. The first can be termed as *stochastic undercoverage*, where the non-probability sample S_A is selected from U_0 and U_0 itself can be viewed as a random sample from U. For example, the contact list of an existing probability survey is used to approach units in the population for participation in the non-probability sample. In this case U_0 consists of units from the probability sample. Another example is a volunteer survey where the target population consists of adults in a specific city/region but the participants are recruited from visitors to major shopping centers in the region over certain period of time. The subpopulation U_0 includes visitors to the chosen locations over the sampling period and it is reasonable to assume that U_0 is a random sample from the target population. Let $D_i = 1$ if $i \in U_0$ and $D_i = 0$ otherwise, $i = 1, 2, \ldots, N.$ We have

$$P\left(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 1\right) > 0 \quad \text{and} \quad P\left(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 0\right) = 0$$

for $i = 1, 2, \ldots, N.$ If the subpopulation U_0 is formed with an underlying stochastic mechanism such that $P\left(D_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$ for all $i \in U,$ we have

$$\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 1\right) P\left(D_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$$

for $i = 1, 2, \ldots, N.$ In other words, the assumption A2 is valid under the scenario of stochastic undercoverage for non-probability samples.

The second scenario is termed as *deterministic undercoverage* where units with certain features will never be included in the non-probability sample. Suppose that participation in the non-probability survey requires internet access and a valid email address, and 20% of the population have neither access to the internet nor an email address, we have an example where the 20% of the population have zero propensity scores. There is no simple fix to the inferential procedures developed under A2. Yilin Chen's PhD dissertation at University of Waterloo (Chen, 2020) contained one chapter dealing with some specific aspects of the scenario.

7.3 Assumption A3

Among all the assumptions, this one is less crucial to the validity of the proposed inferential procedures. Under assumption A3, the full likelihood function for the propensity scores is given in (4.1). For any parametric model on $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right),$ the quasi log-likelihood function $\ell^*(\boldsymbol{\alpha})$ given in (4.2) leads to the quasi score functions $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha},$ which remains unbiased even if assumption A3 is violated. There might be some efficiency loss without assumption A3 in estimating the model parameters $\boldsymbol{\alpha}$ but the estimation methods are still valid under the other three assumptions.

7.4 Assumption A4

It is not difficult to find an existing probability sample from the same target population. It might be very hard, however, to have a probability survey sample which contains the desirable auxiliary variables. Existing probability surveys are designed with specific aims and scientific objectives, and the auxiliary variables included in the survey are not necessarily relevant to the analysis of a particular non-probability survey sample. The ultimate goal for satisfying assumption A4 is to identify and gain access to an existing probability survey sample with a rich collection of demographical variables, social and economic indicators, and geographical variables.

A rich-people's problem (when one has too much money) for assumption A4 may also occur in practice when two or more existing probability survey samples are available. How to combine all of them for more efficient analysis of non-probability survey samples is a research topic that deserves further attention. Some practical guidances on choosing one reference probability sample from available alternatives include following considerations.

(i) Check for availability of important auxiliary variables which are relevant to characterizing the participation behavor or having prediction power to the study variables in the non-probability sample;

(ii) Give first preference to the one with a larger set of variables that are common to the non-probability sample;

(iii) Assign second preference to the probability sample with a larger sample size;

(iv) And lastly, use the probability sample for which the mode of data collection is the same as the one for the non-probability sample.

It was shown by Chen et al. (2020) that two reference probability survey samples with the same set of common auxiliary variables tend to produce very similar IPW estimators but the one with a larger sample size leads to better mass imputation estimators.

8. Concluding remarks

In the early years of the 21st century, Web-based surveys started to become popular, which generated substantial amount of research interest on the topic (Tourangeau, Conrad and Couper, 2013). Issues and challenges faced by web-based and other non-probability survey samples led to the "Summary Report of the AAPOR Task Force on Non-probability Sampling" by Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013). Among other things, the report indicated that (i) unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling; (ii) making inferences for any probability or non-probability survey requires some reliance on modeling assumptions; and (iii) if non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality.

Survey sampling researchers have been answering the call with intensified explorations on statistical inference with non-probability survey samples. The current setting of two samples S_A and S_B, with the non-probability sample S_A having measurements on both the study variable y and auxiliary variables $\mathbf{x}$ and the probability sample S_B providing information on $\mathbf{x}$, was first considered by Rivers (2007) on sample matching using nearest neighbor imputation, which is the original idea leading to the mass imputation method (Kim et al., 2021). The weighted logistic regression using the pooled sample for estimating the propensity scores proposed by Valliant and Dever (2011) was the first serious attempt on the topic, which serves as a motivation for the pseudo maximum likelihood method developed by Chen et al. (2020). Brick (2015) considered compositional model inference under the same setting. Elliott and Valliant (2017) provided informed discussions on inference for non-probability samples. Yang, Kim and Song (2020) addressed issues with high dimensional data in combining probability and non-probability survey samples.

Statistical inference with non-probability survey samples is part of the more general topic on combining data from multiple sources. The term “data integration” is frequently used under this context. Combining information from independent probability survey samples has been studied extensively in the survey literature; see, for instance, Wu (2004), Kim and Rao (2012) and references therein. Inferences with samples from multiple frame surveys are another topic which has been heavily investigated by survey statisticians; see Lohr and Rao (2006) and Rao and Wu (2010a) and references therein. In her recent Waksberg award invited paper, Lohr (2021) provided an overview on multiple-frame surveys and some fascinating discussions on using a multiple-frame structure to serve as an organizing principle for other data combination methods. With emerging new data sources and reshaped views on traditional data sources such as administrative records, data integration has become a very broad area that calls for continued research. Further discussions are provided by Lohr and Raghunathan (2017) on combining survey data with other data sources and by Thompson (2019) on combining new and traditional sources in population surveys. Kim and Tam (2021) and Yang, Kim and Hwang (2021) discussed data integration by combining big data and survey sample data for finite population inference. Yang and Kim (2020) contained a review on statistical data integration in survey sampling.

One of the essential messages that the current paper conveys is the concepts of *validity* and *efficiency* in analyzing non-probability survey samples. Validity refers to the consistency of point estimators and efficiency is measured by the asymptotic variance of the point estimator. Validity is of primary concern and efficiency pursuit is a secondary goal when valid alternative approaches are available. Discussions on validity and efficiency require a suitable inferential framework and rigorous developments of statistical procedures, which is another main message from this paper. Non-probability samples do not fit into the traditional design-based or model-based inferential framework for probability survey samples. Standard statistical concepts and inferential procedures, however, can be built into a suitable framework for valid and efficient inference with non-probability survey samples.

Non-probability samples may have a very large sample size. Large sample sizes are a double-edged sword: when the inferential procedures are valid, large sample sizes lead to more efficient inference; when the estimators are biased, large sample sizes make the bias even more pronounced. A non-probability survey sample with a 80% sampling fraction over the population does not necessarily provide better estimation results than a small probability sample (Meng, 2018).

The large sample sizes also make non-probability samples connected to the modern big data problems. The role of traditional statistical methods in the era of big data was convincingly argued by Richard Lockhart (2018): "*Huge new computing resources do not put an end to the need for careful modelling, for honest assessment of uncertainty, or for good experiment design. Classical statistical ideas continue to have a crucial role to play in keeping data analysis honest, efficient, and effective.*"

Jean-François Beaumont (2020) raised the question "Are probability surveys bound to disappear for the production of official statistics?" The short answer is that probability sampling methods and probability survey samples will remain as an important data collection tool for many fields, including official statistics, and design-based inference will play a crucial role for any evolving inferential framework. The current trend of using non-probability samples and data from other sources will continue. Valid and efficient statistical inference with non-probability samples requires auxiliary information from the target population. A few high quality national probability surveys with carefully designed survey variables can play a pivotal role in analysis of non-probability survey samples.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute. An early version of the paper was presented at the SSC 2021 Annual Meeting as the Special Presidential Invited Address by the Survey Methods Section of the SSC. The author thanks the Editor of *Survey Methodology*, Jean-François Beaumont, for the invitation and for organizing the discussions on the emerging topic of statistical inference with non-probability survey samples. Thanks are also due for the two anonymous reviewers who provided constructive comments on the initial submission which led to improvements of the paper.

References

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

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

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, second edition. Wadsworth & Brooks/Cole Advanced Books & Software.

Brick, J.M. (2015). Compositional model inference. In Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 299-307.

Chen, J., and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*,16, 113-131.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Chen, J., and Sitter, R.R. (1999). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, Y. (2020). *Statistical Analysis with Non-probability Survey Samples*, PhD Dissertation, Department of Statistics and Actuarial Science, University of Waterloo.

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

Chen, Y., Li, P., Rao, J.N.K. and Wu, C. (2022). Pseudo empirical likelihood inference for non-probability survey samples. *The Canadian Journal of Statistics*, accepted.

Chu, K.C.K., and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. Proceedings of the Survey Methods Section of SSC.

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

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1212.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Kim, J.K., and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375-394.

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

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

Kim, J.K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Liu, Z., and Valliant, R. (2021). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. arXiv:2112.00855v1 [stat.ME]. Dec. 2021.

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46, 4-9.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

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

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

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, second edition, New York: Chapman and Hall.

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

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

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

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

Rao, J.N.K., and Wu, C. (2010a). Pseudo empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Rao, J.N.K., and Wu, C. (2010b). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72, 533-544.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 1-26.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*,89, 846-866

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.

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

Thompson, M.E. (1997). *Theory of Sample Surveys*. London: Chapman & Hall.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *International Statistical Review*, 87, S79-89.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

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

Wang, L., Graubard, B.I., Katki, H.A. and Li, Y. (2020). Improving external validity of epidemiologic cohort analysis: A kernel weighting approach. *Journal of the Royal Statistical Society, Series A*, 183, 1293-1311.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā A*, 26, 359-372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, 32, 15-26.

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

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

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

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

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

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

Yuan, M., Li, P. and Wu, C. (2022). Nonparametric estimation of propensity scores for non-probability survey samples. Working paper.

Zhao, P., and Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *International Statistical Review*, 87, S239-256.

Zhao, P., Rao, J.N.K. and Wu, C. (2020a). Empirical likelihood methods for public-use survey data. *Electronic Journal of Statistics*, 14, 2484-2509.

Zhao, P., Ghosh, M., Rao, J.N.K. and Wu, C. (2020b). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B*, 82, 155-174.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 313-318
Statistics Canada, Catalogue No. 12-001-X

Comments on “Statistical inference with non-probability survey samples” – Non-probability samples: An assessment and way forward

Michael A. Bailey[1]

Abstract

Non-probability surveys play an increasing role in survey research. Wu’s essay ably brings together the many tools available when assuming the non-response is conditionally independent of the study variable. In this commentary, I explore how to integrate Wu’s insights in a broader framework that encompasses the case in which non-response depends on the study variable, a case that is particularly dangerous in non-probabilistic polling.

Key Words: Survey sampling; Non-probability polls.

1. Introduction

Surveys are going through massive changes. Gone are the days of random digit dialing phone surveys producing reliably representative samples. Now hardly anyone answers the phone or even responds to emails. Pollsters have responded by coming up with a myriad of clever new ways to generate survey responses in this unwelcoming environment.

The most pervasive innovation is, without a doubt, the use of non-probability samples, often via the internet. While the implementation varies, the approach typically gathers contact information for a large number of people who are willing to respond and then involves selecting a subset from that pool for any given survey. These surveys have proven cost-effective and have often – if, perhaps, not always – produced serviceable results.

But are they believable? Most surveys do not have a ground truth against which to assess results; the lack of such information is, after all, the reason why someone is conducting the survey. Probability samples overcome this problem by relying on theory as the properties of such surveys are well understood. For non-probability samples, however, practice has vastly outpaced theory, meaning that the basis for believing the results is rather speculative.

Wu’s paper therefore is a welcome contribution to our understanding of non-probability surveys. He focuses on the class of estimators that assume ignorable non-response and puts them in context relative to each other and identifies avenues for future work.

One important point made by Wu is that “there must be a more coherent framework and accompanying set of measures for evaluating their quality” (page 305). I heartily concur. In this commentary, I expand on this point in three ways. In Section 2 I explore how to do this within the scope of the research he

[1] Michael A. Bailey, Georgetown University. E-mail: baileyma@georgetown.edu.

examines. In Section 3 I seek to expand the scope of such a framework, noting that the consequences of violations of key assumptions are so much more severe in a non-probability setting that we should build our framework to encompass violations of the key missing-at-random (MAR) assumption. In Section 4 I then explore what, if anything we can do about it. Finally, in Section 5 I provide a few concluding remarks.

2. Non-probability surveys when data is MAR

Wu grounds his analysis with a clear exposition of the four assumptions underlying the models he examines. The most important assumption is that data is MAR, meaning that given a set of covariates the study variable is independent of the decision to respond. (Although the nomenclature is standard in the literature, I cannot resist registering unease with the "missing at random" label. Of course, data is missing at random – something that is true even for MAR's opposite (and also inaptly named) missing-not-at-random (MNAR). I dream of a day when the nomenclature matches the definition, perhaps by replacing MAR with the term "conditional independence" would be a better name. However, I recognize how hard it is to change the accepted terms people use.)

Given these assumptions, Wu divides approaches into those that are model-based, inverse propensity weighting (IPW) based and double robust models. In the model-based approaches, we see the range of efforts to impute from the observed sample, including mass imputation that, broadly conceived, includes flexible sample-matching approaches that allow us to represent a larger population based on observed data points that are "close", variously defined. IPW builds on the same assumptions. Doubly robust estimators tend to be newer and attractive for their ability to give researchers two bites at the apple of relying on correct assumptions; Wu ably documents the headaches these models bring when we try to do inference with them, however.

While Wu has shown the differences in these approaches, it is useful to appreciate that he is fishing in one fairly specific corner of the pond. All models use similar information in similar ways: they all assume MAR and provide tools to model or impute the behavior of unobserved people as direct extrapolations from the observed data. If college graduates differ from non-college graduates and we have too many college graduates, all the MAR-based approaches will extrapolate to the general population directly from the data in the sample on two groups in proportion to these groups' presence in the target population.

My intuition is that the models considered by Wu are roughly equally useful – and also roughly vulnerable to violations of MAR. Or, are there contexts in which we expect the differences across the methods to be substantial? Answering this is not easy, of course, but I would be fascinated to learn Wu's perspective on where the main "action" is in non-probability samples and which of the models he considers would be best suited to accounting for such problems.

One possible focus would be on the flexibility across models. At this point, my intuition is that while these differences could be substantial in theory, in practice these differences are relatively modest. This is especially true if an experienced researcher with domain knowledge specifies a parametric model with a deft touch – including the right interactions and so forth.

3. Non-probability surveys when data is not MAR

We should take very seriously Wu's call for more coherent framework for analyzing non-probability samples. And we should aim big here as a paradigm for non-probability samples is, essentially, a paradigm for the whole field given the importance and trajectory of non-probability samples.

As we think about formulating a framework for polling it is useful to recall George Box's famous aphorism: "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad" (Box, 1976). The tiger in non-probability samples does not live between quota sampling and IPW models. The tiger can almost certainly be found instead in the MAR assumption. The violation of this assumption is the signature weakness of MAR and any framework for non-probability surveys should therefore start there.

The issue is that while MAR violations are a problem in probability sampling (arising due to non-response among the randomly contacted individuals), MAR violations are more serious in a non-probabilistic world. The idea is formalized in Meng (2018) who provides an identity for the error in a survey:

$$\overline{Y}_n - \overline{Y}_N = \underbrace{\rho_{R,Y}}_{\text{data quality}} \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{data quantity}} \underbrace{\sigma_Y}_{\text{data difficulty}} . \tag{3.1}$$

The first term in the equation is $\rho_{R,Y}$, the correlation in the population between R and Y. This quantity can be taken to reflect quality of data with regard to sampling. The second term in the Meng equation, $\sqrt{N - n/n}$, relates to the size of the population (capital N) and the size of the sample (lower case n). The third term in the Meng equation is σ_Y, the standard deviation of Y.

When $\rho_{R,Y} \neq 0$, the sampled mean will be non-zero unless $n = N$ (meaning the sample is the entire population) or $\sigma_Y = 0$ (meaning the value of Y is the same for everyone in the population), neither of which are interesting polling contexts.

This is an identity so even when the expected value of $\rho_{R,Y} = 0$ there will be some error (as in the case of random sampling). However as we move to non-random sampling we can expect the realized correlation of R and Y to grow. The larger $\rho_{R,Y}$, the larger the sampling error, the exact magnitude of which will interact with the other terms.

The most explosive implication of the Meng equation emerge from the interaction of the first two terms. When there is MNAR (meaning there will be specific reason to expect $\rho_{R,Y} \neq 0$ because R depends on Y), the actual error depends on the total population. This result is shocking to modern polling sensibilities but is vital to appreciate in the context of non-random sampling.

We can construct a simple two country world to elaborate on how this works. Suppose that our study variable is covid rates and, for the purposes of our example, that covid rates are the same in both countries. One country is huge (China, perhaps) and the other is small (Luxembourg perhaps). If we *randomly* sampled 1,000 people in each country we could produce estimates with the same precision for each country, despite their massive population differences.

What happens if we are dealing with a *non-random* sample of 1,000 people in each country? Suppose for simplicity that people's eagerness for testing is simply a function of their symptoms and that people with more symptoms are more likely to have covid. This creates MNAR sampling because opting into the sample will be associated with higher expected values of our study variable.

In China we will get the 1,000 sickest people. They will be really sick, as they will be in something like the top 0.00001 percentile. In Luxembourg we will also get the 1,000 sickest people, but you don't have to be as sick to get into this set as you would in a much bigger country. This means that the 1,000 sickest people in Luxembourg will be in roughly the top 0.2 percentile; still very sick relative to the population, but not as skewed as in China. In short, MNAR data will produce an error proportional to the population size for a given sample size.

(Note that true random samples are virtually unheard of given non-response among those who are randomly contacted. The actual practice of probability samples can be described as random contact, defined as surveys in which people are randomly contacted even as the response among those contacted may be non-random. Random contact surveys can violate MAR, but nonetheless have strong virtues. Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Bailey (2023) show how survey error in random contact surveys is proportional to the response rate rather than to the population size.)

MAR violations in non-probability sampling lead to errors that are proportional to population size. To use Box's metaphor, this is where the tigers are. Hence as we pursue Wu's exhortation for more coherence in how we evaluate new forms of polling, we should aim to agree on a framework that encompasses the possibility of MAR violations rather than a framework that assumes away this problem.

4. What to do about MAR violations?

Wu follows much of the literature in shying away from MNAR models. Part of the basis for this is a perception that MNAR non-response is essentially intractable. For example, Wu notes somewhat pessimistically that "it is well understood that the MAR assumption cannot be tested using the sample data itself" (page 302) and that "the biased nature of non-probability samples cannot be corrected by using the sample itself" (page 284).

In terms of guidance for survey researchers concerned about violations of MAR, Wu offers only a modest test, which basically consists of finding another variable that is similar to the study variable but that is available for the whole population. If only it were that easy! Generations of pollsters have scoured data for such variables and yet continue to worry about MNAR, especially when response is non-random.

Wu’s framing understates what we can do about MAR violations. These efforts will require assumptions, of course, but at least we can relax the severe assumption of MAR. The connection to the earlier points is key: since we are going to need assumptions, it is important that we have a framework for thinking about which ones are the most consequential so that we can focus our efforts appropriately. The Meng equation highlights how MAR violations play a central role in creating error in a non-probabilistic sampling world and therefore we should do whatever we can to address that issue.

A widely known example of a model that can tackle MNAR data is the Heckman (1979) selection model. This model allows for – and even estimates the magnitude of – the MAR violations. It is not without problems, of course. As a practical matter it requires an exclusion restriction (an assumption that one or more variables affect response but not the study variable) and many modern scholars are understandably cautious about the Heckman model’s strong parametric assumption.

Scholars have made considerable progress beyond the Heckman model in dealing with MAR violations (Bailey, 2023). The parametric assumption is easy to relax via copula functions (Gomes, Radice, Brenes and Marra, 2019). If we are interested in studying determinants of Y, there is a substantial and growing literature applying highly flexible control functions for MNAR contexts (Das, Newey and Vella, 2003; Liu and Yu, 2022). And if we can identify variables that affect propensity to respond but not the outcome of interest, multiple methods will model and offset MNAR sampling (Peress, 2010; Sun, Liu, Miao, Wirth, Robins and Tchetgen-Tchetgen, 2018).

5. Conclusion

Wu’s paper ably and usefully summarizes the state of the literature of analysis of non-probability survey data under the assumption of MAR. He also highlights a critical need for the field to coalesce around a more coherent framework to evaluate these and other polling innovations.

In this note, I build off Wu’s work to propose a framework that not only encompasses the MAR models analyzed by Wu, but MNAR models as well, as the violation of the MAR assumption is something particularly relevant and harmful for non-probability surveys.

References

Bailey, M.A. (2023). *Polling at a Crossroads: Rethinking Modern Survey Research*, Cambridge University Press – under contract.

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

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600, 695-700.

Das, M., Newey, W.K. and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70, 33-58.

Gomes, M., Radice, R., Brenes, J.C. and Marra, G. (2019). Copula selection models for nongaussian outcomes that are missing not at random. *Statistics in Medicine*, 38, 480-496.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-162.

Liu, R., and Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics*, 226, 321-342.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105, 1418-1430.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen-Tchetgen, E.J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28, 1965-1983.

Comments on "Statistical inference with non-probability survey samples"

Michael R. Elliott[1]

Abstract

This discussion attempts to add to Wu's review of inference from non-probability samples, as well as to highlighting aspects that are likely avenues for useful additional work. It concludes with a call for an organized stable of high-quality probability surveys that will be focused on providing adjustment information for non-probability surveys.

Key Words: Pseudo-weighting; Propensity score; Doubly-robust estimation; Sensitivity analysis.

1. Introduction

Thanks to Dr. Changbao Wu for an excellent review of the previous work and open issues for statistical inference from non-probability samples. Given the large and rapidly developing work in this area, Dr. Wu was understandably unable to cover all of it; my own understanding has blinders as well but I will touch on a few additional approaches that relate to topics he considered. I will also discuss the issue of modeling versus weighting for different inferential targets, and use his discussion and conclusions to highlight the critical importance of probability samples – in particular high-quality studies that focus on estimation of relevant covariates – to improve inference for the profusion of non-probability samples used as replacements for traditional probability samples in many research and official statistics settings. To avoid notation confusion, all notation will follow that of Wu, except where new notation is required.

Section 2 reviews additional approaches to combining data from probability and non-probability surveys. Section 3 briefly reviews the issue of weighting versus modeling when adjusting non-probability survey data. Section 4 reviews some recent developments in sensitivity analyses of standard assumptions for adjusting non-probability survey data using probability survey data. Section 5 concludes with call to systematically design a set of probability surveys with the explicit purpose of adjusting non-probability surveys.

2. Additional approaches to combining data from probability and non-probability surveys

Dr. Wu's paper follows the general prescription of 1) using model estimation and subsequent calibration to probability-sample-estimated covariate distributions, 2) developing propensity score estimates based on discrepancies between the probability- and non-probability sample data, and 3) doubly-

[1] Michael R. Elliott, Department of Biostatistics, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu.

robust methods that combine 1) and 2) in a manner such that only one of the two underlying models needs to be correct.

2.1 Propensity score estimators

Rivers (2007) appears to have been the first to suggest estimating propensity score using logistic regression with membership in the non-probability sample as the outcome and taking the reciprocal of the resulting propensity scores to use as inclusion weights. This approach was formalized further in Valliant and Dever (2011). Separately, using simple results from Bayes' theorem and discriminant analysis first described in Elliott and Davis (2005), Elliott, Resler, Flannagan and Rupp (2010) and Elliott (2013) developed a somewhat different estimator of the form

$$\hat{\pi}_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) = \hat{P}\left(i \in S_A\right) \propto P\left(i \in S_B\right) \frac{\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}{\hat{P}\left(i \in S_B \,\middle|\, i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}. \tag{2.1}$$

$\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ can be obtained using logistic regression, or using one of the suite of machine learning-type approaches such as support vector machines (Soentpiet, 1999), targeted maximum likelihood estimation (Van Der Laan and Rubin, 2006), or Bayesian Additive Regression Trees (BART) (Chipman, George and McCulloch, 2010), and $\hat{P}\left(i \in S_A \,\middle|\, i \in S_B \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ obtained as $1 - \hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$. In principle $P\left(i \in S_B\right) = 1/d_i^B$ is known since sampling probabilities are known for all elements of the population, including those in the non-probability sample, but in practice analysts with access only to public use data may have to estimate this as well. (In addition, d_i^B may include calibration and non-response adjustments that are not known for the non-probability sample elements.) This last point is critical as use of the probability sample to develop propensity scores using only the discrepancies between the non-probability sample and the probability sample will be biased unless the probability sample used an equal probability (epsem) design, as noted by Wu.

In contrast, Chen, Li and Wu (2020) shows that using a pseudo-likelihood approach to estimating $\hat{\pi}_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ directly from the population likelihood for the indicators $I\left(i \in S_A\right)$ as a function of $\mathbf{x}_i$ yields an estimator that does not require $P\left(i \in S_B\right)$ for elements in the non-probability sample under the restriction that $\pi_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ follows a generalized linear model with a canonical link, i.e., logistic regression.

(None of these approaches actually has the correct intercept to obtain a true propensity score; however, as noted in Wu, weighted estimation usually uses Hájek-type estimators [using weights to estimate a population total for denominators; Hájek, 1971] so that propensity scores estimated up to a normalizing constant are sufficient.)

2.2 Doubly-robust estimators

If inference is focused on a particular variable Y available only in the non-probability sample, we can return to the model-assisted estimators that date back to Cassel, Särndal and Wretman (1976), which posit a model for the expectation $E\left(y_i \,\middle|\, \mathbf{x}_i\right) = m_i$. Combining this with propensity score estimates of the probability of being in the non-probability sample (which we are treating as an "unknown probability sample" – more about this under Assumptions below) yields estimators of the form

$$\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}+\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i \tag{2.2}$$

corresponding to $\hat{\mu}_{\mathrm{DR}\,2}$ of (4.11) in Wu. The intuition is that any bias due to the model misspecification in estimation of m_i in $\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i$ will be equal to and opposite in sign of $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ if the model for π_i^A is correctly specified. Conversely, if the model for π_i^A is misspecified but m_i is correctly specified, $y_i-\hat{m}_i$ will be iid with mean zero and consequently $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ will also have mean 0, yielding an unbiased estimator. Chen, Valliant and Elliott (2019) used LASSO for prediction in combination with generalized regression estimators (McConville, Breidt, Lee and Moisen, 2017) when $\mathbf{X}$ is of high dimension. As Wu notes, Wu and Sitter (2001) show the equivalence between GREG applied to predicted values and DR estimators of the form in (2.2), which indicates that the Chen et al. (2019) approach was equivalent to (2.2) with LASSO estimation for m_i and an assumption of simple random sampling for the non-probability sample.

A disadvantage of using (2.1) as opposed to Chen et al. (2020) as the estimator of π_i^A, and thus of d_i^A, is the requirement that the probability sample weights d_i^B be known or at least estimated for the non-probability sample. An advantage of using (2.1), is that non-linear models and machine learning methods can be used in estimation. Rafei, Flannagan and Elliott (2020) uses BART to estimate both m_i and π_i^A, reducing the impact of potential model misspecification. Simulations showed considerable improvement in bias and variance reduction over the method of Chen et al. (2020) when the linear models is misspecified. Variance estimation can proceed by adapting Rubin's multiple imputation rules: from M independent draws from BART, the mean of the variances computed treating the draw of d_i^A as known using standard complex sample design estimators and added to $\frac{M+1}{M}$ times the variance of the point estimates computed across the draws of d_i^A yield an approximately unbiased variance estimator.

An alternative approach to doubly-robust estimation uses the fact that the propensity score is the coarsest possible "balancing score" that contains all of the information about the association between the sampling indicator and the outcome of interest. This has led to the development of mean estimators that use smooth functions of weights to produce consistent estimators that can be more efficient when weights are highly variable or only weakly related to the outcome (Elliott and Little, 2000; Zheng and Little, 2005). Zhou, Elliott and Little (2019) extended this idea into the causal inference setting in non-randomized settings, in which probability of assignment to a treatment or exposure (propensity score) is estimated as a function of covariates $P_Z(\mathbf{x}_i,\boldsymbol{\alpha})$ using logistic regression, and then non-observed potential outcomes Y^z under treatment arm $z_i'\neq z_i$ for observed treatment z_i are imputed from

$$Y_i^Z\sim N\Big(s\Big(\hat{P}_Z^*(\mathbf{x}_i,\hat{\boldsymbol{\alpha}})\,|\,\boldsymbol{\theta}_Z\Big)\Big)+g_Z\Big(\hat{P}^*(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}),\mathbf{x}_i\,|\,\boldsymbol{\beta}_Z\Big),\sigma^2 \tag{2.3}$$

where P^* is the logit transformation of P, $s(\hat{P}_Z^* \,|\, \boldsymbol{\theta}_Z)$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996) of propensity, and $g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ is a general function of covariates including the propensity scores. The resulting estimator is doubly robust in the sense that if either $P_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ or $E(Y^z) = g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ is correctly specified, $Y^{(z)}$ will be approximately unbiased; see Zhang and Little (2009). This can be implemented in the non-probability setting by replacing $\hat{P}_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ in the mean model for (2.3) with $\hat{\pi}_i^A$ estimated using (2.1) to obtain a draw of $Y_i^{(b)}$. (Note this requires obtaining $\hat{\pi}_i^A$ for the probability sample elements requiring prediction.) Inference can proceed by obtaining $b = 1, \ldots, B$ draws from the posterior distribution of the estimated population quantity of interest, e.g., for the population mean

$$Y^{(b)} = \frac{\sum_{i \in S_R} N_i^{(b)} Y_i^{(b)} + \sum_{i \in S_A} \left(y_i - Y_i^{(b)} \right)}{N}$$

where now $N_i^{(b)}$ is a estimate of the population represented by the weight d_i^R obtained from a finite population Bayesian bootstrap (Little and Zheng, 2007); more complete FBPP extensions to complex sample designs that include clustering and stratification are available in Dong, Elliott and Raghunathan (2014).

As in the estimation of (2.1), the non-parametric (spline) component of (2.3) can be replaced with other machine-learning estimators; see Chapter 4 of Rafei (2021) for implementation using Gaussian processes. Also, extensions to non-normal models are direct, although not necessarily computational easy.

2.3 Poststratified estimators

Wu also describes the use of poststratified estimators in the context of quota sampling, which is not only a very old form of non-probability sampling but indeed the standard before Neyman made the case for stratified random sampling (Neyman, 1934). Wu's Section 5 suggests a robust alternative to the propensity score estimates obtained by ordering observations in the probability sample by $\hat{\pi}_i$, stratifying into K strata based on this ordering, and computing the predicted proportion of the population belonging to the k^{th} stratum as proportion of the sample weights W_k in this stratum using the probability sample, with

$$\hat{\mu}_{\text{PST}} = \sum_k \hat{W}_k \bar{y}_k \tag{2.4}$$

where $\bar{y}_k$ is the mean within the k^{th} stratum in the non-probability sample. Wu notes the tradeoff between choosing K to be large enough to retain homogeneity within units but small enough to obtain stable estimates of $\bar{y}_k$, suggesting 30 as the old "rule of thumb" for "large [enough] sample sizes". I would add that a more formal approach discussed in Little (1986) suggests a method to generate strata (there in the context of non-response adjustment) that minimizes mean square error by maximizing the between-stratum-to-within-stratum variance. It would seem such an approach would be appropriate to consider in the non-probability post-stratified estimator as well.

A more direct approach to obtain estimates using a post-stratified type estimator is multilevel regression and poststratification (Wang, Rothschild, Goel and Gelman, 2015; Downes and Carlin, 2020). Here only data from the non-probability sample is used in the outcome model:

$$E\left(Y_{k[i]}\right) = \beta_0 + \mathbf{x}_k^T \boldsymbol{\beta} + \sum_j a_{l[k]}^j \tag{2.5}$$

where $k = 1, \ldots, K$ indexes the poststratum developed from $j = 1, \ldots, J$ variables, $a_{l[k]}^j \sim N(0, \sigma_j^2)$ for $l = 1, \ldots, L_j$ and $l[k]$ maps the postratum cell k to the appropriate category l of variable j. The poststratifed estimator is still given by (2.4) with $\hat{W}_k$ now replaced with known population totals W_k; posterior inference is obtained though posterior draws of β_0, $\boldsymbol{\beta}$, and $a_{l[k]}^j$ to obtain a draw of

$$\hat{\mu}_{\text{PST}}^{(b)} = \sum_k W_k \left[\frac{1}{n_k} \sum_{i \in k} \left(\beta_0^{(b)} + \mathbf{x}_k^T \boldsymbol{\beta}^{(b)} + \sum_j a_{l[k]}^{j(b)}\right)\right].$$

Though not technically doubly-robust, it has been shown to work well in some applications where J is large enough to capture all of the important discrepancies between the probability and non-probability sample, and the non-probability sample is sufficiently large to allow reasonably accurate estimation of $a_{l[k]}^j$. In the absence of known joint distributions of a high dimensional $\mathbf{X}$, this approach has the weakness of relying on estimated distributions, which are unstable. A possible alternative might be replace the simple $\bar{y}_k$ with (2.5) in Wu's poststratified estimator (2.4), using the fact that the sampling weights d_i^R summarize the information about $\mathbf{X}$ in the probability sample similar to that of the propensity score for non-probability sample.

3. Weighting vs. modeling for the general user

Wu's paper and the above addendums tend to follow the long-trodden path regarding weighting versus modeling in the finite population inference setting, dating back at least to Hansen, Madow and Tepping (1983). In thinking about this choice I believe it is important to distinguish between models used to derive so-called descriptive parameters – in the sense of Kalton (1983) – and models that are of interest in and of themselves, so-called analytic parameters in regression models, latent classes analysis, etc. For the former distinguishing a descriptive target of interest Y from potential modeling covariates $\mathbf{X}$ has the advantage of creating doubly-robust estimators that are targeted to a single descriptive parameter. This also requires assumptions such as A1 in Section 2.1 (propensity score does not depend on Y conditional on $\mathbf{X}$). When models themselves are the targets of interest, it may be that developing weights via propensity scores to account for selection bias and, as Wu notes, employing standard weighted estimating equations may be the most sensible choice, since typically a wide number of models may be considered. This comes at the cost of double robustness, since there is usually no attempt to model the analytic parameter directly. Developing ways to extend double-robustness into a broader class of model parameter estimates may be a fruitful exercise.

4. Unverifiable assumptions: Recent developments in sensitivity analysis

Wu provides four key assumptions required to correct for selection bias in non-probability surveys using data from probability surveys: they can be roughly summarized as “selection at random” or SAR (covariates in the non-probability sample explain the probability of selection in the non-probability sample); “positivity” (all elements in the population have a non-zero probability of selection into the non-probability sample); “independence” (elements are selected independently into the non-probability sample); and “common covariates” (there exists a probability survey with covariates whose subset matched the covariates required for the MAR assumption to hold). It might be worth noting that the first two assumptions basically require the non-probability survey to be a probability survey “in disguise” – that is, there really are non-zero probabilities of selection into the non-probability survey for all elements in the population, but we as analysts just do not know what they are.

In practice neither of these assumptions probably hold precisely. Some recent work has focused on the failure of the first, the SAR assumption. Some existing measures borrowed from the non-response literature have been repurposed here: for example, the R-indicator measure (Schouten, Cobben and Bethlehem, 2009), which in this context is the measure of the variability in the probabilities of selection in the non-probability sample:

$$\hat{R} = 1 - 2\sqrt{\frac{1}{n_a - 1}\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A - \sum_{j=1}^{n_A}\hat{\pi}_j^A \big/ n_a\right)^2}$$

$\hat{R}$ can range between 0 and 1, where 1 is achieved when probabilities of selection are constant – suggesting something akin to a simple random sample, with less chance for selection bias – and 0 – suggesting all elements are either included with probability 1 or 0, maximizing the risk of selection bias.

Of course, in the absence of the outcome Y in the probability sample, there is no way to directly assess selection bias. Hence recent work has extended Andridge and Little (2011), which develops a sensitivity analysis using a pattern-mixture model, wherein selection into non-probability sample is allowed to depend entirely on a scalar reduction to the covariates $\mathbf{X}$, entirely on the outcome Y, or some convex combination thereof. Little, West, Boonstra and Hu (2020), Andridge, West, Little, Boonstra and Alvarado-Leiton (2019), and West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton (2021) consider sensitivity to this assumption in the estimation of the mean of a normally distributed variable, the mean of a binary outcome, and the regression parameters in a linear regression model, respectively, in non-probability samples. By varying the convex mixing parameter ϕ, sensitivity to the SAR assumption can be assessed. Boonstra, Little, West, Andridge and Alvarado-Leiton (2021) finds that these "standard measures of bias" (SMB) compare favorably with alternatives such as $\hat{R}$ in a simulation study. An important point to note is that the methods that extend Andridge and Little (2011) do not depend on assumption of common covariates in a probability sample. This suggests that methods that use information available in the probability sample to assess SAR are an open area for development.

The second assumption – positivity – is also unlikely to exist precisely in many practical settings. My own work in this area has focused on naturalistic driving studies, which typically involve convenience samples in a limited geographical area: for example, the Second Strategic Highways Research Program (SHRP2) recruited drivers in six specific geographic regions across the United States (Transportation Research Board (TRB) of the National Academy of Sciences, 2013). This corresponds to the second scenario given by Wu in Section 7.2, where only a subpopulation has any chance of being selected into the non-probability sample, which as he notes has "no simple fix". Following his notation of D providing an indicator of membership in the subpopulation, it would seem that if $D_i \perp \mathbf{X}_i, Y_i \mid \pi_i^A$ – that is, if the distribution of $\mathbf{X}, Y$ is the same for $D = 0$ and $D = 1$ after weighting for π_i^A within the $D = 1$ stratum – then lack of positivity would have no impact on inference. This is likely a tall order in the most general settings but might be reasonably well approximated if the analysis of interest involves a subset of $\mathbf{X}, Y$ that is only weakly associated with D even before adjustment.

Finally, regarding the fourth assumption – existence of a probability sample with available $\mathbf{X}$ – I very much second Wu's observation that methods to take advantage of multiple probability surveys need more development. However, it remains more likely that a researcher will struggle to find a single probability sample with sufficient covariates than struggle with a surfeit of options (Wu's "rich person's problem"). To this end I will conclude with a call to action by the survey community.

5. Probability sampling in the 21st century: Now more than ever

I learned statistics, and particularly survey statistics, near the end of the 20th century, when probability sampling was the unchallenged touchstone of survey design. I was first introduced to the problem of making inference from non-probability samples in the late 00's in the context of injury analysis using Crash Injury Research (CIREN) data, where analysts were treating a highly-restricted sample of individuals in passenger vehicle crashes as if they were a random sample of crash victims and consequently finding non-sensible results (Elliott et al., 2010). About the same time web surveys were exploding in popularity and survey statisticians were somewhat at a loss as to how to make inference from such data. I will admit to a rather paternalistic attitude at the time – I almost avoided trying to do research in this area because I thought it would only encourage "bad behavior" regarding sample design. I did not think I could single-handedly stop it, but I did not want to participate in what I perceived as the downgrading of science. I came to recognize, however, that many of these new data sources have advantages beyond what can be achieved through the traditional probability sample, certainly within limited budgets. This is above and beyond the increasing challenges to implementing probability surveys, especially in general populations, due to non-response, lack of adequate sampling frames, etc.

However, I remain concerned that the idea that we have developed methods to deal with the limitations of non-probability surveys means that probability sampling is passe is becoming entrenched among scientists and policy makers with limited statistical training, despite efforts like those of Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Marek, Tervo-Clemmens, Calabro et al. (2022). However, as Wu's review notes, the absence of probability samples unmoors the non-probability sample from the possibility of even partial calibration or other adjustment approaches (although sensitivity analyses such as those SMB approaches noted above do not require benchmarking probability samples). Hence I believe it is increasingly critical for an organized and ideally government funded stable of high-quality probability surveys to be put into place for routine data collection. Some of these obviously already exist – the US Census' American Community Survey and the National Center for Health Statistics National Health Interview Survey premier among them – but going forward I believe it would be valuable for statistical agencies to explicitly coordinate around the need for high quality probability surveys to serve a role as analytic partners to the non-probability survey world rather than just as stand-alone products. This means thinking carefully about important covariates across a variety of public health and social science roles in which survey data play a role. Choices will have to be made given limited budget constraints, and at the same time provisions should be made for sufficient funding to retain the quality needed for adjustment. Finally, while some methods do not require microdata and thus can use summary measures such as those avaiable in the American Communities Survey, other will require such data, which likely means new areas of research to be explored in the fields of privacy and confidentiality research as applied to the combining of data from probability and non-probability surveys.

References

Andridge, R.R., and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society*, C68, 1465-1483.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.

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

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

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

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266-298.

Dong, Q., Elliott, M.R. and Raghunathan, T.E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40, 1, 29-46. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14003-eng.pdf.

Downes, M., and Carlin, J.B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, 62, 479-491.

Eilers, P.H., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Elliott, M.R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. *Journal of the Royal Statistical Society*, C54, 595-609.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Elliott, M.R., Resler, A., Flannagan, C.A. and Rupp, J.D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference*, 236.

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

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175-188.

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

Little, R.J.A., and Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8, 1-20.

Little, R.J.A., West, B.T., Boonstra, P.S. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.

Marek, S., Tervo-Clemmens, B., Calabro, F.J. et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, in press.

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

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

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. University of Michigan Thesis. Accessible at https://www.overleaf.com/project/6228db145a47be05f8da3777.

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

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*. Available at https://static.texastribune.org/media/documents/Rivers_matching4.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.

Soentpiet, R. (1999). *Advances in Kernel Methods: Support Vector Learning*. Boston: MIT Press.

Transportation Research Board of the National Academy of Sciences (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.*

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

Van Der Laan, M.J., and Rubin, D.R. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 980-991.

West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

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

Zhang, G., and Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 911-918.

Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Zhou, T., Elliott, M.R., and Little, R.J.A. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114, 1-19.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 331-338
Statistics Canada, Catalogue No. 12-001-X

Comments on “Statistical inference with non-probability survey samples”

Sharon L. Lohr[1]

Abstract

Strong assumptions are required to make inferences about a finite population from a nonprobability sample. Statistics from a nonprobability sample should be accompanied by evidence that the assumptions are met and that point estimates and confidence intervals are fit for use. I describe some diagnostics that can be used to assess the model assumptions, and discuss issues to consider when deciding whether to use data from a nonprobability sample.

Key Words: Convenience sample; Diagnostics; Imputation; Probability sample; Survey quality; Survey weights.

1. Introduction

Many thanks to Changbao Wu for his stimulating review and assessment of methods for making inferences from nonprobability samples. I especially appreciate his thoughtful examination of the strong assumptions needed to derive the bias and variance of estimates.

Wu reviews three approaches for estimating the finite population mean μ_y of a variable y that is measured in a nonprobability sample S_A of size n_A. Because this sample is not representative of the population (and hence the sample mean $\bar{y}_A$ is likely biased for estimating μ_y), each approach relies on information from a high-quality probability sample S_B of size n_B: S_B does not measure y but it contains a set of auxiliary variables $\mathbf{x}$ that are also observed in S_A.

In the model-based predictive approach, a model is developed on S_A to predict y from $\mathbf{x}$. The mass imputation (MI) estimator, for example, uses the model to impute an estimate y_i^* of y_i for every member of the probability sample S_B. Then the population total of y is estimated by $\sum_{i \in S_B} d_i^B y_i^*$ where d_i^B is the design weight of unit i in S_B.

In the inverse propensity weighting (IPW) approach, a model is developed predicting the probability π_i^A that population unit i appears in S_A as a function of $\mathbf{x}$. Then unit i in S_A is assigned weight $w_i^A = 1 / \hat{\pi}_i^A$ and the population total is estimated by $\sum_{i \in S_A} w_i^A y_i$.

Wu also reviews a “doubly robust” estimator of μ_y that, by combining the predictive and IPW estimators, is approximately unbiased under the assumptions if either model is correctly specified. In this discussion, I will concentrate on the predictive and IPW approaches because these methods generalize more easily for multivariate analyses and estimating population characteristics other than means.

In Section 2, I explore assumptions needed for inference from nonprobability samples and diagnostics for assessing them. Then, in Section 3, I look at some questions to ask when deciding which approach (if any) to use for inference.

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

2. Model assumptions and diagnostics

Probability sampling gained widespread use after the theory was developed in the 1930s and 1940s because it provided a mathematically justified solution to the problem of how to generalize from a sample to a population. Under minimal assumptions, a full-response probability sample produces approximately unbiased estimates of population quantities, accompanied by confidence intervals that have approximately correct coverage probabilities. It is the *only* method that is guaranteed to produce accurate confidence intervals without making assumptions about the unsampled members of the population. A probability sample is representative because of the procedure by which it is drawn.

All other methods require huge assumptions. The major assumptions for the predictive and IPW methods, given in Section 2.1 of Wu's article, are: (A1) y and the random variable indicating participation in S_A are independent given $\mathbf{x}$, (A2) every unit in the population has $\pi_i^A > 0$, and (A3) the random variables indicating participation in S_A are independent given $\mathbf{x}$. These assumptions imply that the auxiliary information $\mathbf{x}$ is rich enough to develop inverse propensity weights that remove selection bias for y, and that a model developed on S_A to predict y from $\mathbf{x}$ will also apply to units not in S_A.

Statistical properties of the estimators are developed assuming that (A1) - (A3) are true and that the models adopted for weighting or imputation are correctly specified. Under those conditions, the estimated population mean is approximately unbiased with variance given by the appropriate theorem. But, as Wu points out, that variance estimate is conditional on the assumptions being satisfied; if the assumptions are not met, it will severely underestimate the true mean squared error and give a misleading impression of the estimate's trustworthiness. If n_A and n_B are large but (A1) is violated, the bias might be 10 percentage points but the reported standard error of an MI or IPW estimate will be close to zero. In practice, many nonprobability samples will violate the assumptions: Mercer, Lau and Kennedy (2018) found, when weighting online opt-in samples with rich auxiliary information, that "even the most effective adjustment strategy was only able to remove about 30% of the original bias".

The assumptions cannot be fully tested because they involve missing data – population members missing from S_A and y values missing from S_B. But, as with nonresponse adjustments in probability samples (Lohr, 2022, Chapter 8), one can perform model checks and diagnostics using available information, with the recognition that these might not catch all model deficiencies.

Compare statistics from the nonprobability sample with those from other data sources

Wu suggests comparing empirical distribution functions of variables in $\mathbf{x}$ from S_A with the survey-weighted empirical distribution functions from S_B. Differences may indicate that observations in S_A have unequal propensity scores or that the $\mathbf{x}$ variables are measured differently in S_A than in S_B (see Section 3). One can also compare empirical distributions from S_A with those from another probability survey S_C.

If IPW is used, one can also compare propensity-score-weighted empirical distribution functions from S_A with those from S_B and other surveys. This should be done only for variables not used in the weighting, since the propensity score weights have already adjusted for imbalances in weighting variables. Dutwin and Buskirk (2017), for example, constructed propensity weights for a nonprobability sample through raking on marginal totals and then compared the cross-tabulations of those raking variables.

Wu also suggests treating a variable z that is measured in both S_A and S_B as a response variable, and comparing conditional models for $z \mid \mathbf{u}$ fitted on S_A and S_B, where $\mathbf{u}$ is a subset of $\mathbf{x}$ (excluding z). Differences in the two models can indicate that z is needed as an auxiliary variable, and may also raise questions of how well the set of measured auxiliary variables satisfy assumption (A1).

In an example from Kim, Park, Chen and Wu (2021), the estimated percentage of persons who volunteer was 24.8% from the Current Population Survey (the gold-standard estimate), but the MI and IPW estimates from S_A were both close to 50% with reported standard error less than one percentage point. The standard error, computed under the model assumptions, did not account for the selection bias of S_A with respect to volunteerism – a bias that could not be removed using demographics, home ownership, and medical insurance as model covariates.

Compare results from the IPW and MI approaches

An alternative to using the doubly robust estimator for analysis is to use each model to identify potential deficiencies of the other. Possible investigations include comparing the empirical distribution of y from S_A (using the inverse propensity weights) with the empirical distribution of y^* from S_B (using the imputed values and the survey weights). Similarly, as suggested by Chipperfield, Chessman and Lim (2012), one can compare estimated domain means from S_A and S_B for a set of domains $d = 1, \ldots, D.$ One might also compare imputations for y fit to the unweighted data set S_A with imputations developed on S_A with inverse propensity weights.

Simulation studies are valuable for checking the small-sample behavior when the assumptions are met, but are of limited value for exploring sensitivity to model assumptions. These explore model deviations devised by the investigators, but real surveys can diverge from the model in many unanticipated ways.

Perform model diagnostics

Of course, for either the IPW or model-predictive approach, analysts should employ standard regression diagnostics such as examining residuals and influential observations to examine model fit and sensitivity to outliers, and document the checks that were done.

For the IPW approach, it is also desirable to examine characteristics of the final weights. The coefficient of variation of the weights provides a rough measure of the amount of adjustments that were needed to make sample S_A “representative”. A low coefficient of variation, however, does not necessarily mean the sample is representative; this may merely reflect inadequacy of the available auxiliary information for developing weights. For example, suppose a quota sample from an opt-in internet panel is drawn to match the population with respect to the auxiliary variables. The inverse propensity weights will have little variation because the $\mathbf{x}$ variables were used to form the quota classes, but the sample may still produce biased estimates of y variables such as internet usage or volunteering.

The graphical methods proposed by Makela, Si and Gelman (2014) for assessing weight adjustments in surveys can be used with IPW as well. Brick (2015) suggested looking at the magnitude of the IPW adjustments in the weighting cells. One can also examine the distribution of the weights within domains of interest.

The inverse propensity weights can also provide information about assumption (A2). A domain that has high weights relative to other domains may have undercoverage in S_A. Dever (2018) proposed investigating assumption (A2) by identifying individuals in S_B who have no close match in S_A.

Bondarenko and Raghunathan (2016) reviewed and proposed graphical and numerical diagnostic tools for assessing and improving imputation models. None of these diagnostics, however, will test the assumption that the regression model fit on S_A applies to units not in S_A. Just as $\bar{y}_A$ may be a biased estimator of μ_y, regression coefficients derived from S_A may also be biased, and the model constructed from S_A to predict y from $\mathbf{x}$ might not apply to other parts of the population.

Take a small probability sample to investigate assumptions

The preceding steps can identify some model deficiencies, but cannot fully test assumptions (A1) and (A2). But one can test the imputation model by obtaining data about y on a probability subsample of S_B. Similarly, one could take a probability sample from population members not in S_A to check inferences from the IPW approach, or observe y on a subsample of units in S_B that are similar to those with high weights in S_A, or that have no close match in S_A.

3. When should one use nonprobability samples?

Wu describes methods for combining information from probability and nonprobability samples after the decision has been made to do so. A first question, however, is whether the operation should be done at all. It may be desired to use a nonprobability sample because no high-quality probability sample measures y, and it is thought that “any information is better than no information”. But is that true?

Suppose that, despite the careful model-fitting and model-checking, key statistics are still biased. Could reporting a flawed statistic be worse than reporting no statistic? Bad statistics, once published, can circulate for a long time – even after more rigorous studies show that they are biased. In 1975, advice columnist Ann Landers asked her readers to respond to the question “If you had it to do over again, would you have children?” About 70% of the 10,000 persons who mailed a response said they would not have children in a do-over. This statistic is still cited, even though it is from a convenience sample, has been contradicted by numerous other studies, and is nearly 50 years old (Lohr, 2022). It is also unlikely that predictive modeling or IPW would have corrected the selection bias affecting Landers’ statistic, which occurred within all demographic groups.

With these issues in mind, here are some questions that could be asked when deciding whether to use estimates from a nonprobability sample and, if so, which statistical method to use for making inferences.

- How will the statistics be used? Estimates from the nonprobability sample might serve well for developing a marketing strategy or for an exploratory sociological study, but might not be deemed reliable enough for estimating unemployment or the number of persons requiring food assistance. Statistics from a nonprobability sample should be accompanied by evidence that the estimates are fit for use.
- What is the quality of the data in S_A? Administrative records such as tax records have a different quality profile than a survey of volunteers recruited through an internet advertisement.

  If the population for S_A is well-defined (for example, tax filers), it may be better to report statistics for that population than to attempt to generalize to the population of S_B. For tax records, many persons below preset income thresholds have $\pi_i^A = 0$ and assumption (A2) is violated. Instead, a multiple-frame approach might be adopted, where a different data source is used to estimate μ_y for the parts of the population not in S_A (Lohr, 2021).

  Since all of the models rely on auxiliary information $\mathbf{x}$, it is important to have S_A and S_B measure the $\mathbf{x}$ variables the same way. If income is used as an auxiliary variable, the same questions should be used to define income in both surveys, and income should be measured for the same unit (person or household).

  Kennedy (2022) suggested that some respondents to opt-in online surveys may provide incorrect demographic information or bogus answers to questions; if that occurs, model predictions will be flawed. It may even be possible for outsiders desiring a specific outcome to manipulate the data in S_A – for example, an organization might arrange for the survey to be taken by a set of volunteers whose claimed demographic characteristics match those of the population but who give the “desired” answer for y. Some proponents of nonprobability samples argue that low-response-rate probability samples also require weighting adjustments or imputation, but there is one important difference: the probability survey may have nonresponse, but the initial sample is selected randomly and cannot be manipulated by outside organizations.

  If the data in S_A are low-quality, is it worth spending the time to construct models? As Louis (2016) said, “Space-age procedures will not rescue stone-age data”.

- How detailed is the auxiliary information? If S_A is large, and the auxiliary information is specific enough to be able to identify specific records, then linking records between S_A and S_B would be a better method for combining the data. Imputation or IPW would be used if the auxiliary information $\mathbf{x}$ is rich enough to give good predictions of y_i or π_i^A, but not rich enough to permit accurate linkage. If there is little auxiliary information, however, then one would expect low variation in the propensity scores or imputed values, and the methods may give poor predictions – with little information to diagnose potential problems.
- What analyses are desired? Wu discusses estimating the population mean, but the analyst may also want to look at relationships between y and other variables, or estimate means or medians

for subgroups. The choice of method depends in part on the variables that are available in S_A and S_B. If S_A contains many response variables whose relationship is of interest, the IPW approach might be preferred.

If it is desired to explore relationships between y and variables measured only in S_B, imputation might be a better choice. Here, though, the analyst should be careful to acknowledge the imputation when presenting results – if, say, linear regression is used for the imputation, the correlation calculated on S_B is not between variable u and variable y, but between u and $\mathbf{x}^T\hat{\boldsymbol{\beta}}$.

- What are the implications for data equity? Jagadish, Stoyanovich and Howe (2021) defined "representation equity" as "increasing the visibility of underrepresented groups that have been historically disadvantaged or suppressed in the data record".

  Nonprobability samples have the potential to improve data equity. They can increase the sample size and visibility of rare population subgroups – a large data set S_A might contain 10,000 members of the subgroup, while even a full-response probability survey with $n_B = 60{,}000$ might contain only ten. Or the nonprobability sample may contain population members who are underrepresented in the probability survey because they are out of scope, undercovered in the sampling frame, or prone to nonresponse. In these situations, S_A provides information about groups that are not as well represented in the probability survey.

  On the other hand, historically disadvantaged groups may be underrepresented in all data sources, including S_A. For example, a large nonprobability sample of electronic health records will be able to generate estimates for more population subgroups than a small probability sample about health. But persons without health insurance or access to medical care are underrepresented. In this situation, relying on S_A to produce population estimates may reinforce inequities. If the estimates are used to distribute resources, then, as the program is implemented, more data will be collected in the areas getting those resources and will validate their needs, but no such follow-up will be done in areas that are inaccurately determined to receive no resources. The feedback loop will propagate the inequitable representation in data sources.

  The MI and IPW methods have different data equity implications. Imputation assigns a predicted value of y to each observation in S_B, and the imputed y value may differ from the y value the respondent would have supplied if asked – particularly if the respondent is in a subgroup that is unrepresented or misrepresented in S_A. Will the model give accurate predictions for historically underrepresented subgroups? Did the respondents to S_B give informed consent for y to be imputed?

  IPW assumes that the propensity scores can be estimated from auxiliary information. Is that information rich enough to give accurate weights? Are some subgroups unrepresented in S_A? It may be useful to compare the results from the two methods, and from other data sources if available, for historically underrepresented population subgroups.

Wu's critical review raises many important issues for persons interested in using nonprobability samples to make inferences about the population. I especially appreciate his assessment of the strong assumptions needed for the model-based methods, and applaud the emphasis on addressing these problems during the survey design stage.

References

Bondarenko, I., and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, 35(17), 3007-3020.

Brick, J.M. (2015). Compositional model inference. In *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 299-307.

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

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

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81(S1), 213-239.

Jagadish, H.V., Stoyanovich, J. and Howe, B. (2021). COVID–19 brings data equity challenges to the fore. *Digital Government: Research and Practice*, 2(2), 1-7.

Kennedy, C. (2022). Exploring the assumption that online opt-in respondents are answering in good faith. Paper presented at the 2022 Morris Hansen Lecture, March 1, 2022.

Kim, J.-K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

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

Louis, T.A. (2016). Discussion of combining information from survey and non-survey data sources: Challenges and opportunities. 130th CNSTAT Meeting Public Seminar; Washington, DC. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_172505.pdf.

Makela, S., Si, Y. and Gelman, A. (2014). Statistical graphics for survey weights. *Revista Colombiana de Estadística,* 37(2), 285-295.

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

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 339-360
Statistics Canada, Catalogue No. 12-001-X

Comments on “Statistical inference with non-probability survey samples” – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples

Xiao-Li Meng[1]

Abstract

Non-probability samples are deprived of the powerful *design probability* for randomization-based inference. This deprivation, however, encourages us to take advantage of a natural *divine probability* that comes with any finite population. A key metric from this perspective is the *data defect correlation (ddc)*, which is the model-free finite-population correlation between the individual’s sample inclusion indicator and the individual’s attribute being sampled. A data generating mechanism is equivalent to a probability sampling, in terms of design effect, if and only if its corresponding *ddc* is of $N^{-1/2}$ (stochastic) order, where N is the population size (Meng, 2018). Consequently, existing valid linear estimation methods for non-probability samples can be recast as various strategies to miniaturize the *ddc* down to the $N^{-1/2}$ order. The quasi design-based methods accomplish this task by diminishing the variability among the N inclusion propensities via weighting. The super-population model-based approach achieves the same goal through reducing the variability of the N individual attributes by replacing them with their residuals from a regression model. The doubly robust estimators enjoy their celebrated property because a correlation is zero whenever one of the variables being correlated is constant, regardless of which one. Understanding the commonality of these methods through *ddc* also helps us see clearly the possibility of “double-plus robustness”: a valid estimation without relying on the full validity of either the regression model or the estimated inclusion propensity, neither of which is guaranteed because both rely on *device probability*. The insight generated by *ddc* also suggests *counterbalancing sub-sampling*, a strategy aimed at creating a miniature of the population out of a non-probability sample, and with favorable quality-quantity trade-off because mean-squared errors are much more sensitive to *ddc* than to the sample size, especially for large populations.

Key Words: Data defect index; Design probability; Divine probability; Device probability; Design-based inference; Model-assisted survey estimators; Non-response bias.

1. Distinguish among design, divine, and device probabilities

1.1 What can statistics/statisticians say about non-probability samples?

Dealing with non-probability samples is a delicate business, especially for statisticians. Those who believe statistics is all about probabilistic reasoning and inference may question if statistics has anything useful to offer to the non-probabilistic world. Whereas such questioning may reflect the inquirers’ ignorance about or even hostility towards statistics, taking the question conceptually, it deserves statisticians’ introspection and extrospection. What kind of probabilities are we referring to when the sample is non-probabilistic? The entire probabilistic sampling theory and methods are built upon the randomness introduced by powerful sampling mechanisms, which then yields the beautiful designed-based inferential framework without having to *conceive* anything else is random (Kish, 1965; Wu and Thompson, 2020; Lohr, 2021). When that power – and beauty – is taken away from us, what’s left for statisticians?

[1] Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, MA 02138. E-mail: meng@stat.harvard.edu.

A philosophical answer by some statisticians would be to dismiss the question altogether by declaring that there is no such thing as probability sample in real life. (I was reminded by Andrew Gelman about this sentiment when I sought his comments on this discussion article. See https://statmodeling.stat.columbia.edu/2014/08/06 for a related discussion.) By the time the data arrive at our desk or disk, even the most carefully designed probability sampling scheme would be compromised by the imperfections in execution, from (uncontrollable) defects in sampling frames to non-responses at various stages and to measurement errors in the responses. In this sense, the notion of probability sample is always a theoretical one, much like efficient market theory in economics, which offers a mathematically elegant framework for idealization and for approximations, but should never be taken literally (e.g., Lo, 2017).

The timely article by Professor Changbao Wu (Wu, 2022) provides a more practical answer, by showcasing how statisticians have dealt with non-probability samples in the long literature of sample surveys and (of course) observational studies, especially for causal inference; see Elliott and Valliant (2017) and Zhang (2019) for two complementary overviews addressing the same challenge. To better understand how probability theory is useful for non-probability samples, it is important to recognize (at least) three types of probabilistic constructs for statistical inference, as listed in Section 1.2. Non-probability samples take away only one of the three, and as a result, they typically force a stronger reliance on the other two.

With these conceptual issues clarified, the rest sections discuss a unified strategy for dealing with non-probability samples. Section 2 reviews a fundamental identity for estimation error, which has led to the construction of data defect correlation (Meng, 2018). Section 3 then discusses how this construct suggests the unified strategy. Section 4 demonstrates the strategy respectively for the qp and ξp settings in Wu (2022). Section 5 then applies the strategy to the two settings simultaneously to reveal an immediate insight into the celebrated double robustness, as reviewed in Wu (2022). Inspired by the same construct, Section 6 explores *counterbalancing sampling* as an alternative strategy to weighting. Section 7 concludes with a general call to treat probability sampling theory as an aspiration instead of the centerpiece of survey and sampling research.

1.2 A trio of probability constructs

The first of the three named constructs below, design probability, is self-explanatory. It is at the heart of sampling theory and reified by practical implementation, however imperfect the implementation might be. The distinction between the next two, divine probability and device probability, may be more nuanced especially at practical levels. But their conceptual differences are no less important than distinguishing between an estimand and an estimator. Fittingly, the data recording or inclusion indicator, a key quantity in modeling non-probability samples, provides a concrete illustration of all three probabilistic constructs; see the leading paragraph of Section 4.

Design Probability. A paramount concept and tool for statistics – and for general science – is randomized replications (Craiu, Gong and Meng, 2022). By designing and executing a probabilistic mechanism to generate randomized replications, we create probabilistic data that can be used directly for making verifiable inferential statements. Besides probabilistic sampling in surveys, randomization in clinical trials, bootstraps for assessing variability, permutation tests for hypothesis testing, and Monte Carlo simulations for computing are all examples of statistical methods that are built on design probability. Non-probability samples, by definition, do not come with such design probability, at least not an identified one. Hence, the phrase non-probability samples should be understood as a short hand for “samples without an identified design probability construct”.

It is worth to remind ourselves, however, that there is a potential for design probabilities to come back in a substantial way especially for large non-probability data sets, such as administrative data, due to the adoption of differential privacy (Dwork, 2008), for example by US Census Bureau (see the editorial by Gong, Groshen and Vadhan, 2022, and the special issue in *Harvard Data Science Review* it introduces). Differential privacy methods inject well-designed random noise into data for the purpose of protecting data privacy while not unduly sacrificing data utility. Like the design probability used for probabilistic sampling, the fact that the noise-injecting mechanism is designed by the data curator, and is made publicly known, renders the transparency that is critical for valid statistical inference by the data user (Gong, 2022). The area of how to properly analyze non-probability data with differential privacy protection is wide open. Even more so is the fascinating area of how to take into account the existing defects in non-probability data when designing probabilistic protection mechanisms for data privacy to avoid adding unnecessary noise. Readers who are interested in forming a big picture of the statistical issues involved in data privacy should consult the excellent overview article by Slavkovic and Seeman (2022) on the general area of “statistical data privacy”.

Divine Probability. In the absence of design probability for randomization-based inference, in order to conduct a (conventional) statistical inference, we typically conceptualize that the data at hand is a realization of a generative probabilistic mechanism given by nature or God. (I learned about the term “God’s model” during my PhD training, which I took as an expression for faith or something beyond human control, rather than reflecting one’s religious belief. The phrase “divine” is adopted here with a similar connotation.) We do so regardless of whether we believe or not that the world is intrinsically deterministic or stochastic (e.g., see David Peat, 2002; Li and Meng, 2021). We need to assume this divine probability primarily because of the restrictive nature of the probabilistic framework to which we are so accustomed. For example, in order to invoke the assumption of missing at random, we need to conjure a probabilistic mechanism under which the concept “missing at random” (Rubin, 1976) can be formalized. As Elliott and Valliant (2017) emphasized, the quasi-randomization approach, which corresponds to the qp framework of Wu (2022), “assumes that the nonprobability sample actually does have a probability sampling mechanism, albeit one with probabilities that have to be estimated under identifying assumptions". That is, we replace the design probability by a divine probability that we have faith for its existence, which then typically is treated as the "truth" or at least as an estimand.

Conceptually, therefore, we need to recognize that the assumption of any particular kind of divine probability is not innocent, as otherwise we will not need to rely on our faith to proceed. Nor is it always necessary. Any finite population provides a natural histogram for any quantifiable attributes or a contingency table for any categorizable attributes of its constituents, and hence it induces a divine probability without referencing any kind of randomness, conceptualized or realized, *if our inferential target is the finite population itself* (not a super-population that generates it, for example). The empirical likelihood approach takes advantage of this natural probability framework, which also turns out to be fundamental for quantifying data quality via data defect correlation (see Meng, 2018). The same emphasis was made by Zhang (2019), whose unified criterion was based on the same identity for building data defect correlation; see Section 2 below.

Device Probability. By far, most probabilities used in statistical modeling are devices for expressing our belief, prior knowledge, assumptions, idealizations, compromises, or even desperation (e.g., imposing a prior distribution to ensure identifiability since nothing else works). Whereas modeling reality has always been a key emphasis in the statistical literature, we inevitably must make a variety of simplifications, approximations, and some times deliberate distortions in order to deal with practical constraints (e.g., the use of variational inference for computational efficiency; see Blei, Kucukelbir and McAuliffe (2017)). Consequently, many of these device probabilities do not come with a requirement of being realizable, or even coherent mathematically (e.g., the employment of incompatible conditional probability distributions for multiple chain imputation; see Van Buuren and Oudshoorn (1999)). Nor are they easy or even possible to be validated, as Zhang (2019) investigated and argued in the context of non-probability sampling, especially with the superpopulation modeling approach, which corresponds to the ξp framework of Wu (2022). Nevertheless, device probabilities are the workhorse for statistical inferences. Both quasi-randomization approach and super-population modeling rely on such device probabilities to operate, as shown in Wu (2022) and further discussed in Sections 4-5 below. The lack of design probability can only encourage more device probabilities to make headway. To paraphrase Box's famous quote "all models are wrong, but some are useful", all device probabilities are problematic, but some are problem-solving.

1.3 Let's reduce "Garbage in, package out"

In a nutshell, probabilistic constructs are more needed for non-probability samples than probability ones precisely because of the deprivation of the design probability. Therefore, dealing with non-probability samples is not a new challenge for statisticians. If anything is new, it is the availability of massive amounts of large and non-probabilistic data sets, such as administrative data and social media data, and the accelerated need to combine multiple sources of data, most of which inherently are non-probabilistic because they are not collected for statistical inference purposes (e.g., Lohr and Rao, 2006; Meng, 2014; Buelens, Burger and van den Brakel, 2018; Beaumont and Rao, 2021). Contrary to common belief, the large sizes of "big data" can make our inference much worse, because of the "big data paradox" (Meng, 2018; Msaouel, 2022) when we fail to take into account the data quality in assessing the errors and uncertainties in our analyses; see Section 6.1.

It is therefore becoming more pressing than ever to greatly increase the general awareness of, and literacy about, the critical importance of data quality, and how we can use statistical methods and theories to help to reduce the data defect. The central concern here goes beyond the common warning about "garbage in, garbage out" – if something is recognized as garbage, it would likely be treated as such (likely, but not always, because as Andrew Gelman reminded me that "many researchers have a strong belief in *procedure* rather than *measurement*, and for these people the most important thing is to follow the rules, not to look at where their data came from"). The goal is to prevent "garbage in, package out" (Meng, 2021), where low quality data are auto-processed by generic procedures to create a cosmetically attractive "AI" package and sold to uninformed consumers or worse, to those who seek "data evidence" to mislead or disinform. Properly handling non-probability samples obviously does not resolve all the data quality issues, but it goes a very long way in addressing an increasingly common and detrimental problem of lack of data quality control in data science.

I therefore thank Professor Changbao Wu for a well timed and designed in-depth tour of "the must-sees" of the large sausage-making factory for processing non-probability samples. It adds considerably more detailed and nuanced exhibitions to the general tour by Elliott and Valliant (2017), which includes excellent illustrations on many forms and shapes of non-probability samples as well as their harms. It also showcases theoretical and methodological milestones for us to better appreciate the millstones displayed in the intellectual tour by Zhang (2019), which challenges statisticians and data scientists in general to understand better the quality, or rather the lack thereof, of the products we produce and promote. Together, this trio of overview articles form an informative tour for anyone who wants to join the force to address the ever-increasing challenges of non-probability data. Perhaps the best tour sequence starts with Elliott and Valliant (2017) to form a general picture, with Wu (2022)'s as the main exhibition of methodologies, and ends with Zhang (2019) to generate deep reflections on some specific challenges. For additional common methods for dealing with non-probability samples, such as multilevel modeling and poststratification, readers are referred to Gelman (2007), Wang, Rothschild, Goel and Gelman (2015) and Liu, Gelman and Chen (2021).

As a researcher and educator, I have been beating similar drums but often frustrated by the lack of time or energy to engage deeply. I am therefore particularly grateful to Editor Jean-François Beaumont for inviting me to help to ensure Professor Wu's messages are loud and clear: data cannot be processed as if they were representative unless the observed data are genuinely probability samples (which is extremely rare); many remedies have been proposed and tried, but many more need to be developed and evaluated. Among them, the concept of data defect correlation is a promising general metric to be explored and expanded, as demonstrated below.

2. A finite-population deterministic identity for actual error

To demonstrate the fruitfulness of the finite-population framework, consider the estimation of the population mean, denoted by $\bar{G}$, of $\{G_i = G(X_i): i \in \mathcal{N}\}$, where $\mathcal{N} = \{1, \ldots, N\}$ indexes a finite population, and the X_i's are data collected on individual i. For each i, let $R_i = 1$ if G_i (or rather X_i) is recorded in our sample, and $R_i = 0$ otherwise; hence the sample size is $n_R = \sum_{i=1}^N R_i$. We stress that this is an all-encompassing indicator, which can (and should) be decomposed into $R_i = r_i^{(1)}, \ldots, r_i^{(J)}$, when the data collection consists of J stages (e.g., $r_i^{(1)}$ indicates whether or not the i^{th} individual is sampled, and $r_i^{(2)}$ whether the individual responded or not once sampled).

Let $\{W_i, i \in S\}$ be a set of weights to be determined, where the index set $S = \{i : R_i = 1\}$, such that $\sum_{i \in S} W_i > 0$. Let $\bar{G}_W$ be the weighted sample average, expressible in three ways:

$$\bar{G}_W = \frac{\sum_{i \in S} W_i G_i}{\sum_{i \in S} W_i} = \frac{\sum_{i=1}^N R_i W_i G_i}{\sum_{i=1}^N R_i W_i} = \frac{\mathrm{E}_I(\tilde{R}_I G_I)}{\mathrm{E}_I(\tilde{R}_I)}, \tag{2.1}$$

where $\tilde{R}_I = R_I W_I$, and E_I is taken with respect to the uniform distribution over the index set $\mathcal{N}$. The first expression in (2.1) simply defines a weighted sample average. With the help of R_i, the second expression turns the sample averages into finite-population averages. This trivial re-expression is fundamental because it explicates the role of R_i in influencing the behavior of $\bar{G}_W$ as an estimator of $\bar{G}$. The third expression reveals a divine probability through I, the finite-population index (FPI) variable, by utilizing the fact that averaging is the same as taking expectation over a uniformly distributed random index I. All finite-population moments then can be expressed via E_I.

In particular, we can express the actual error of $\bar{G}_W$ via the following identity, where the first expression can be traced back to Hartley and Ross (1954), who used it to express biases in ratio estimators. The second expression was given in Meng (2018) with a slightly different (but equivalent) expression:

$$\bar{G}_W - \bar{G} = \frac{\mathrm{Cov}_I(\tilde{R}_I, G_I)}{\mathrm{E}_I[\tilde{R}_I]} = \rho_{\tilde{R},G} \times \sqrt{\frac{N - n_W}{n_W}} \times \sigma_G. \tag{2.2}$$

Here $\rho_{\tilde{R},G} = \mathrm{Corr}_I(\tilde{R}_I, G_I)$ is the *finite-population correlation* between $\tilde{R}_I$ and G_I, σ_G^2 is the finite-population variance of G_I, and n_W is the effective sample size due to using weights (Kish, 1965)

$$n_W = \frac{n_R}{1 + \mathrm{CV}_W^2}, \tag{2.3}$$

with CV_W being the coefficient of variation (i.e., standard deviation/mean) of $\{W_i, i \in S\}$.

The expression (2.2) is an algebraic identity because it holds for any instances of $\{(G_i, R_i W_i), i \in \mathcal{N}\}$. Hence no model assumptions are imposed, not even the assumption that R (or any quantity) is random, echoing the comment by Mary Thompson, as quoted in Wu (2022), that "the sample inclusion indicator R is a random variable is itself an assumption". The only requirement is that the recorded G_i is unchanged from the G_i's in the target population. (But note this requirement has two components: (1) there is no over-coverage, that is, everyone in the sample belongs to the target population, e.g., no non-eligible voters are surveyed when the target population is eligible voters, and (2) there is no measurement error; extensions to the cases with measurement errors are available, but not pursued in this article.) When we use equal weights, the three factors on the right-hand side of (2.2) reflect respectively (from left to right) data defect, data sparsity, and problem difficulty, as detailed in Meng (2018) and further illustrated in Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) in the context of COVID-19 vaccination surveys.

In particular, when all weights are equal, $\rho_{\tilde{R},G}$ is termed as *data defect correlation* (*ddc*) in Meng (2018) because it measures the lack of representativeness of the sample via capturing the dependence of inclusion/recording indicator on the attributes – the higher the dependence, the more biased the sample average becomes for estimating population averages. With the basic strategies of probabilistic sampling or inverse probability weighting, *ddc* will be zero on average because $\mathrm{E}(W_i R_i) = 1,$ and it is of $O_p(N^{-1/2})$ order because it is essentially an average of N independent terms (Meng, 2018). Our general goal here therefore is to bring down *ddc* to $O_p(N^{-1/2})$ for non-probability samples, which we shall refer to as "miniaturizing *ddc*" because $N^{-1/2}$ is typically a minuscule number in practice.

When we use weights, the first term $\rho_{\tilde{R},G}$ captures the data defect that still exists after the weighting adjustment, since no weights are perfect in practice. Identity (2.2) shows the impact of the weights on both data quality and data quantity. The impact on the *nominal* effective sample size n_W is never positive because $n_W \leq n_R$ as seen in (2.3). Incidentally, the exactness of (2.3) reveals that this well-known expression is in fact not an approximation (which is often attributed to Kish (1965)), but an exact formula for the reduction of the sample size due to weighting *if the weighting had no impact on ddc*. However, weighting can have a major positive impact on reducing the overall error by judiciously choosing weights to significantly decrease *ddc*, though apparently at the price of $n_W < n_R$. Of course, this is exactly the aim of the quasi-randomization framework, as discussed below. Most importantly, however, (2.2) leads to a unified insight about the variety of methods reviewed in Wu (2022), including an intuitive explanation of the doubly robust property, which has been receiving increased attention for integrating data sources including both probability and non-probability samples (e.g., Yang, Kim and Song, 2020).

Indeed, Zhang (2019, Section 3.1) used the first expression in (2.2) to define a unified non-parametric asymptotic (NPA) non-informativeness assumption, which requires that the numerator $\mathrm{Cov}_I(\tilde{R}_I, G_I)$ goes to zero, while keeping the denominator $\mathrm{E}_I[\tilde{R}_I]$ positive, as $N \to \infty.$ This unification permits Zhang (2019) to evaluate the quasi-randomization approach and regression modeling via a common criterion. The *ddc* framework echoes this unification, as discussed in Section 3 below, with Section 4 stressing the same broad message as emphasized by Zhang (2019). Section 5 harvests another low-hanging fruit of the *ddc* formulation, since it provides an immediate explanation of the celebrated double robustness. Section 6 then ventures into a much harder area of engineering a more representative sub-sample out of a large non-representative sample, a worthwhile trade-off because data quality is far more important than data quantity (Meng, 2018), as briefly reviewed below.

3. A unifying strategy based on data defect correlation

In the setup of Wu (2022), for each individual $i,$ we have a set of attributes $A_i = \{y_i, \mathbf{x}_i\},$ where y is the attribute of interest, and $\mathbf{x}$ is auxiliary, which is useful in two ways. First, reducing the sampling bias due to non-probability sampling becomes possible when the non-probability mechanism can be (fully) explained by $\mathbf{x}$. Second, by taking advantage of the relationships between y_i and $\mathbf{x}_i,$ we can improve the efficiency of our estimation. As a starting point, Wu (2022) assumes that we have two data sources available, which we denote via two recording indicators, R and $R^*.$ The main source of the data is a non-probability sample, where we observe both y_i and $\mathbf{x}_i$ for $i \in S \equiv \{i : R_i = 1\},$ but the recording indicator R_i is determined by a mechanism uncontrolled by any (known) design probability. A second source is (assumed to be) a probability sample, where we observe $\mathbf{x}_i$ only, for $i \in S^* \equiv \{i : R_i^* = 1\}.$ This second sample provides information to estimate population auxiliary information that is useful for estimating population quantities about $y,$ such as its mean. Hence this setup is closely related to the setup where $S \cup S^* = \mathcal{N};$ see Tan (2013).

Now for any function $m(\mathbf{x}),$ let $z_i = y - m(\mathbf{x}_i), i \in \mathcal{N}.$ Clearly we can estimate the population mean $\bar{y}_N = \mathrm{E}_I(y_I)$ via estimating $\bar{z} = \mathrm{E}_I(z_I)$ and $\bar{m} = \mathrm{E}_I\left[m(\mathbf{x}_I)\right].$ From the second sample, $\bar{m}$ can be estimated unbiasedly since it involves $\mathbf{x}$ only. We therefore can focus on estimating $\bar{z},$ while recognizing that a more principled approach is to set up a likelihood or Bayesian model to estimate all unknown quantities jointly (Pfeffermann, 2017). Applying identity (2.2) with $G = z$ then tells us that our central task is to choose the weight $\{W_i, i \in S\}$ and/or the m function to miniaturize the *ddc* $\rho_{\tilde{R},z}.$ For our current discussion, it is easier to explain everything via the covariance

$$c_{\tilde{R},z} \equiv \mathrm{Cov}_I(\tilde{R}_I, z_I) = \mathrm{Cov}_I(W_I R_I, y_I - m(\mathbf{x}_I)) = \frac{1}{N}\sum_{i=1}^{N} W_i R_i (z_i - \bar{z}) \tag{3.1}$$

instead of the correlation $\rho_{\tilde{R},z}$ because $\mathrm{Cov}_I(\tilde{R}_I, z_I)$ is a bi-linear function in R_I and $z_I.$ However, $\rho_{\tilde{R},z},$ being standardized, is more appealing theoretically and for modelling purposes; see Sections 6 and 7.

The expression in (3.1) tells us immediately how to make it zero in expectations operationally, and in what sense conceptually. For whatever probability we impose on R_i (to be specified in late sections), let $\pi_i = \Pr(R_i = 1 \mid \mathbf{A}),$ which we assume will depend on A_i only. Then the linearity of the covariance operator implies that the average covariance with respect to the randomness in R_i is given by

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{A}] = \mathrm{Cov}_I\left(W_I \pi_I, y_I - m(\mathbf{x}_I)\right), \tag{3.2}$$

where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Similarly, if one is willing to posit a joint model for $\{(R_i, y_i), i \in \mathcal{N}\}$ conditioning on $\mathbf{X}$ in the independence form $\Pi_{i=1}^{N} P\left(R_i, y_i \,|\, \mathbf{x}_i\right)$, then

$$\mathrm{E}[c_{\tilde{R},z} \,|\, \mathbf{X}] = \mathrm{Cov}_I\left(W_I \pi_I, \mathrm{E}(y_I \,|\, \mathbf{x}_I) - m(\mathbf{x}_I)\right). \tag{3.3}$$

Very intuitively, one can ensure a zero covariance or correlation between two variables by making either of them a constant. The two choices then would lead to respectively the quasi-randomization approach by making $W_I \pi_I \propto 1$ and the super-population approach by making $\mathrm{E}[y_I \,|\, \mathbf{x}_I] - m(\mathbf{x}_I)$ a constant (e.g., zero). The fact that either one is sufficient to render zero covariance (under the joint model) yields the double robustness, because it does not matter which one. But clearly these are not the only methods to achieve a zero correlation/covariance or double robustness, an emphasis of Kang and Schafer (2007) in their attempt to demystify the doubly robust approach (Robins, Rotnitzky and Zhao, 1994; Robins, 2000; Scharfstein, Rotnitzky and Robins, 1999). See also Tan (2007, 2010) for discussions and comparisons of an array of estimators, including those corresponding to only the quasi-randomization approach or only the super-population approach, some of them are doubly robust.

Indeed, because formula (2.2) is an identity for the actual error, any asymptotically unbiased (linear) estimators of the population mean must imply its corresponding *ddc* is asymptotically unbiased for zero, and vice versa, with respect to the randomness in R or in $\{R, y\}$. However, it is possible for *ddc* to be asymptotically unbiased for zero, without assuming any model is correctly specified – see Section 5 for an example. (This “double-plus robustness” is different from the “multiple robustness” of Han and Wang (2013), which still needs to assume the validity of at least one of the posited multiple models.) These two observations suggest that any general sufficient and necessary strategy for ensuring asymptotically consistent/unbiased (linear) estimators for the population mean would be equivalent to miniaturizing *ddc*.

As an example of a unified insight that otherwise might not be as intuitive, expression (3.2) suggests that we should include our estimate of π_I as a part of the predictor in the regression model $m(\mathbf{x}_I)$, since that can help to reduce the correlation between $W_I \pi_I$ and $z_I = y_I - m(\mathbf{x}_I)$, especially when we use constant weights W_I. Using $\hat{\pi}_I$ as a predictor for y is generally hard to motivate purely from the regression perspective, especially when we assume y and R are independent given $\mathbf{x}$ (typically a necessary condition to proceed, as discussed in the next section). However, expression (3.2) tells us that for the purpose of estimating the mean of y, it is not absolutely necessary to fit the correct regression model $m(\mathbf{x})$. Rather, it is sufficient to ensure the “residual” z_I is as uncorrelated with $W_I \pi_I$ as I varies. However, it is critically important to recognize that it is not sufficient to ensure zero or small correlation only among the observed data, because $\mathrm{Cov}_I(W_I \pi_I, z_I \,|\, R_I = 1)$ tells us little about $\mathrm{Cov}_I(W_I \pi_I, z_I \,|\, R_I = 0)$. In the setting of Wu (2022), our ability to extrapolate from $R_I = 1$ to $R_I = 0$ depends on the availability of the (independent) auxiliary data indexed by $R_I^* = 1$, which allow us to observe some x_I’s for which $R_I = 0$.

The strategy of including propensity estimates as a predictor has been found beneficial in related literature. For example, Little and An (2004) included the logit of $\hat{\pi}$ in their imputation model, and reported the inclusion enhanced the robustness of the imputed mean to the misspecification of the imputation model. The method was further developed and enhanced by Zhang and Little (2009) and by Tan, Flannagan and Elliott (2019), who used the term "Robust-squared" to emphasize the enhanced robustness. In a more recent article on such a strategy for non-probability samples, Liu et al. (2021) emphasized the importance of including the estimated propensity $\hat{\pi}_i$ "as a predictor" in $m(x, \hat{\pi})$ (using notation in this article). Furthermore, in the literature of targeted maximum likelihood estimation (TMLE) for semi-parametric models for dealing with non-probability data (van der Laan and Rubin, 2006; Luque-Fernandez, Schomaker, Rachet and Schnitzer, 2018) (also see Scharfstein et al. (1999); Tan (2010)), the variables $R_I / \hat{\pi}_I$ and $(1 - R_I) / (1 - \hat{\pi}_I)$ are called *clever covariates* and are used in the regression models for y_I. The implementations and theories of TMLE, and the related Collaborative TMLE (van der Laan and Gruber, 2009, 2010), are mathematically more involved than those under finite-population settings as discussed below, but the insights gained from (3.2)-(3.3) can provide us with helpful intuitions on understanding the essence of such methods.

4. Quasi-randomization *or* super-population implementations

In a nutshell, the quasi-randomization approach focuses on making $W_I \pi_I$ a constant variable (induced by FPI I). When our sample is genuinely selected by a probabilistic scheme by design, then $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i),$ for $i \in \mathcal{N},$ is a design probability, free of $y_i,$ but it can depend on $\mathbf{x}_i$ for example when $\mathbf{x}_i$ includes a stratifying variable. When the design probability is unavailable, we first need to invoke a divine probability. This could be a natural one given by the finite population, such as the propensity $\pi_i = \Pr_I(R_I = 1 \mid A_I = A_i)$ induced by FPI, where $A_i = \{y_i, \mathbf{x}_i\},$ or an imagined super-population one such as the R_i's being generated independently from $\mathrm{Ber}(\pi_i),$ where $\pi_i = \Pr(R_i = 1 \mid A_i) > 0.$ This positivity assumption is necessary if the finite population is pre-specified, or its imposition defines the finite population that can be studied. (This is a practically rather relevant consideration, such as in election polling, where the finite population may not be always pre-specified even theoretically.) Since these divine probabilities are unknown and serve as our estimand, we need to assume some device probabilities, such as via a generalized linear model $\pi_i = g(y_i, \mathbf{x}_i)$ to proceed, even though we don't really believe in any particular choice of $g.$

For our current discussion, suppose our divine probability is given by the super-population Bernoulli model. Let $n_R = \sum_{i=1}^{N} R_i,$ and $\tilde{p}(\mathbf{A}) = \Pr(n_R > 0 \mid \mathbf{A}) = 1 - \Pi_{i \in N}(1 - \pi_i),$ where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}.$ Because the R_i here is controlled by a divine probability, the sample size n_R is no longer a design variable to be conditioned upon in our replication scheme; it is generally no longer an ancillary statistic. Nevertheless, we should condition on $n_R > 0,$ a universal requirement for constructing data-driven estimates for $\bar{G}.$ Fortunately this conditioning does not create mathematical complications to the simplicity granted by the independence among $\pi_i, i \in \mathcal{N}$ as functions of $A_i.$ This is because $\tilde{\pi}_i(\mathbf{A}) \equiv \Pr(R_i = 1 \mid \mathbf{A}, n_R > 0) = \pi_i / \tilde{p}(\mathbf{A}),$ but the normalizing constant $\tilde{p}(\mathbf{A})$ – which depends on the entire $\mathbf{A}$ – is not relevant for the developments in this article, such as assigning weights that are proportional to $\tilde{\pi}_i^{-1}(\mathbf{A})$.

Consequently, under this divine probability, which corresponds to (the true model for) the q-model setting in Wu (2022), we have for any chosen W_I, by (3.1)

$$\begin{aligned} \mathrm{E}(c_{\tilde{R},z} \mid \mathbf{A}, n_R > 0) &= \mathrm{Cov}_I(W_I \mathrm{E}[R_I \mid \mathbf{A}, n_R > 0], y_I - m(\mathbf{x}_I)) \\ &= \tilde{p}^{-1}(\mathbf{A}) \, \mathrm{Cov}_I(W_I \pi_I, y_I - m(\mathbf{x}_I)), \end{aligned} \tag{4.1}$$

where E is with respect to the (unknown) divine probability over R_I (for fixed I). It follows then that, regardless of whether we want to ensure zero expectation in (3.2) or in (4.1), we will impose $W_I \pi_I \propto 1$, that is, $W_I \propto \pi_I^{-1}$, the well-known inverse probability weighting. Therefore, if our postulated model q permits us to reliably capture π_i in reality, then $c_{\tilde{R},z} = O_p(N^{-1/2})$ because it has mean zero (with respect to the divine probability), and it is a weighted average of N essentially independent Bernoulli variables, as seen in (3.1).

This is a randomization oriented approach because it treats the entire finite population attribute values $\mathbf{A}$ as fixed, and the hypothetical replications are generated only by repeated realizations of the recording indicator R_I. Of course, in general, the values of $\{\pi_i, i \in \mathcal{N}\}$ are unknown, and worse they are inestimable from a non-probability sample without further assumptions. To proceed, we pose assumptions such as missing at random, i.e., $\Pr(R_i = 1 \mid A_i) = \Pr(R_i = 1 \mid \mathbf{x}_i)$, and the requirement of an auxiliary sample so that we have some values of $\mathbf{x}_i$ with $R_i = 0$. We also have choices on how to estimate the inclusion propensity $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$, parametrically or non-parametrically. These assumptions, requirements, and estimation methods are all essential for practical implementation, as carefully reviewed and discussed by Wu (2022); also see Tan (2010) for a detailed comparison of various estimation strategies. Nevertheless, the overarching idea of quasi-randomization methods is to choose W_I to free $\tilde{R}_I = W_I R_I$ from I in expectation over the posited hypothetical replications, to regain the freedom guaranteed by probability sampling.

Complementarily, the super-population approaches aim to miniaturize $c_{\tilde{R},z}$ via making the other variable in $c_{\tilde{R},z}$, that is, z_I free of I in expectation, but over a different hypothetical replication scheme. Here the idea is to choose an $m(\mathbf{x}_i)$ that is a good approximation to y_i such that the residual $z_i = y_i - m(\mathbf{x}_i)$ will be zero in expectation conditioning on $\mathbf{x}$. Typically, this is done by considering a joint model for $\{R_i, y_i\}$ given $\mathbf{x}_i$, and with a specific regression model $\xi(y \mid \mathbf{x})$, using the notation in Wu (2022). It is important to recognize that, although we only specify the regression model y_i given $\mathbf{x}_i$, we must include R_i in the replications in order to capture the possible dependence of R_i on the entire $A_i = \{y_i, \mathbf{x}_i\}$, which is the key concern for non-probability samples. Indeed, it is this joint specification that permits the adoption of the missing at random assumption to reduce $P(y_i \mid \mathbf{x}_i, R_i) = P(y_i \mid \mathbf{x}_i)$, which in turn permits us to focus on specifying a single regression model $\xi(y_i \mid \mathbf{x}_i)$ for both observed and unobserved individuals. Therefore, when we write E_ξ, we mean the expectation with respect to

$$P(R_i, y_i \mid \mathbf{x}_i) = P(R_i \mid \mathbf{x}_i)\, P(y_i \mid R_i, \mathbf{x}_i) = \pi_i^{R_i} (1-\pi_i)^{1-R_i} \xi(y_i \mid \mathbf{x}_i), \tag{4.2}$$

where $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ is left unspecified, unlike with the quasi-randomization approach.

It follows then that, conditioning on $\mathbf{X} = \{\mathbf{x}_i, i \in \mathcal{N}\}$ and $n_R > 0$, which does not alter $P(y \mid \mathbf{X})$ because y and R are independent given $\mathbf{X}$, we have

$$\mathrm{E}(c_{\tilde{R},z} \mid \mathbf{X}, n_R > 0) = [\tilde{p}(\mathbf{X})]^{-1} \mathrm{Cov}_I (W_I \pi_I, \mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)). \tag{4.3}$$

Clearly, (4.3) becomes zero when we choose $m(\mathbf{x}_I) = \mathrm{E}_\xi [y_I \mid \mathbf{x}_I]$ and that the ξ model is (first-order) correctly specified, that is, $\mathrm{E}_\xi [y_I \mid \mathbf{x}_I] = \mathrm{E}[y_I \mid \mathbf{x}_I]$. This summarizes the super-population approach, and it renders $c_{\tilde{R},z} = O_p(N^{-1/2})$ for similar reasons as given for the quasi-randomization framework.

5. Quasi-randomization *and* super-population implementations

Once a joint model for $\{R_i, y_i\}$ is set up, of course we can use it for estimating both π_i and the regression function $m(\mathbf{x})$, each of which is made possible by the availability of the auxiliary probability sample, and the assumption of missing at random. But as shown before, correctly specifying and estimating one of them is sufficient for miniaturizing $c_{\tilde{R},z}$. However, from (4.3), in order for the covariance/correlation to be zero, neither multiplicative correction to π_I via W_I nor the additive adjustment for $\mathrm{E}(y_I \mid \mathbf{x}_I)$ via $m(\mathbf{x}_I)$ need to be correct. All we need is that, after the correction or adjustment, what is left would be uncorrelated with each other. The aforementioned framework of Collaborative TMLE was built essentially on this insight (e.g., see Section 3.1 of van der Laan and Gruber, 2009), though the heavy mathematical treatments in its literature might have discouraged readers to seek such intuitive understanding.

To provide a simple illustration, consider a finite population that is an i.i.d. sample from a super-population model:

$$\mathrm{E}[y \mid x] = \sum_{k=0}^{3} \beta_k x^k, \quad x \sim N(0,1). \tag{5.1}$$

The non-probability sample is generated by a mechanism R such that $\Pr\big(R = 1 \mid y, x\big) = \pi\big(\lvert x \rvert\big)$, that is, it is determined by the magnitude of x only. Suppose we mis-specify the function form for π (e.g., the divine model may not be monotone in $\lvert x \rvert$, but the device model such as the conventional logistic link is), as well the regression model by choosing $m(x) = b_0 + b_1 x + b_2 x^2$. Since x^2 is uncorrelated with x or x^3 under $x \sim N(0,1)$, we know that our least-square estimator for b_2 would still be valid for β_2 even under the mis-specified regression model. This turns out to be sufficient to ensure the asymptotic unbiasedness (as $N \to \infty$) of the following "doubly robust" estimator for $\mu = \bar{y}_N$, the finite-population mean,

$$\hat{\mu}_+ = \frac{\sum_{i=1}^N R_i w(|x_i|)(y_i - \hat{m}(x_i))}{\sum_{i=1}^N R_i w(|x_i|)} + \frac{\sum_{i=1}^N R_i^* \hat{m}(x_i)}{\sum_{i=1}^N R_i^*}, \tag{5.2}$$

where R^* indicates the auxiliary sample (of $\mathbf{x}$ only). Or equivalently,

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I\left(R_I w(|x_I|), y_I - \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I w(|x_I|)\right)} + \frac{\mathrm{Cov}_I\left(R_I^*, \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I^*\right)}, \tag{5.3}$$

which makes it clearer that any bias in $\hat{\mu}_+$ is controlled by the covariance (or correlation) involving R, since the covariance involving R^* is already miniaturized by the assumption that the auxiliary sample is probabilistic (which, for simplicity, is assumed to be a simple random sample).

Here $w(x)$ is any weight function such that $\mathrm{E}_\phi\left[|x|^3 w(|x|)\right] < \infty$, where the expectation is with respect to $x \sim N(0,1)$, and $\hat{m}(x) = b_0 + b_1 x + \hat{\beta}_2 x^2$, with $\hat{\beta}_2$ being the least-square estimator for β_2 from the biased sample, and b_0 and b_1 can be chosen arbitrarily. Because the finite-population covariance/correlation between $\pi(|x_I|) w(|x_I|)$ and x_I^k is $O_p(N^{-1/2})$, for $k = 1$ and $k = 3$, the misfitted parts for π or m do not contribute to the *ddc* (asymptotically) since they are uncorrelated with each other under the super-population model, leading to further robustness going beyond "double robustness". This of course does not mean that we can misfit a model arbitrarily and still obtain valid estimators, but it does imply that having at least one model being correct is a sufficient, but not necessary, condition for the validity of the doubly robust estimators.

It is also worth stressing that, in formatting the regression model, we do not necessarily need to invoke a device probability, e.g., a super-population regression model, because the FPI variable provides a finite-population regression via applying the least-squares method to regress y_i on $\mathbf{x}_i, i \in \mathcal{N}$. This regression fitting itself says little about whether the resulting regression line $y = \hat{m}(\mathbf{x})$ is a good fit to $(y_i, \mathbf{x}_i)$ or not. However, the example above indicates that, for the purpose of estimating the population average of y, the lack of fit may not matter that much, as long as the "residual" $z_I = y_I - \hat{m}(\mathbf{x}_I)$ has little correlation with $W_I \pi_I$, as two functions of the FPI variable I. Indeed, as discussed in Section 3, we can consider including $\hat{\pi}_I$ in the regression model $\hat{m}(\mathbf{x}_I, \hat{\pi}_I)$. How effective this strategy is in general is a topic of further research.

6. Counterbalancing sub-sampling

6.1 The devastating impact of data defect on effective sample size

A key finding, which has surprised many, from studying the data quality issue is how small the size of our "big data" is when we take into account the data defect. To prove this mathematically, we can equate the mean-squared error (MSE) of $\bar{G}_W$ in (2.1), with the MSE of a simple random sampling estimator of size n_{eff}. This yields (see Meng (2018) for derivation):

$$n_{\mathrm{eff}} \approx \frac{f_W}{1-f_W}\frac{1}{\mathrm{E}[\rho^2_{\tilde{R},G}]} \approx \frac{f_W}{1-f_W}\frac{1}{\rho^2_{\tilde{R},G}}, \tag{6.1}$$

where $f_W = n_W / N$ and the expectation E is with respect to the conditional distribution of $\tilde{R}$ given n_W. It is worthwhile to note that this (conditional) distribution can involve all three types of probability discussed in Section 1.2 because the variations in $\tilde{R}$ can come from multiple sources. For example, in typical opinion surveys, there will be (1) design probability in the sampling indicator, (2) divine probability in formulating the non-response mechanism, and (3) device probability for estimating the mechanism and the weights.

Expression (6.1) is the weighted version/extension of the expression given in Meng (2018) with equal weights, which reveals the devastating impact of a seemingly tiny *ddc*. Suppose our sample is 1% of the population, and it suffers from a half-percent *ddc*. Applying (6.1) (with equal weights) with $f_W = 0.01$ and $\rho_{\tilde{R},G} = 0.005$ yields $n_{\mathrm{eff}} \approx 404$ *regardless of the sample size* n_R. In the case of the 2020 US presidential election, 1% of the voting population is about 1.55 million people, and hence the loss of sample size due to a half percent *ddc* is about 1 - (404 / 1,550,000) > 99.97%. Such seemingly impossible losses have been reported in both election studies (Meng, 2018) and COVID vaccination studies (Bradley et al., 2021). A most devastating consequence of such losses is the “big data paradox”: the larger the (apparent) data size, the surer we fool ourselves because our false confidence (in both technical and literal sense) goes up with the erroneous data size, while the actual coverage probability of the incorrectly constructed confidence intervals become vanishingly small (Meng, 2018; Msaouel, 2022).

A positive implication from this revelation, however, is that we can trade much data quantity for data quality, and still end up having statistically more accurate estimates. Of course, in order to reduce the bias, we will need some information about it. If we have reliable information on the value of *ddc*, we can directly adjust for the bias in estimating the population average corresponding to the *ddc*, for example by a Bayesian approach, similar to that taken by Isakov and Kuriwaki (2020) in their scenario analysis. Furthermore, if we have sufficient information to construct reliable weights, we can use the weights to adjust for selection biases as commonly done. Nevertheless, even in such cases, it may still be useful to create a representative miniature of the population out of a biased sample for general purposes, which for example can eliminate many practitioners’ anxiety and potential mistakes for not knowing how to properly use the weights. Indeed, few really know how to deal with weights, because “Survey weighting is a mess” (Gelman, 2007).

However, creating a representative miniature out of a biased sample in general is a challenging task, especially because *ddc* can (and will) vary with the variable of interest. Nevertheless, just as weighting is popular tool despite it being far from perfect, let us explore representative miniaturization and see how far we can push the idea. The following example therefore is purely for brainstorming purposes, by looking into a common but challenging scenario, where we have reasonable information or understanding on the direction of the bias, that is, the sign of the *ddc*, but rather vague information about its magnitude. A good example is non-representativeness of election polls because voters tend to not want to disclose their preferences when they plan to vote for a socially unpopular candidate; we therefore know the direction of the bias, but not much about its degree other than some rough guesses (e.g., a range of 10 percentage points).

6.2 Creating a less biased sub-sample

The basic idea is to use such partial information about the selection bias to design a *biased* sub-sampling scheme to *counterbalance* the bias in the original sample, such that the resulting sub-samples have a *high likelihood* to be less biased than the original sample from our target population. That is, we create a sub-sampling indicator $S_I,$ such that with high likelihood, the correlation between $S_I R_I$ and G_I is reduced, compared to the original $\rho_{R,G},$ to such a degree that it will compensate for the loss of sample size and hence reduce the MSE of our estimator (e.g., the sample average). We say with *high likelihood*, in its non-technical meaning, because without full information on the response/recording mechanism, we can never guarantee such a counterbalance sub-sampling (CBS) would always do better. However, with judicious execution, we can reduce the likelihood of making serious mistakes.

To illustrate, consider the case where y is binary. Let $\Delta = r_1 - r_0,$ where r_y is the propensity of responding/reporting for individuals whose responses will take value $y:$ $r_y = \Pr_I\,(R_I = 1 \mid y_I = y).$ If the sample is representative, then like $\rho_{R,G},$ Δ is miniaturized, meaning that it is on the order of $N^{-1/2}.$ This is most clearly seen via the easily verifiable identity (see (4.1) of Meng, 2018)

$$\Delta = \frac{\mathrm{Cov}_I\,(y_I, R_I)}{p(1-p)} = \rho_{R,y}\sqrt{\frac{f_R(1-f_R)}{p(1-p)}}, \tag{6.2}$$

where $p = \Pr_I\,(y_I = 1)$ and $f_R = \Pr_I\,(R_I = 1),$ which is the original sampling rate. A key ingredient of CBS is to determine $s_y = P_I\,(S_I = 1 \mid y_I = y, R_I = 1)$ for $y = 0, 1,$ that is, the sub-sampling probabilities of individuals who reported $y = 1$ and $y = 0,$ respectively.

To determine the beneficial choices, let $f_S = \Pr_I\,(S_I = 1 \mid R_I = 1)$ be the sub-sampling rate, and $\Delta_S = s_1 r_1 - s_0 r_0.$ Then by applying (2.2) (with equal weights) and (6.2) to both the sample average and the sub-sample average, we see that the sub-sample average has smaller (actual) error in magnitude if and only if

$$\left(\frac{\Delta_S}{f_S f_R}\right)^2 < \left(\frac{\Delta}{f_R}\right)^2 \Leftrightarrow f_S^2 > \left(\frac{\Delta_S}{\Delta}\right)^2. \tag{6.3}$$

Writing $r = r_1/r_0$ and $s = s_1/s_0,$ the right-hand side of (6.3) becomes

$$\left[sp^* + (1-p^*)\right]^2 > \left(\frac{rs-1}{r-1}\right)^2, \tag{6.4}$$

where $p^* = \Pr_I(y_I = 1 | R_I = 1)$ is observed in the original sample, which should remind us that p^* may be rather different from the p we seek, because of the biased R-mechanism.

An immediate choice to satisfy (6.4) is to set $s = r^{-1}$, which of course typically is unrealistic because if we know the value of r, then the problem would be a lot simpler. To explore how much leeway we have in deviating from this ideal choice, let $\delta = r - 1$, we can then show that (6.4) is equivalent to

$$(s-1)\left\{[1+(1+p^*)\,\delta]\,(s-1)+2\delta\right\} \;<\; 0. \tag{6.5}$$

This tells precisely the permissible choices of s without over-correcting (in the magnitude of the resulting bias):

(i) When $r > 1$, i.e., $\delta > 0$, we can take any s such that

$$\frac{[1-(1-p^*)\,\delta]_+}{1+(1+p^*)\,\delta} \;\le\; s \;<\; 1; \tag{6.6}$$

(ii) When $r < 1$, i.e., $\delta < 0$, we can take any s such that

$$1 \;<\; s \;\le\; \frac{1-(1-p^*)\,\delta}{[1+(1+p^*)\,\delta]_+}. \tag{6.7}$$

This pair of results confirms a number of our intuitions, but also offers some qualifications that are not so obvious. Since we sub-sample to compensate for the bias in the original sample, s and r must stay on the opposite side of 1, i.e., $(s-1)(r-1) = (s-1)\,\delta < 0$, as seen in (6.6)-(6.7). To prevent over corrections, some limits are needed, but it is also possible that the initial bias is so bad that no sub-sampling scheme can make things worse, which is reflected by the positivizing function $[x]_+$ in the two expressions above. However, the expressions for the limits as well as for the thresholds to activate the positivizing functions are not so obvious. Nor is it obvious that these expressions depend on the unknown p indirectly via the observed p^*, and hence only prior knowledge of r is required for implementing or assessing CBS.

This observation suggests that it is possible to implement a beneficial CBS when we can borrow information from other surveys (or studies) where the response/recording behaviors are of similar nature. For example, we may learn that a previous similar survey had $r =$ 1.5 (e.g., those with $y = 1$ had 6% of chance to be recorded, and those with $y = 0$ had only 4% chance). Taking into account the uncertainty in the similarity between the two surveys, we might feel comfortable to place (1.2, 1.8) as the plausible range for r in the current study. Suppose we observe $p^* =$ 0.6, this means that the maximum – over the range $r \in$ (1.2, 1.8) – of the lower bound on the permissible s as given in (6.6) is

$$\frac{[1-(1-0.6)\,(r-1)]_+}{1+1.6\,(r-1)} \;=\; \frac{[1.4-0.4r]_+}{1.6r-0.6} \;\le\; \frac{1.4-0.4\times 1.2}{1.6\times 1.2-0.6} \;=\; 0.7. \tag{6.8}$$

Therefore, as long as we choose $s \in [0.7, 1)$, we are unlikely to over-correct. The price we pay for this robustness is that the resulting sub-sample is not as good quality as it can be, for example, when the underlying r for the current study is indeed 1.5 (in expectation). Choosing any $s \in [0.7, 1)$ will not provide the full correction as provided by $s = 1/r = 0.67$, that is, the sub-sample average will still have a positive bias but with a smaller MSE compared to the original sample average. Of course both the feasibility and effectiveness of such CBS need to be carefully investigated before it can be recommended for general consumption, especially going beyond binary y. The literature on inverse sampling (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003) is of great relevance for such investigations, because it also aims to produce simple random samples via subsampling, albeit with a different motivation (to turn complex surveys into simple ones for ease of analysis).

7. Probability sampling as aspiration, not prescription

As it should be clear from the definition of *ddc*, it is not directly estimable from the biased sample alone. One therefore naturally would (and should) question how useful *ddc* is or could be. The answer turns out to be an increasingly long one thanks to *ddc* being model-free and hence a versatile data quality metric for both probability samples and non-probability samples. Its usefulness for generating theoretical insights is demonstrated by its role in quantifying the data quality-quantify trade-off via effective sample size as seen in (6.1), in understanding simulation errors in quasi-Monte Carlo as explored in Hickernell (2016), and in anticipating the "double-plus robustness" phenomenon as presented in Section 5. Its methodological usages are illustrated by the scenario analyses for the 2020 US Presidential election (Isakov and Kuriwaki, 2020) and for the COVID-19 vaccination assessments (Bradley et al., 2021). Its practical implications can be found in epidemiological studies (Dempsey, 2020), particle physics (Courtoy, Houston, Nadolsky, Xie, Yan and Yuan, 2022), and political polling (Bailey, 2023).

Not surprisingly, these practical applications found the notion of *ddc* and the underlying error decomposition (2.2) helpful because of the non-probability samples they need to deal with, either due to distortions to the probability samples such as by a biased non-response mechanism or due to selection biases in the first place such as selective COVID-19 testing. Professor Wu's overview, and the many references cited there and in this discussion, should make it clear that non-probability samples are *almost surely* everywhere. I am invoking this strong probabilistic phrase not merely for its humorous value. When we consider the unaccountably many possible values for the mean of *ddc*, the probability – however we construct it to capture the wild west of data collection processes out there – that it will land precisely on zero must be zero. This zero mean is a necessary condition for the sample to be a probability sample, because a probability sample implies that *ddc* must be of the order of $N^{-1/2}$ order (Meng, 2018), which is impossible when its mean is non-zero (asymptotically). This observation suggests that we should move away from our tradition of treating probability sampling as a centerpiece and then try to model the much larger world of non-probability samples as "deviations" from it. Instead, we should start with studying samples with general collection mechanisms using tools or concepts such as *ddc*, and then treat (design) probability samples as the very special, ideal case – always an aspiration, but never the only prescription for action.

Acknowledgements

I am grateful to Editor Jean-François Beaumont for inviting me to discuss Changbao Wu's timely and thought-provoking overview. I thank James Bailie, Radu Craiu, Adel Daoud, Andrew Gelman, Stas Kolenikov, Rod Little, Cory McCartan, Kelly McConville, James Robins, Zhiqiang Tan, and Li-Chun Zhang for moral endorsement and for constructive criticisms. I also thank NSF for partial financial support, and Steve Finch for careful proofreading.

References

Bailey, M.A. (2023). *Polling at a Crossroads – Rethinking Modern Survey Research*. Cambridge University Press.

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

Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, Z.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695-700.

Buelens, B., Burger, J. and van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *International Statistical Review*, 86(2), 322-343.

Courtoy, A., Houston, J., Nadolsky, P., Xie, K., Yan, M. and Yuan, C.-P. (2022). Parton distributions need representative sampling. *arXiv preprint arXiv:2205.10444.*

Craiu, R.V., Gong, R. and Meng, X.-L. (2022). Six statistical senses. *arXiv preprint arXiv:2204.05313.*

David Peat, F. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century*. Joseph Henry Press.

Dempsey, W. (2020). The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts. *arXiv preprint arXiv:2005.10425*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, Springer, 1-19.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Special Issue 2), June 24, 2022. https://hdsr.mitpress.mit.edu/pub/ld4smnnf.

Gong, R., Groshen, E.L. and Vadhan, S. (2022). Harnessing the known unknowns: Differential privacy and the 2020 Census. *Harvard Data Science Review*, (Special Issue 2), June 24 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

Han, P., and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100(2), 417-430.

Hartley, H.O., and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270-271.

Hickernell, F.J. (2016). The trio identity for Quasi-Monte Carlo error. In International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer, 3-27.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-21. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3101-eng.pdf.

Isakov, M., and Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective Lens from 2016. *Harvard Data Science Review*, 2(4), Nov. 3, 2020. https://hdsr.mitpress.mit.edu/pub/cnxbwum6.

Kang, J.D.Y., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

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

Li, X., and Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533), 353-367.

Little, R., and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949-968.

Liu, Y., Gelman, A. and Chen, Q. (2021). Inference from non-random samples using Bayesian machine learning. *arXiv preprint arXiv:2104.05192*.

Lo, A.W. (2017). Adaptive markets. In *Adaptive Markets*. Princeton University Press.

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

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Luque-Fernandez, M.A., Schomaker, M., Rachet, B. and Schnitzer, M.E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530-2546.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science*, (Eds., Lin et al.), CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4), 1161-1175.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, 1-27.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, 69(1), 35-63.

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

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, Indianapolis, IN, 1999, 6-10.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussions). *Journal of the American Statistical Association*, 94(448), 1096-1146.

Slavkovic, A., and Seeman, J. (2022). Statistical data privacy: A song of privacy and utility. *arXiv preprint arXiv:2205.03336*.

Tan, Y.V., Flannagan, C.A.C. and Elliott, M.R. (2019). “Robust-Squared” imputation models using Bart. *Journal of Survey Statistics and Methodology*, 7(4), 465-497.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560-568.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661-682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2), 399-415.

Van Buuren, S., and Oudshoorn, K. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO.

van der Laan, M.J., and Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 246.

van der Laan, M.J., and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M.J., and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wu, C. (2022). Statistical inference with non-probability survey samples (with discussions). *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer.

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465.

Zhang, G., and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3), 911-918.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2), 103-113.

Comments on “Statistical inference with non-probability survey samples”

Zhonglei Wang and Jae Kwang Kim[1]

Abstract

Statistical inference with non-probability survey samples is a notoriously challenging problem in statistics. We introduce two new methods of nonparametric propensity score technique for weighting in the non-probability samples. One is the information projection approach and the other is the uniform calibration in the reproducing kernel Hilbert space.

Key Words: Information projection; Uniform function calibration; Data integration.

1. Introduction

We would like to congratulate Dr. Changbao Wu on the outstanding work in non-probability sampling. Even though probability sampling served as a golden standard tool for finite population inference in the past decades, it has recently become tarnished gold due to low response rates and high costs. Non-probability sampling, on the other hand, is popular due to its feasibility and low cost (Couper, 2000; Kaplowitz, Hadlock and Levine, 2004). More importantly, non-probability sampling, such as a web survey, can quickly gather up-to-date information when compared to a probability sample. However, because the selection mechanism is unavailable for non-probability sampling, failing to correct the selection bias in analyzing a non-probability sample may result in inefficiency or even erroneous inference. As a result, adjusting the selection bias for a non-probability sample is a fundamental topic for survey sampling researchers, and this work presents the most comprehensive answers to this subject.

Dr. Wu’s research, in particular, includes a thorough examination of propensity score (PS) techniques. Those PS techniques, on the other hand, have drawbacks. First, even for a correctly specified PS model, the inverse probability weighting estimator may be inefficient due to small estimated propensity scores. One alternative is post-stratification, as stated in Section 5 of the paper, although there is no clear guidance on how to choose K. Furthermore, in practice, correctly specifying a PS model is difficult. While doubly robust estimation can help to safeguard a bad PS model, the final estimator is problematic when both the PS and regression models are incorrect (Kang and Schafer, 2007).

To overcome the misspecification of the PS model, Dr. Wu has mentioned several nonparametric methods, including a kernel method and a tree-based method. In this discussion, we would like to expand on this direction and provide two more methods to augment the study. One is based on a density ratio model using information projection (Csiszár and Shields, 2004), and the other is by uniformly calibrating functions over a reproducing kernel Hilbert space (RKHS). As explained by Wahba (1990), RKHS is a very flexible function space for approximation. Instead of estimating the propensity scores, we aim at

[1] Zhonglei Wang, Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen, Fujian, People's Republic of China; Jae Kwang Kim, Iowa State University, Ames, IA 50011, USA. E-mail: jkim@iastate.edu.

estimating the sampling weights $\left\{(\pi_i^A)^{-1} : i \in S_A\right\}$ to avoid possible inefficiency due to small estimated propensity scores.

Denote S_A and S_B to be the index sets for the non-probability and reference probability samples, respectively, and the corresponding sample sizes are n_A and n_B. Let $\left\{(y_i, \mathbf{x}_i) : i \in S_A\right\}$ and $\left\{(\mathbf{x}_i, d_i^B) : i \in S_B\right\}$ be available, where y_i and $\mathbf{x}_i$ are the study variable and auxiliary vector for the i^{th} unit and d_i^B is the design weight for $i \in S_B$.

The paper is organized as follows. In Section 2, we introduce the information projection approach. In Section 3, we introduce the basic idea of uniform calibration. Some concluding remarks are made in Section 4.

2. Information projection approach

Suppose that we are interested in estimating parameter $\boldsymbol{\theta}_0$ defined through $E_N\left\{U(\boldsymbol{\theta};\ \mathbf{X}, Y)\right\} = 0,$ where $E_N(\cdot)$ is the expectation with respect to the population empirical distribution $\Pr\{(\mathbf{X}, Y) = (\mathbf{x}_i, y_i)\} = N^{-1}$ for $i = 1, \ldots, N$ and 0 otherwise, and $U\left(\boldsymbol{\theta};\ \mathbf{x}, y\right)$ is a certain estimating function. For example, $U\left(\theta;\ \mathbf{x}, y\right) = y - \theta$ corresponds to $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ in the paper. We wish to obtain an estimator of $(\pi_i^A)^{-1},$ $\pi_i^A = \Pr\left(R_i = 1 \,|\, \mathbf{x}_i, y_i\right),$ and $R_i = 1$ if $i \in S_A$ and 0 otherwise.

To estimate $\left\{(\pi_i^A)^{-1} : i \in S_A\right\},$ we may use the relationship in the density ratio function. First, we consider a super-population model $\xi,$ and let $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y)$ be the density functions of $(\mathbf{x}, y)$ given $R = 0$ and $R = 1,$ respectively. Denote the density ratio function to be

$$r(\mathbf{x}, y) \;=\; \frac{f_0(\mathbf{x}, y)}{f_1(\mathbf{x}, y)},$$

and by the Bayes formula, we have

$$(\pi_i^A)^{-1} \;=\; 1 \;+\; \frac{\Pr(R_i = 0)}{\Pr(R_i = 1)}\; r(\mathbf{x}_i, y_i). \tag{2.1}$$

Thus, there is a one-to-one relationship between $(\pi_i^A)^{-1}$ and $r(\mathbf{x}_i, y_i).$

Under assumption A1, we can show that $r(\mathbf{x}, y) = r(\mathbf{x}).$ In this section, we make a more general assumption that there exists $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_L(\mathbf{x}))^{\mathrm{T}}$ such that

$$R \perp Y \,\big|\, \mathbf{b}(\mathbf{x}). \tag{2.2}$$

Rosenbaum and Rubin (1983) called $\mathbf{b}(\mathbf{x})$ in (2.2) balancing scores.

To estimate the density ratio function $r(\mathbf{x}),$ we minimize the Kullback-Leibler divergence

$$Q(f_0) \;=\; \int \log(f_0 / f_1)\, f_0 \,\mathrm{d}\mu \tag{2.3}$$

with respect to f_0 subject to some constraint, where both f_0 and f_1 are absolutely continuous with respect to a σ-finite measure $\mu.$ Regarding the constraint, we may use the following one

$$\Pr(R_i = 1) \int \mathbf{b}(\mathbf{x}) f_1(\mathbf{x}) \mu(d\mathbf{x}) \;+\; \Pr(R_i = 0) \int \mathbf{b}(\mathbf{x}) f_0(\mathbf{x}) \mu(d\mathbf{x}) \;=\; E_\xi\{\mathbf{b}(\mathbf{X})\}, \tag{2.4}$$

where $E_\xi(\cdot)$ is the expectation with respect to the super-population model $\xi.$ That is, given $f_1(\mathbf{x}),$ we can find $f_0(\mathbf{x})$ to minimize (2.3) under a calibration constraint with respect to $\mathbf{b}(\mathbf{x}).$

By Lemma 3.1 of Wang and Kim (2021), the optimized conditional density function satisfies

$$f_0^*(\mathbf{x}) \;=\; f_1(\mathbf{x}) \frac{\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}}{E_1\left[\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}\right]}, \tag{2.5}$$

where $\boldsymbol{\lambda}_1$ is chosen to satisfy (2.4). Note that the solution (2.5) is equivalent to

$$\log\{r(\mathbf{x};\, \boldsymbol{\lambda})\} \;=\; \lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}) \tag{2.6}$$

for the density ratio function $r(\mathbf{x}),$ where $\boldsymbol{\lambda} = (\lambda_0, \boldsymbol{\lambda}_1^{\mathrm{T}})^{\mathrm{T}},$ and λ_0 is a normalizing constant satisfying $\int r(\mathbf{x};\, \boldsymbol{\lambda}) f_1(\mathbf{x}) \mu(d\mathbf{x}) = 1.$ Thus, the information projection finds the best model for propensity score function.

Once the model is determined as in (2.6), we need to estimate the model parameters. Because of the moment constraints in (2.4), the sample-version estimating equation for $\boldsymbol{\lambda}$ is the calibration equation given by

$$\frac{n_A}{N} \sum_{i=1}^{N} R_i \left[1, \mathbf{b}(\mathbf{x}_i)\right] \left[1 + \frac{1 - n_A}{n_A} \exp\{\lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}\right] \;=\; \left[1, \frac{1}{N} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i)\right]. \tag{2.7}$$

Here, since $E_\xi\{\mathbf{b}(\mathbf{X})\}$ is not available, we use its estimate $N^{-1} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i).$ Once the parameter estimate $\hat{\boldsymbol{\lambda}}$ is obtained, we can construct

$$\hat{\omega}_i \;=\; 1 \;+\; \frac{1 - n_A}{n_A} \exp\{\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}$$

as the final PS weights. The parameter of interest can be estimated by solving $N^{-1} \sum_{i \in S_A} \hat{\omega}_i U(\boldsymbol{\theta};\, \mathbf{x}_i, y_i) = 0$ for $\boldsymbol{\theta}.$

Wang and Kim (2021) developed this framework under the non-probability sampling setup where $\mathbf{x}_i$ are available throughout the finite population. Consistency and the asymptotic normality can be developed under the assumption that $E\{U(\boldsymbol{\theta};\, \mathbf{x}, Y) \mid \mathbf{x}\}$ lies in the linear space generated by $\{b_1(\mathbf{x}), \ldots, b_L(\mathbf{x})\}.$ Instead of assuming the availability of $\{\mathbf{x}_i : i = 1, \ldots, N\}$ as in Wang and Kim (2021), there only exists a reference probability sample $\{(\mathbf{x}_i, d_i^B) :\, i \in S_B\}.$ If the probability sample S_B is a census, then the method above reduces to the one considered by Wang and Kim (2021), except that we consider a finite population parameter $\boldsymbol{\theta}_0.$ In Section 11.2 of Kim and Shao (2021), the information projection approach is called the maximum entropy method and applied to the data integration problem. In the simulation study presented in example 11.1 of the book, the proposed information projection method shows better performance than the methods of Chen, Li and Wu (2020) and Elliott and Valliant (2017).

3. Uniform calibration approach

Calibration is commonly used to improve the representativeness of a non-probability sample, but existing methods, including the information projection approach mentioned in Section 2, are based on calibrating a set of pre-specified functions. However, it is hard to correctly specify them for calibration in practice. In this section, we propose a general framework for uniformly calibrating functions in an RKHS. Instead of considering a parametric form for $E_\xi(Y \mid \mathbf{x})$ in (3.1), we only assume $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i)$, where $m(\mathbf{x})$ is a smooth function satisfying certain conditions.

We still consider (2.1) under the assumption A1. Instead of assuming a set of pre-specified functions $\mathbf{b}(\mathbf{x})$, we propose to estimate $\{r_i : i \in S_A\}$ by the following optimization,

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \geq 0}{\operatorname{argmin}} \left[\sup_{u \in H} \left\{ \frac{S(\boldsymbol{\gamma}, u)}{\|u\|_2^2} - \lambda_1 \frac{\|u\|_H^2}{\|u\|_2^2} \right\} + \lambda_2 Q_A(\boldsymbol{\gamma}) \right], \tag{3.1}$$

where $\boldsymbol{\gamma} = (r_1, \ldots, r_N)$, $r_i = 0$ for $i \notin S_A$, $\boldsymbol{\gamma} \geq 0$ is equivalent to $r_i \geq 0$ for $i = 1, \ldots, N$, H is an RKHS,

$$S(\boldsymbol{\gamma}, u) = \left[N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) r_i \right\} u(\mathbf{x}_i) - N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i) \right]^2, \tag{3.2}$$

$\|u\|_2^2 = (n_A + n_B)^{-1} \sum_{i \in S_A \cup S_B} u(\mathbf{x}_i)^2$, $\|u\|_H$ is the norm associated with the RKHS, $Q_A(\boldsymbol{\gamma})$ is a general penalty on $\boldsymbol{\gamma}$ to avoid overfitting, and λ_1 and λ_2 are two tuning parameters; see Wahba (1990) for a detailed introduction about the RKHS.

The intuition for the optimization (3.1) is briefly discussed. First, if r_i approximates the true density ratio $r(\mathbf{x}_i)$ well, the bias of the first term in (3.1) is negligible for estimating $N^{-1} \sum_{i=1}^N u(\mathbf{x}_i)$ for $u \in H$. Besides, $N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i)$ is design-unbiased. Thus, $S(\boldsymbol{\gamma}, u)$ balances two estimators for $N^{-1} \sum_{i=1}^N u(\mathbf{x}_i)$, and it is small if r_i approximately equals $r(\mathbf{x}_i)$ for $i \in S_A$. However, $S(\boldsymbol{\gamma}, u)$ is not scale invariant, and we have $S(\boldsymbol{\gamma}, cu) = c^2 S(\boldsymbol{\gamma}, u)$ for $c \in \mathbb{R}$. Thus, we use $\|u\|_2^2$ to make it scale-invariant. The term $\lambda_1 \|u\|_H^2$ is used to penalize the smoothness of the function u for $u \in H$. There exist different choices for $Q_A(\boldsymbol{\gamma})$. For example, $Q_A(\boldsymbol{\gamma}) = \sum_{i \in S_A} \{1 + (N n_A^{-1} - 1) r_i\}^2$ corresponds to penalizing extreme values for the sampling weights, and Wong and Chan (2018) investigated a similar problem assuming the availability of $\{\mathbf{x}_i : i = 1, \ldots, N\}$. The optimization (3.1) can be viewed as a "minmax" problem, and if $m \in H$, the estimated density ratios $\{\hat{r}_i : i \in S_A\}$ may lead to a reasonably good estimator

$$\hat{\mu}_{uc} = N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) \hat{r}_i \right\} y_i. \tag{3.3}$$

Uniform calibration is a new method for non-probability sampling, and there are some technical challenges in (3.1). For example, how to incorporate the design properties of S_B when establishing the theoretical properties of (3.3) has not be fully investigated, and we have finished a working paper about this topic (Wang, Mao and Kim, 2022). The kernel-based method is computationally expensive, especially when the sample sizes are large. It may be interesting to propose a more computationally efficient algorithm for the uniform calibration problem. One possible answer is to consider some other functional spaces, such as the one spanned by B-splines. In addition, it is also of interest to consider how to incorporate more than one reference probability sample, and how to formulate a uniform calibration if we have different covariates in different reference probability samples.

4. Concluding remarks

Propensity score weighting is an important tool for correcting selection bias in the nonprobability sampling. Dr. Changbao Wu made significant contributions on this important topic. In addition to the two additional methods, the empirical likelihood (EL) approach of Qin, Leung and Shao (2002) is potentially useful as another tool for propensity score weighting. In particular, the EL-based weighting method is applicable even under informative sampling. Further investigation on this direction will be explored elsewhere.

References

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

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Csiszár, I., and Shields, P.C. (2004). *Information Theory and Statistics: A Tutorial*.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Kang, J.D., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22(4), 523-539.

Kaplowitz, M.D., Hadlock, T.D. and Levine, R. (2004). A comparison of Web and mail survey response rates. *Public Opinion Quarterly*, 68(1), 94-101.

Kim, J.K., and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*, second edition. CRC press.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

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.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, H., and Kim, J.K. (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469*, 1-34.

Wang, Z., Mao, X. and Kim, J.K. (2022). Functional calibration under non-probability survey sampling. Submitted (https://arxiv.org/abs/2204.09193).

Wong, R.K., and Chan, K.C.G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1), 199-213.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 367-373
Statistics Canada, Catalogue No. 12-001-X

Author's response to comments on "Statistical inference with non-probability survey samples"

Changbao Wu[1]

Abstract

This response contains additional remarks on a few selected issues raised by the discussants.

Key Words: Data defect correlation; Double robustness; Inverse probability weighting; Model assumptions; Model-based prediction; Validation sample.

Let me start by thanking the Editor of *Survey Methodology*, Jean-François Beaumont, for organizing the discussions and putting together a glamour array of discussants. Each discussant looked at the topic of non-probability survey samples, and more generally topics on data integration and combining data from multiple sources, with some unique perspectives. I have enjoyed reading the discussions and I believe they are significant contributions to dealing with non-probability and other types of samples with selection bias. In what follows, I will make some additional remarks on a few selected issues raised by the discussants.

Michael A. Bailey

Dr. Bailey focused on the limitations of the estimation methods I presented under the assumptions A1-A4, and called for further development when these assumptions, and the so-called "MAR assumption" A1 in particular, are violated. Bailey used non-probabilistic polling as an example to argue that "non-response (can indeed) depends on the study variable" and the danger of A1 being violated is real.

While the criticism on the limitations of the methods reviewed in my paper is fair and square, the statements "(Wu) is fishing in one fairly specific corner of the pond" and "shying away from MNAR models" seem to show significant underappreciation on the importance of methodological development under the standard assumptions A1-A4 which were used by several authors on non-probability survey samples. First of all, the assumption A1 is on the participation (or inclusion/selection) mechanism for non-probability samples, which is not the same as "non-response". There are many scenarios where these assumptions can indeed be justified, especially for surveys using web- or phone-panels where the initial participation in those panels depends largely on certain demographic variables. Second, participation behaviour in non-probability surveys can be confounded by certain study variables during data collection in the same way we face in probability surveys on non-response, which is exactly how the current literature on non-probability surveys has been evolving in dealing with those issues. Third, any methodological advances in addressing the so-called "MNAR models" for non-probability surveys would require the foundation and thorough understanding established under the assumptions A1-A4.

[1] Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1.E-mail: cbwu@uwaterloo.ca.

Bailey also stated that “while MAR violations are a problem in probability sampling (arising due to non-response among randomly contacted individuals), MAR violations are more serious in a non-probability world”. I heartily concur. As a matter of fact, violations of the positivity assumption A2 are as serious as violations of the “MAR assumption” A1, and the two are intercorrelated. Violations of A2 imply that $\pi_i^A = P(i \in S_A \mid \mathbf{x}_i, y_i) = 0$ for some units in the target population, leading to the undercoverage problem that is as notorious as non-response. When A2 is violated but A1 holds, it is often believed that model-based prediction estimators can mitigate the biases due to undercoverage. Under the assumption A1 the sample inclusion indicator variable R and the study variable y are conditionally independent given $\mathbf{x}$, which implies that

$$E(y_i \mid \mathbf{x}_i, R_i = 1) = E(y_i \mid \mathbf{x}_i). \tag{1}$$

It follows that a valid prediction model $y \mid \mathbf{x}$ can be built using the observed data $\{(y_i, \mathbf{x}_i), i \in S_A\}$ (i.e., units with $R_i = 1$). Unfortunately, the equation (1) implicitly requires $P(R_i = 1 \mid \mathbf{x}_i) > 0$, and prediction-based estimators are not immune to potential undercoverage biases. Bailey’s call for “a framework that encompasses the possibility of MAR violations” is in line with some of the current research effort on dealing with undercoverage and “non-ignorable” participation mechanisms for non-probability survey samples. See, for instance, Chen, Li and Wu (2023), Cho, Kim and Qiu (2022) and Yuan, Li and Wu (2022), among others. In a nutshell, valid statistical inferences under those scenarios require either external data such as a validation sample or additional assumptions such as the existence of instrumental variables.

I am on the exact same page of discontent as Bailey with the “missing at random” label, since the term might be confused with “randomly missing” (Wu and Thompson, 2020, page 195). The term “ignorable” is also an unfortunate choice of terminology for missing data and causal inference literature, since it certainly cannot be ignored by the data analyst (Rivers, 2007). I use the standard term “propensity scores” for non-probability samples, while several other authors are in favour of “participation probabilities”, including Beaumont (2020) and Rao (2021).

Michael R. Elliott

Dr. Elliott discussed several issues with augmented materials and an expanded list of references. They are important additions to the current topic, especially the reviews on “additional approaches to combining data from probability and non-probability surveys” and sensitivity analysis on “unverifiable assumptions”.

Elliott’s discussions on distinctions between descriptive parameters and analytic parameters and weighting versus modelling raised the critical issue of efficiency of the IPW estimators in practice. It has been known for probability survey samples that the inverse probability weighted Horvitz-Thompson estimator of the population total T_y is extremely inefficient (in terms of large variance) when the sample selection probabilities π_i are unequal but have very weak correlation to the study variable y, although the estimator remains unbiased under such scenarios. Basu’s elephant example (Basu, 1971) described a "convincing case" where the inverse probability weighted and unbiased Horvitz-Thompson estimator failed miserably, leading to the dismissal of the circus statistician. Discussions on weighting versus modelling, i.e., the IPW estimators versus model-based prediction estimators for descriptive population parameters, are highly relevant for both theoretical developments and practical applications. Our job as a statistician in dealing with non-probability survey samples could be very much in limbo unless we develop solid guidelines and diagnostic tools for choosing suitable approaches with the given dataset and inferential problems.

Elliott echoed my call for a few large scale probability surveys with rich information on auxiliary variables with the statement "it is increasingly critical for an organized and ideally government funded stable of high-quality probability surveys to be put into place for routine data collection". His comments on new areas of research on issues with privacy and confidentiality due to the need for microdata under the context of analyzing non-probability survey samples are a visionary call and deserve an increased amount of attention from the research community.

Zhonglei Wang and Jae Kwang Kim

Dr. Wang and Dr. Kim presented two new approaches to propensity score based estimation, one uses the so-called information projection through a density ratio model and the other employs uniformly calibration functions over a reproducing kernel Hilbert space. These are new adventures in the field, and Kim and his collaborators have the experience and the analytic power to move the research forward.

The starting point for both approaches is the following equation which connects the propensity scores to the density ratios,

$$\frac{1}{P(R_i = 1 \mid \mathbf{x}_i, y_i)} = 1 + \frac{P(R_i = 0)}{P(R_i = 1)} \frac{f_0(\mathbf{x}_i, y_i)}{f_1(\mathbf{x}_i, y_i)}.$$

The propensity scores $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i)$ only require the model on $R_i = 1$ given $\mathbf{x}_i$ and y_i. Justification of the equation given above, however, requires a joint randomization framework involving both the model q for the propensity scores and the superpopulation model ξ on $(\mathbf{x}, y)$. From a consistency view point regarding the final estimator of the finite population mean of y, the joint framework imposes very little restrictions if the density ratios are modelled nonparametrically. The consequential impact of the approach is on variance and variance estimation. Variance of an estimator under a joint randomization framework involves more than one component, and variance estimation has further complications if nonparametric procedures are involved. Efficiency comparisons between the proposed methods and some of the existing methods need to be carried out under suitable settings. I am eager to see further developments on the proposed methods.

Sharon L. Lohr

Dr. Lohr's extended discussions on diagnostic tools for assessing model assumptions are highly valuable to the topic. Her explorations of existing ideas and methods and the adaptations to the current setting highlight the seemingly different but deeply connected issues faced by both nonprobability and probability survey samples. One such issue is the undercoverage problem (i.e., violations of assumption A2) and the interweave of assumptions A1 and A2. Lohr was rightfully concerned with prediction based estimators where the prediction model of y given $\mathbf{x}$ is built based on the nonprobability sample S_A and the mass imputation estimator is computed using observed $\mathbf{x}$ in the reference probability sample S_B, a scenario where each of the two assumptions A1 and A2 does not stand alone. The undercoverage problem is an example where "space-age procedures will not rescue stone-age data". Lohr advocated to "take a small probability sample to investigate assumptions", which is of necessity in theory since rigorously defendable methods under certain scenarios require validation samples. Developments of compromising strategies with existing data sources, however, are more appealing but also more challenging in practice.

Lohr's observation "nonprobability samples have the potential to improve data equity" is an important one, since inclusion of units from groups which may be invisible in probability samples can be boosted relatively easily for nonprobability samples. Lohr also observed that "historically disadvantaged groups may be underrepresented in all data sources, including (nonprobability samples)". Addressing the issue of data equity with nonprobability survey samples presents both opportunities and challenges.

Lohr's question "when should one use nonprobability samples" is a tough one. The same question can be asked for any other statistical methods. We do not seem to always question the validity of the methods and the usefulness of the results in many other scenarios due to our unchecked confidence that the required assumptions seem to be reasonable. For nonprobability samples, we have a more vulnerable situation regarding assumptions, and assessments and diagnostics of these assumptions are more difficult than cases with controlled experiments and/or more structured data. From this view point, Lohr's extended discussion on assessing assumptions should be read with deep appreciation. In practice, an important confidence booster on the assumptions is the thorough investigation at the "design stage", if such a stage can be conceived prior to data collection, on variables which might be related to participation behaviour, and to include these variables as part of the sample with further exploration of existing data sources containing these variables.

Xiao-Li Meng

Dr. Meng's discussion, with the formal title "Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples", should be a standalone discussion paper itself. Meng weaved through a number of issues in estimating a finite population mean with a nonprobability sample, and explored strategies and directions for constructing an approximately unbiased estimator using the central concept of the so-called *data defect correlation (ddc)*. The discussions are fascinating and thought-provoking, and will surely generate more discussions and research endeavours on implications of the *ddc*. I would like to use this opportunity to comment briefly on the *ddc* in relation to three basic concepts in probability sampling: *sampling strategy*, *undercoverage*, and *model-assisted estimation*. It is not a nostalgia for the good old days when probability sampling was the golden standard but rather an appreciation of how research in survey sampling has been evolving and the potential usefulness of the *ddc* in dealing with nonprobability survey samples.

The term *sampling strategy* refers to the pair of sampling design and estimation method (Thompson, 1997, Section 2.4; Rao, 2005, Section 3.1). The two components go hand in hand and are the backbone of conventional probability survey sampling theory. For the estimation of the population total T_y of the study variable y using a probabiliity sample S with first order inclusion probabilities π_i, the Horvitz-Thompson estimator $\hat{T}_{y\text{HT}} = \sum_{i \in S} d_i y_i$ with the weight $d_i = \pi_i^{-1}$ is the unique unbiased estimator among a class of linear estimators (Wu and Thompson, 2020). The theoretical argument for the result is straightforward due to the known inclusion probabilities π_i under the given sampling design. Using the notation of Meng, the *ddc* involves three variables: the study variable G, the weight variable W, the sample inclusion indicator R, and is defined as the finite population correlation coefficient between $\tilde{R} = RW$ and G. The *ddc* implicitly puts R and W as an inseparable pair for any *inference strategy*, with R corresponding to the unknown "design" and W for the "estimation method". With the unknown "design" characterized by the unknown "divine probabilities" π_I for the nonprobability sample, Meng showed through his equation (3.3) that $W_I \propto \pi_I^{-1}$ is essentially a required condition for unbiased estimation of $\bar{G}$ if nothing is assumed on the outcome regression model. The result provides a justification of the use of inverse probability weighted (IPW) estimator for nonprobability samples as the only sensible choice if a superpopulation model on the study variable is not involved.

The problem of *undercoverage* has been discussed extensively in the existing literature on probability sampling. For nonprobability samples the issue is closely related to the violation of the positivity assumption A2 as discussed in Section 7.2 of my paper and my comments to the discussions of Bailey, Elliott and Lohr, with additional details given in Chen et al. (2023). Let $U = U_0 \cup U_1$, where U_1 is the uncovered subpopulation with $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i) = 0$. Let $N = N_0 + N_1$, where N_0 and N_1 are respectively the sizes of the two subpopulations U_0 and U_1. Let Cov_I and $\text{Cov}_I^{(0)}$ denote respectively the covariance with respect to the discrete uniform distribution over U and U_0. It can be shown that

$$\text{Cov}_I(\tilde{R}_I, G_I) = \omega_0 \left\{ \text{Cov}_I^{(0)}(\tilde{R}_I, G_I) - \omega_1 (\bar{G}_1 - \bar{G}_0)\, \hat{N}_0 \big/ N_0 \right\}, \qquad (2)$$

where $\omega_k = N_k / N$ for $k = 0, 1$, $\hat{N}_0 = \sum_{i \in S} W_i$, S is the set of units for the nonprobability sample, and $\bar{G}_0$ and $\bar{G}_1$ are respectively the population means of U_0 and U_1 for the study variable G. Equation (2) has two immediate implications. First, if the estimation method is valid in the sense that the value of $\text{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ is small, then the bias of the estimator $\bar{G}_W$ due to undercoverage depends on ω_1 (i.e., the size of the uncovered subpopulation U_1) and $\bar{G}_1 - \bar{G}_0$ (i.e., the difference between U_0 and U_1), a statement which has previously been established under probability sampling. Second, the equation reveals a scenario for potential *counterbalancing*: A biased estimator $\bar{G}_W$ for the "sampled population mean" $\bar{G}_0$ can be less biased for the target population mean $\bar{G}$ if $\text{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ and $\bar{G}_1 - \bar{G}_0$ have the same plus or minus sign.

Meng's discussions on quasi-randomization and/or super-population using the *ddc* provided a much deeper understanding on doubly robust estimation. Historically, *model-assisted estimation* started to emerge in survey sampling in the early 1970s, and the approach has the same spirit of double robustness. The generalized difference estimator of the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ as discussed in Cassel, Särndal and Wretman (1976) is given by

$$\hat{\mu}_{y\text{GD}} = \frac{1}{N}\left\{\sum_{i\in S}\frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i\right\}, \tag{3}$$

where S is a probability sample, the π_i's are the first order inclusion probabilities, and $\{c_1, c_2, \ldots, c_N\}$ is an arbitrary sequence of known numbers. The estimator $\hat{\mu}_{y\text{GD}}$ is exactly unbiased for μ_y under the probability sampling design p for any sequence c_i, and is also model-unbiased if we choose $c_i = m_i = E_\xi(y_i \mid \mathbf{x}_i)$. Cassel et al. (1976) showed a main theoretical result that the choice $c_i = m_i$ is optimal leading to minimum model-based expectation of the design-based variance $E_\xi\{V_p(\hat{\mu}_{y\text{GD}})\}$ when the model has certain structure in variance. The first part of the results on unbiasedness is under $(p$ or $\xi)$; the second part on optimality is under $(p$ and $\xi)$. Note that the estimator $\hat{\mu}_{y\text{GD}}$ with the choice $c_i = \hat{m}_i$ has exactly the same structure of the doubly robust estimator discussed extensively in the missing data and causal inference literature since the 1990s, with the "divine probabilities" π_i being unknown and estimated in the latter cases.

The use of *ddc* in practice requires additional information from the population. Meng's proposal of creating a representative miniature out of a biased sample echoes the call for a validation sample with a small size, since such a sample "can (also) eliminate many practitioners's anxiety and potential mistakes for not knowing how to properly use the weights".

"There is no such thing as probability sample in real life" is probably a defendable statement for human populations. Probability samples, however, do exist in other fields such as business and establishment surveys, agricultural surveys, and natural resource inventory surveys; see Wu and Thompson (2020) for further detail. For humans, any rigorous rules and precise procedures are *almost surely* as aspiration, not prescription.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part One. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto, 203-242.

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

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

Chen, Y., Li, P. and Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Survey Methodology*, under review.

Cho, S., Kim, J.K. and Qiu, Y. (2022). *Multiple Bias Calibration for Valid Statistical Inference with Selection Bias*. Working paper.

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

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

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 1-26.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, London.

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

Yuan, M., Li, P. and Wu, C. (2022). *Inference with Non-Ignorable Sample Inclusion for Non-Probability Survey Samples*. Working paper.

Are deep learning models superior for missing data imputation in surveys? Evidence from an empirical comparison

Zhenhua Wang, Olanrewaju Akande, Jason Poulos and Fan Li[1]

Abstract

Multiple imputation (MI) is a popular approach for dealing with missing data arising from non-response in sample surveys. Multiple imputation by chained equations (MICE) is one of the most widely used MI algorithms for multivariate data, but it lacks theoretical foundation and is computationally intensive. Recently, missing data imputation methods based on deep learning models have been developed with encouraging results in small studies. However, there has been limited research on evaluating their performance in realistic settings compared to MICE, particularly in big surveys. We conduct extensive simulation studies based on a subsample of the American Community Survey to compare the repeated sampling properties of four machine learning based MI methods: MICE with classification trees, MICE with random forests, generative adversarial imputation networks, and multiple imputation using denoising autoencoders. We find the deep learning imputation methods are superior to MICE in terms of computational time. However, with the default choice of hyperparameters in the common software packages, MICE with classification trees consistently outperforms, often by a large margin, the deep learning imputation methods in terms of bias, mean squared error, and coverage under a range of realistic settings.

Key Words: Deep learning; Hyperparameter selection; Missing data; Multiple imputation by chained equations; Simulation studies; Survey data.

1. Introduction

Many sample surveys suffer from missing data, arising from unit nonresponse, where a subset of participants do not complete the survey, or item nonresponse, where missing values are concentrated on particular questions. In opinion polls, nonresponse may reflect either refusal to reveal a preference or lack of a preference (De Leeuw, Hox and Huisman, 2003). If not properly handled, missing data patterns can lead to biased statistical analyses, especially when there are systematic differences between the observed data and the missing data (Rubin, 1976; Little and Rubin, 2019). Complete case analysis on units with completely observed data is often infeasible and may lead to large bias in most situations (Little and Rubin, 2019). As a result, many analysts account for the missing data by imputing missing values and then proceeding as if the imputed values are true values.

Multiple imputation (MI) (Rubin, 1987) is a popular approach for handling missing values. In MI, an analyst creates $L > 1$ completed datasets by replacing the missing values in the sample data with plausible draws generated from the predictive distribution of probabilistic models based on the observed data. In each completed dataset, the analyst can then compute sample estimates for population estimands of interest, and combine the sample estimates across all L datasets using MI inference methods developed by Rubin (1987), and more recently, Rubin (1996), Barnard and Meng (1999), and Reiter and

[1] Zhenhua Wang is PhD student in the Department of Statistics, University of Missouri, Columbia, MO 65211. E-mail: zhenhua.wang@mail.missouri.edu; Olanrewaju Akande is research scientist at Meta Platforms, Inc. E-mail: akandelanre13@gmail.com; Jason Poulos is Postdoctoral Associate in the Department of Health Care Policy, Harvard Medical School, Boston, MA. E-mail: poulos@hcp.med.harvard.edu; Fan Li is Professor in the Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708. E-mail: fl35@duke.edu.

Raghunathan (2007), and Harel and Zhou (2007). In MI, the estimated variance of an estimand consists of both within-imputation and between-imputation variances, and thus takes into account the inherent variability of the imputed values. Note that in survey studies, single imputation, e.g., via matching or regression, remains to be common for dealing with missing data, where the variance is estimated via the delta method or resampling methods (Chen and Haziza, 2019; Haziza and Vallée, 2020).

1.1 Model-based imputation

There are two general modeling strategies for MI. The first strategy, known as *joint modeling* (JM), is to specify a joint distribution for all variables in the data, and then generate imputations from the implied conditional (predictive) distributions of the variables with missing values (Schafer, 1997). The JM strategy aligns with the theoretical foundation of Rubin (1987), but it can be challenging to specify a joint model with high-dimensional variables of different types. Indeed, most popular JM approaches, such as "PROC MI" in SAS (Yuan, 2011), and "AMELIA" (Honaker, King and Blackwell, 2011) and "norm" in `R` (Schafer, 1997), make a simplifying assumption that the data follow multivariate Gaussian distributions, even for categorical variables, which can lead to bias (Horton, Lipsitz and Parzen, 2003). Recent research developed flexible JM models based on advanced Bayesian nonparametric models such as Dirichlet Process mixtures (Manrique-Vallier and Reiter, 2014; Murray and Reiter, 2016). However, these methods are computationally expensive, and often struggle to scale up to high-dimensional cases.

The second strategy is called *fully conditional specification* (FCS, van Buuren, Brand, Groothuis-Oudshoorn and Rubin (2006)), where one separately specifies a univariate conditional distribution for each variable with missing values given all the other variables and imputes the missing values variable-by-variable iteratively, akin to a Gibbs sampler. The most popular FCS method is multiple imputation by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn, 2011), usually implemented with specifying generalized linear models (GLMs) for the univariate conditional distributions (Raghunathan, Lepkowski, Van Hoewyk and Solenberger, 2001; Royston and White, 2011; Su, Gelman, Hill and Yajima, 2011). Recent research indicates that specifying the conditional models by classification and regression trees (CART, Breiman, Friedman, Olshen and Stone (1984) and Burgette and Reiter (2010)) comprehensively outperforms MICE with GLM (Akande, Li and Reiter, 2017). A natural extension of MICE with CART is to use ensemble tree methods such as random forests, rather than a single tree (Breiman, 2001; Doove, Van Buuren and Dusseldorp, 2014).

MICE is appealing in large-scale survey data because it is simple and flexible in imputing different types of variables. However, MICE has a key theoretical drawback that the specified conditional distributions may be incompatible, that is, they do not correspond to a joint distribution (Arnold and Press, 1989; Gelman and Speed, 1993; Li, Yu and Rubin, 2012). Despite this drawback, MICE works remarkably well in real applications and numerous simulations have demonstrated it outperforms many theoretically sound JM-based methods; see van Buuren (2018) for case studies. However, MICE is also computationally intensive (White, Royston and Wood, 2011) and generally cannot be parallelized. Moreover, popular software packages for implementing MICE with GLMs, e.g., `mice` in `R` (van Buuren and Groothuis-Oudshoorn, 2011), often crash in settings with high dimensional non-continuous variables, e.g., categorical variables with many categories (Akande et al., 2017).

1.2 Imputation with deep learning models

Recent advances in deep learning greatly expand the scope of complex models for high-dimensional data. This advancement brings the hope that a new generation of missing data imputation methods based on deep learning models may address the theoretical and computational limitations of existing statistical methods. For example, deep generative models such as generative adversarial networks (GANs, Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio (2014)) are naturally suitable for producing multiple imputations because they are designed to generate data that resemble the observed data as much as possible. A method in this stream is the generative adversarial imputation network (GAIN) of Yoon, Jordon and Schaar (2018). Multiple imputation using denoising autoencoders (MIDA, Gondara and Wang (2018) and Lu, Perrone and Unpingco (2020)), is another generative method based on deep neural networks trained on corrupted input data in order to force the networks to learn a useful low-dimensional representation of the input data, rather than its identity function (Vincent, Larochelle, Bengio and Manzagol, 2008; Vincent, Larochelle, Lajoie, Bengio, Manzagol and Bottou, 2010). Several methods have been proposed for missing value imputation in time-series data using variational autoencoders (Fortuin, Baranchuk, Rätsch and Mandt, 2020) or recurrent neural networks (Lipton, Kale and Wetzel, 2016; Monti, Bronstein and Bresson, 2017; Cao, Wang, Li, Zhou, Li and Li, 2018; Che, Purushotham, Cho, Sontag and Liu, 2018; Yoon, Zame and van der Schaar, 2018).

Deep learning based MI methods have several advantages, at least theoretically, over the traditional statistical models, including (i) they avoid making distributional assumptions; (ii) can readily handle mixed data types; (iii) can model nonlinear relationships between variables; (iv) are expected to perform well in high-dimensional settings; and (v) can leverage graphics processing unit (GPU) power for faster computation. Several papers report encouraging performance of deep learning based MI methods compared to MICE (e.g., Yoon, Jordon and Schaar, 2018). However, such conclusions are made based on limited evidence. First, the studies are usually based on small simulations or several well-studied public "benchmark" datasets, such as those described in Section 5, which do not resemble survey data. Second, the evaluations are usually based on a few overall performance metric, e.g., the overall predictive mean squared error or accuracy. Such metrics may not give a full picture of the comparisons and sometimes can be even misleading, as will be illustrated later. Third, given the uncertainty of the missing data process, it is crucial to examine the repeated sampling properties of imputation methods, but these have been rarely evaluated. Finally, hyperparameter tuning is crucial for machine learning models and different tuning can result in dramatically different results, but few details are provided on hyperparameter tuning and its consequences on the performance of imputation methods.

Motivated by these limitations, in this paper we carry out extensive simulations based on real survey data to evaluate MI methods with a range performance metrics. Specifically, we conduct simulations based on a subsample from the American Community Survey to compare repeated sampling properties of four aforementioned MI methods: MICE with CART (MICE-CART), MICE with random forests (MICE-RF), GAIN, and MIDA. We find that deep learning based MI methods are superior to MICE in terms of computational time. However, MICE-CART consistently outperforms, often by a large margin, the deep learning methods in terms of bias, mean squared error, and coverage, under a range of realistic settings. This contradicts previous findings in the machine learning literature, and raises questions on the appropriate metrics for evaluating imputation methods. It also highlights the importance of assessing repeated-sampling properties of imputation methods. Though we focus on multiple imputation in this paper, we note that the aforementioned MI methods are readily applicable to generate single imputation when L is set to 1. Extensive empirical evidences suggest that the within-imputation variance usually dominates the between-imputation variance in MI. As such, we expect the patterns between different imputation methods observed here also stand if these methods are used for single imputation.

The remainder of this article is organized as follows. In Section 2, we review the four MI methods used in our evaluation. In Section 3, we describe a framework with several metrics for evaluating imputation methods. In Section 4, we describe the simulation design and results with large-scale survey data, and in Section 5 we summarize evaluation results on the benchmark datasets used in machine learning literature. Finally, in Section 6, we conclude with a practical guide for implementation in real applications.

2. Missing data imputation methods

We first introduce notation. Consider a sample with n units, each of which is associated with p variables. Let Y_{ij} be the value of variable j for individual $i,$ where $j=1,\ldots,p$ and $i=1,\ldots,n.$ Here, Y can be continuous, binary, categorical or mixed binary-continuous. For each individual $i,$ let $\mathbf{Y}_i=\left(Y_{i1},\ldots,Y_{ip}\right).$ For each variable $j,$ let $\mathbf{Y}_j=\left(Y_{1j},\ldots,Y_{nj}\right).$ Let $\mathbf{Y}=\left(\mathbf{Y}_1,\ldots,\mathbf{Y}_n\right)$ be the $n\times p$ matrix comprising the data for all records included in the sample. We write $\mathbf{Y}=\left(\mathbf{Y}_{\text{obs}},\mathbf{Y}_{\text{mis}}\right),$ where $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ are respectively the observed and missing parts of $\mathbf{Y}.$ We write $\mathbf{Y}_{\text{mis}}=\left(\mathbf{Y}_{\text{mis},1},\ldots,\mathbf{Y}_{\text{mis},p}\right),$ where $\mathbf{Y}_{\text{mis},j}$ represents all missing values for variable $j,$ with $j=1,\ldots,p.$ Similarly, we write $\mathbf{Y}_{\text{obs}}=\left(\mathbf{Y}_{\text{obs},1},\ldots,\mathbf{Y}_{\text{obs},p}\right)$ for the corresponding observed data.

In MI, the analyst generates values of the missing data $\mathbf{Y}_{\text{mis}}$ using pre-specified models estimated with $\mathbf{Y}_{\text{obs}},$ resulting in a completed dataset. The analyst then repeats the process to generate L completed datasets, $\left\{\mathbf{Y}^{(l)}\colon l=1,\ldots,L\right\},$ that are available for inference or dissemination. For inference, the analyst can compute sample estimates for population estimands in each completed dataset $\mathbf{Y}^{(l)},$ and combine them using MI inference rules developed by Rubin (1987), which will be reviewed in Section 3.

2.1 MICE with classification tree models

Under MICE, the analyst begins by specifying a separate univariate conditional model for each variable with missing values. The analyst then specifies an order to iterate through the sequence of the conditional models, when doing imputation. We write the ordered list of the variables as $\left(\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(p)}\right).$ Next, the analyst initializes each $\mathbf{Y}_{\text{mis},(j)}.$ The most popular options are to sample from (i) the marginal distribution of the corresponding $\mathbf{Y}_{\text{obs},(j)}$, or (ii) the conditional distribution of $\mathbf{Y}_{(j)}$ given all the other variables, constructed using only available cases.

After initialization, the MICE algorithm follows an iterative process that cycles through the sequence of univariate models. For each variable j at each iteration t, one fits the conditional model $(\mathbf{Y}_{(j)} \big| \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}: k < j\}, \{\mathbf{Y}_{(k)}^{(t-1)}: k > j\})$. Next, one replaces $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ with draws from the implied model $(\mathbf{Y}_{\text{mis},(j)}^{(t)} \big| \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}: k < j\}, \{\mathbf{Y}_{(k)}^{(t-1)}: k > j\})$. The iterative process continues for T total iterations until convergence, and the values at the final iteration make up a completed dataset $\mathbf{Y}^{(l)} = \left(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(T)}\right)$. The entire process is then repeated L times to create the L completed datasets. We provide pseudocode detailing each step of the MICE algorithm in the supplementary material.

Under MICE-CART, the analyst uses CART (Breiman et al., 1984) for the univariate conditional models in the MICE algorithm. CART follows a decision tree structure that uses recursive binary splits to partition the predictor space into distinct non-overlapping regions. The top of the tree often represents its root and each successive binary split divides the predictor space into two new branches as one moves down the tree. The splitting criterion at each leaf is usually chosen to minimize an information theoretic entropy measure. Splits that do not decrease the lack of fit by an reasonable amount based on a set threshold are pruned off. The tree is then built until a stopping criterion is met; e.g., minimum number of observations in each leaf.

Once the tree has been fully constructed, one generates $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ by traversing down the tree to the appropriate leaf using the combinations in $(\{\mathbf{Y}_k^{(t)}: k < j\}, \{\mathbf{Y}_k^{(t-1)}: k > j\})$, and then sampling from the $Y_{(j)}^{\text{obs}}$ values in that leaf. That is, given any combination in $(\{\mathbf{Y}_k^{(t)}: k < j\}, \{\mathbf{Y}_k^{(t-1)}: k > j\})$, one uses the proportion of values of $\mathbf{Y}_j^{\text{obs}}$ in the corresponding leaf to approximate the conditional distribution $(\mathbf{Y}_{(j)} \big| \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}: k < j\}, \{\mathbf{Y}_{(k)}^{(t-1)}: k > j\})$. The iterative process again continues for T total iterations, and the values at the final iteration make up a completed dataset.

MICE-RF instead uses random forests for the univariate conditional models in MICE (e.g., Stekhoven and Bühlmann, 2012; Shah, Bartlett, Carpenter, Nicholas and Hemingway, 2014). Random forests (Ho, 1995; Breiman, 2001) is an ensemble tree method which builds multiple decision trees to the data, instead of a single tree like CART. Specifically, random forests constructs multiple decision trees using bootstrapped samples of the original, and only uses a sample of the predictors for the recursive partitions in each tree. This approach can reduce the prevalence of unstable trees as well as the correlation among individual trees significantly, since it prevents the same variables from dominating the partitioning process across all trees. Theoretically, this decorrelation should result in predictions with less variance (Hastie, Tibshirani and Friedman, 2009).

For imputation, the analyst first trains a random forests model for each $\mathbf{Y}_{(j)}$ using available cases, given all other variables. Next, the analyst generates predictions for $\mathbf{Y}_{\text{mis},j}$ under that model. Specifically, for any categorical $\mathbf{Y}_{(j)}$, and given any particular combination in $(\{\mathbf{Y}_k^{(t)}: k < j\}, \{\mathbf{Y}_k^{(t-1)}: k > j\})$, the analyst first generates predictions for each tree based on the values $\mathbf{Y}_j^{\text{obs}}$ in the corresponding leaf for that tree, and then uses the most commonly occurring majority level of among all predictions from all the trees. For a continuous $\mathbf{Y}_{(j)}$, the analyst instead uses the average of all the predictions from all the trees. The iterative process again cycles through all the variables, for T total iterations, and the values at the final iteration make up a completed dataset. A particularly important hyperparameter in random forests is the maximum number of trees d.

For our evaluations, we use the `mice` R package to implement both MICE-CART and MICE-RF, and retain the default hyperparameter setting in the package to mimic the common practice in real world applications. Specifically, we set the minimum number of observations in each terminal leaf to 5 and the pruning threshold to 0.0001 in MICE-CART. In MICE-RF, the maximum number of trees d is set to be 10.

2.2 Generative Adversarial Imputation Network (GAIN)

GAIN (Yoon, Jordon and Schaar, 2018) is an imputation method based on GANs (Goodfellow et al., 2014), which consist of a generator function G and a discriminator function D. For any data matrix $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, we replace $\mathbf{Y}_{\text{mis}}$ with random noise, Z_{ij}, sampled from a uniform distribution. The generator G inputs this initialized data and a mask matrix $\mathbf{M}$, with $M_{ij} \in \{0, 1\}$ indicating observed values of $\mathbf{Y}$, and outputs predicted values for both the observed data and missing data, $\hat{\mathbf{Y}}$. The discriminator D utilizes $\hat{\mathbf{Y}} = (\mathbf{Y}_{\text{obs}}, \hat{\mathbf{Y}}_{\text{mis}})$ and a hint matrix $\mathbf{H}$ of the same dimension to identify which values are observed or imputed by G, which results in a predicted mask matrix $\hat{\mathbf{M}}$. The hint matrix, sampled from the Bernoulli distribution with p equal to a "hint rate" hyperparameter, reveals to D partial information about $\mathbf{M}$ in order to help guide G to learn the underlying distribution of $\mathbf{Y}$.

We first train D to minimize the loss function, $L_D(\mathbf{M}, \hat{\mathbf{M}})$, for each mini-batch of size n_i:

$$L_D(\mathbf{M}, \hat{\mathbf{M}}) = \sum_{i=1}^{n_i} \sum_{j=1}^{J} M_{ij} \log(\hat{M}_{ij}) + (1 - M_{ij}) \log(1 - \hat{M}_{ij}). \tag{2.1}$$

Next, G is trained to minimize the loss function (2.2), which is composed of a generator loss, $L_G(\mathbf{M}, \hat{\mathbf{M}})$, and a reconstruction loss, $L_M(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M})$. The generator loss (2.3) is minimized when D incorrectly identifies imputed values as being observed. The reconstruction loss (2.4) is minimized when the predicted values are similar to the observed values, and is weighted by the hyperparameter β:

$$L(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}, \hat{\mathbf{M}}) = L_G(\mathbf{M}, \hat{\mathbf{M}}) + \beta L_M(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}), \tag{2.2}$$

$$L_G(\mathbf{M}, \hat{\mathbf{M}}) = \sum_{i=1}^{n_i} \sum_{j=1}^{J} M_{ij} \log(1 - \hat{M}_{ij}), \tag{2.3}$$

$$L_M(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}) = \sum_{i=1}^{n_i} \sum_{j=1}^{J} (1 - M_{ij}) L_{\text{rec}}(Y_{ij}, \hat{Y}_{ij}), \tag{2.4}$$

where

$$L_{\text{rec}}(Y_{ij}, \hat{Y}_{ij}) = \begin{cases} (\hat{Y}_{ij} - Y_{ij})^2 & \text{if } Y_{ij} \text{ is continuous} \\ -Y_{ij} \log \hat{Y}_{ij} & \text{if } Y_{ij} \text{ is categorical.} \end{cases} \tag{2.5}$$

In our experiments, we model both G and D as fully-connected neural networks, each with three hidden layers, and θ hidden units per hidden layer. The hidden layer weights are initialized uniformly at random with the Xavier initialization method (Glorot and Bengio, 2010). We use leaky ReLU activation function (Maas, Hannun and Ng, 2013) for each hidden layer, and a softmax activation function for the output layer for G in the case of categorical variables, or a sigmoid activation function in the case of numerical variables and for the output of D. We facilitate this choice of output layer for numerical variables by transforming all continuous variables to be within range (0, 1) using the MinMax normalization: $Y_{ij}^* = \left\{Y_{ij} - \min(Y_{\cdot j})\right\} \big/ \left\{\max(Y_{\cdot j}) - \min(Y_{\cdot j})\right\}$, where $\min(Y_{\cdot j})$ and $\max(Y_{\cdot j})$ are the minimum and maximum of variable j, respectively. After imputation, we transform each value back to its original scale. We generate multiple imputations using several runs of the model with varying initial imputation of the missing values.

To implement GAIN in our evaluations, we use the same architecture as the one in Yoon, Jordon, and Schaar (2018). We set $\beta = 100$, θ equal to the number of features of the input data, and tune the hint rate on a single simulation. Following the common practice in the GAN literature (Berthelot, Schumm and Metz, 2017; Ham, Jun and Kim, 2020), we track the evolution of GAIN's generator and discriminator losses, and manually tune the hint rate so that the two losses are qualitatively similar. Specifically, we first coarsely select the hint rate among {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. Then we determine the final value by an additional fine tuning step. In the MAR scenario, for example, after observing that the optimal value is in the range (0.1, 0.2), we perform a search among {0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19}. Finally, we set the optimal hint rate for MCAR and MAR scenarios to be 0.3 and 0.13, respectively. We train the networks for 200 epochs using stochastic gradient descent (SGD) and mini-batches of size 512 to learn the parameter weights. We use the Adam optimizer to adapt the learning rate, with an initial rate of 0.001 (Kingma and Ba, 2014).

2.3 Multiple Imputation using Denoising Autoencoders (MIDA)

MIDA (Gondara and Wang, 2018; Lu et al., 2020) extends a class of neural networks, denoising autoencoders, for MI. An autoencoder is a neural network model trained to learn the identity function of the input data. Denoising autoencoders intentionally corrupt the input data in order to prevent the networks from learning the identity function, but rather a useful low-dimensional representation of the input data. The MIDA architecture consists of an encoder and decoder, each modeled as a fully-connected neural network with three hidden layers, with θ hidden units per hidden layer. We first perform an initial imputation on missing values using the mean for continuous variables and the most frequent label for categorical variables, which results in a completed data $\mathbf{Y}_0$. The encoder inputs $\mathbf{Y}_0$, and corrupts the input data by randomly dropping out half of the variables. The corrupted input data is mapped to a higher dimensional representation by adding Θ hidden units to each successive hidden layer of the encoder. The decoder receives output from the encoder, and symmetrically scales the encoding back to the original input dimension. All hidden layers use a hyperbolic tangent (tanh) activation function, while the output layer of the decoder uses a softmax (sigmoid) activation function in the case of categorical (numerical) variables. Multiple imputations are generated by using multiple runs with the hidden layer weights initialized as a Gaussian random variable.

Following Lu et al. (2020), we train MIDA in two phases: a primary phase and fine-tuning phase. In the primary phase, we feed the initially imputed data to MIDA and train for N_{prime} epochs. In the fine-tuning phase, MIDA is trained for N_{tune} epochs on the output in the primary phase, and produces the outcome. The loss function is used in both phases and closely resembles the reconstruction loss in GAIN:

$$L\left(Y_{ij_0}, \hat{Y}_{ij}, M_{ij}\right) = \begin{cases} \left(1-M_{ij}\right)\left(Y_{ij_0}-\hat{Y}_{ij}\right)^2 & \text{if } Y_{ij} \text{ is continuous} \\ -\left(1-M_{ij}\right) Y_{ij_0} \log \hat{\mathbf{Y}}_{ij} & \text{if } Y_{ij} \text{ is categorical.} \end{cases} \tag{2.6}$$

To implement MIDA in our evaluations, we use the same architecture and tune the hyperparameters in a single simulation as in Lu et al. (2020). We plot the evolution of loss function L, and select the number of additional units Θ among {1, 2, 3, 4, 5, 6, 7 ,8, 9, 10} to reduce the loss. In our experiments, we set θ equal to the number of features of the input data and add $\Theta = 7$ hidden units to each of the three hidden layers of the encoder. We train the model for $N_{\text{prime}} = 100$ epochs in the primary phase and $N_{\text{tune}} = 2$ epochs in the fine-tuning phase. Similar as in GAIN, we learn the model parameters using SGD with mini-batches of size 512, and use the Adam optimizer to adapt the learning rate with the initial rate being 0.001.

3. Simulation-based evaluation of imputation methods

Methods for missing data imputation are usually evaluated via real-data based simulations (van Buuren, 2018). Namely, one creates missing values from a complete dataset according to a missing data mechanism (Little and Rubin, 2014), imputes the missing values by a specific method, and then compares these imputed values with the original "true" values based on some metrics.

We first give a quick review of Rubin's MI combination rules. Let Q be the target estimand in the population, and $q^{(l)}$ and $u^{(l)}$ be the point and variance estimate of Q based on the l^{th} imputed dataset, respectively. The MI point estimate of Q is $\bar{q}_L = \sum_{l=1}^{L} q^{(l)} / L$, and the corresponding estimate of the variance is equal to $T_L = (1 + 1/L) b_L + \bar{u}_L$, where $b_L = \sum_{l=1}^{L} \left(q^{(l)} - \bar{q}_L\right)^2 / (L-1)$, and $\bar{u}_L = \sum_{l=1}^{L} u^{(l)} / L$. The confidence interval of Q is constructed using $\left(\bar{q}_L - Q\right) \sim t_v\left(0, T_L\right)$, where t_v is a t-distribution with $v = (L-1)\left(1 + \bar{u}_L / \left[(1+1/L)\, b_L\right]\right)^2$ degrees of freedom.

The first step in our simulation-based evaluation procedure is choosing a dataset with all values observed, which is taken as the "population". We then choose a set of target estimands Q and compute their values from this population data, which are taken as the "ground truth". The estimands are usually summary statistics of the variables or parameters in a down-stream analysis model, e.g., a coefficient in a regression model (Tang, Song, Belin and Unützer, 2005; Huque, Carlin, Simpson and Lee, 2018). Second, we randomly draw without replacement H samples of size n from the population data, and in each of sample $(h = 1, \ldots, H)$ create missing data according to a specific missing data mechanism and pre-fixed proportion of missingness. Third, for each simulated sample with missing data, we create L imputed datasets using the imputation method under consideration and construct the point and interval estimate of each estimand using Rubin's rules. Lastly, we compute performance metrics of each estimand from the quantities obtained in the previous step.

In the empirical application, we select a large complete subsample from the American Community Survey (ACS) – a national survey that bears the hallmarks of many big survey data – as our population. Since discrete variables are prevalent in the ACS, as well as in most survey data, we focus on the marginal probabilities of binary and categorical variables; e.g., a categorical variable with K categories has $K-1$ estimands. To evaluate how well the imputation methods preserve the multivariate distributional properties, similar to Akande et al. (2017), we also consider the bivariate probabilities of all two-way combinations of categories in binary and categorical variables. Another useful metric is the finite-sample pairwise correlations between continuous variables. For continuous variables, the common estimands are mean, median or variance. To facilitate meaningful comparisons of the results between the categorical and continuous variables, we propose to discretize each continuous variable into K categories based on the sample quantiles. We then evaluate these binned continuous variables as categorical variables based on the aforementioned estimands of marginal and bivariate probabilities.

For each estimand Q, we consider three metrics. The first metric focuses on bias. To accommodate close-to-zero estimands that are prevalent in probabilities of categorical variables, we consider the absolute standardized bias (ASB) of each estimand Q:

$$\text{ASB} = \sum_{h=1}^{H} \left| \bar{q}_L^{(h)} - Q \right| \Big/ \left(H \cdot Q \right), \tag{3.1}$$

where $\bar{q}_L^{(h)}$ is the MI point estimate of Q in simulation h.

The second metric is the relative mean squared error (Rel.MSE), which is the ratio between the MSE of estimating Q from the imputed data and that from the sampled data before introducing the missing data:

$$\text{Rel.MSE} = \frac{\sum_{h=1}^{H} \left(\bar{q}_L^{(h)} - Q \right)^2}{\sum_{h=1}^{H} \left(\tilde{Q}^{(h)} - Q \right)^2}, \tag{3.2}$$

where $\bar{q}_L^{(h)}$ is defined earlier, and $\tilde{Q}^{(h)}$ is the prototype estimator of Q, i.e., the point estimate from the complete sampled data in simulation h.

The third metric is coverage rate, which is the proportion of the $\alpha\%$ (e.g., 95%) confidence intervals, denoted by CI_h^α $\left(h = 1, \ldots, H \right)$, in the H simulations that contain the true Q:

$$\text{Coverage} = \sum_{h=1}^{H} \mathbf{1} \left\{ Q \in \text{CI}_h^\alpha \right\} \Big/ H. \tag{3.3}$$

We recommend conducting a large number of simulations (e.g., $H \geq 100$) to obtain reliable estimates of MSE and coverage. This would not be a problem for deep learning algorithms, which can be typically completed in seconds even with large sample sizes. However, it can be computationally prohibitive for the

MICE algorithms when each of the simulated data is large (e.g., $n = 100{,}000$ in some of our simulations). In the situation that one has to rely on only a few or even a single simulation for evaluation, we propose a modified metric of bias. Specifically, for each categorical variable or binned continuous variable j, we define the weighted absolute bias (WAB) as the sum of the absolute bias weighted by the true marginal probability in each category:

$$\text{Weighted absolute bias} = \sum_{k=1}^{K} Q_{jk} \left| \bar{q}_{jk}^{(h)} - Q_{jk} \right|, \tag{3.4}$$

where K is the total number of categories, Q_{jk} is the population marginal probability of category k in variable j, and $\bar{q}_{jk}^{(h)}$ is its corresponding point estimate in simulation h. We can also average the weighted absolute bias over a number of repeatedly simulated samples.

The above procedure and metrics differ from the common practice in the machine learning literature. For example, many machine learning papers on missing data imputation conduct simulations on benchmark datasets, but these data often have vastly different structure and features from survey data and thus are less informative for the goal of this paper. One such dataset is the Breast Cancer dataset in the UCI Machine Learning Repository (Dua and Graff, 2017), which has only 569 sample units and no categorical variables. Also, these simulations are usually based on randomly creating missing values of a single dataset repeatedly rather than on drawing repeated samples from a population, and thus fails to account for the sampling mechanism. Moreover, these evaluations often use metrics focusing on accuracy of individual predictions rather than distributional features. Specifically, the most commonly used metrics are the root mean squared error (RMSE) and accuracy (Gondara and Wang, 2018; Yoon, Jordon and Schaar, 2018; Lu et al., 2020). Both metrics can be defined in an overall or variable-specific fashion, but the machine learning literature usually focuses on the overall version. The overall RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j} M_{ij} \left(\hat{Y}_{ij} - Y_{ij} \right)^2}{\sum_{i=1}^{n} \sum_{j} M_{ij}}}, \tag{3.5}$$

where Y_{ij} is the value of continuous variable j for individual i in the complete data before introducing missing data, and $\hat{Y}_{ij}$ is the corresponding imputed value. For non-missing values (i.e., $M_{ij} = 1$), $Y_{ij} = \hat{Y}_{ij}$. The (overall) accuracy is defined for categorical variables, namely it is the proportion of the imputed values being equal to the corresponding original "true" value:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \sum_{j \in S_{\text{cat}}} M_{ij} \mathbf{1}\left(\hat{Y}_{ij} = Y_{ij} \right)}{\sum_{i=1}^{n} \sum_{j \in S_{\text{cat}}} M_{ij}}, \tag{3.6}$$

where S_{cat} is the set of categorical variables.

A number of caveats are in order for the RMSE and accuracy metrics. First, they are usually computed on a single imputed sample as an overall measure of an imputation method, but this ignores the uncertainty of imputations. Second, both RMSE and accuracy are single value summaries and do not capture the multivariate distributional feature of data. Third, RMSE does not adjust for the different scale of variables and can be be easily dominated by a few outliers; also, it is often computed without differentiating between continuous and categorical variables. Lastly, when there are multiple (L) imputed data, a common way is to use the mean of the L imputed value as $\hat{Y}_{ij}$ in (3.5), but the statistical meaning of the resulting metrics is opaque. This is particularly problematic for categorical variables. For these reasons, we warn against using the overall RMSE and accuracy as the only metrics for comparing imputation methods, and one should exercise caution when interpreting them.

4. Evaluation based on ACS

In this section, we evaluate the four imputation methods described in Section 2 following the procedure and metrics described in Section 3. For simplicity, in the following discussions we use CART and RF to denote MICE-CART and MICE-RF, respectively.

4.1 The "population" data

We use the one-year Public Use Microdata Sample from the 2018 ACS to construct our population. The 2018 ACS data contains both household-level variables – for example, whether or not a house is owned or rented – and individual-level variables – for example, age, income and sex of the individuals within each household. Since individuals nested within a household are often dependent, and the imputation methods we evaluate generally assume independence across all observations, we set our unit of observation at the household-level, where independence is more likely to hold. We first remove units corresponding to vacant houses. Next, we delete units with any missing values, so that we only keep the complete cases. Within each household, we also retain individual-level data corresponding only to the household head and merge them with the household-level variables, resulting in a rich set of variables with potentially complex joint relationships.

It is often challenging to generate plausible imputations for ordinal variables with many levels when there is very low mass at the highest levels, as is the case for some variables in the ACS data. Following Li, Baccini, Mealli, Zell, Frangakis and Rubin (2014), we treat ordinal variables with more than 10 levels as continuous variables. We also follow the approach in Akande et al. (2017) to exclude binary variables where the marginal probabilities violate $np > 10$ or $n(1-p) > 10$; this eliminates estimands where the central limit theorem is not likely to hold. For each categorical variable with more than two levels but less than 10 levels where this might also be a problem, we merge the levels with a small number of observations in the population data. For example, for the household language variable, we recode the levels from five to three (English, Spanish, and other), because the probability of speaking neither English nor Spanish in the full population is less than 8.8%.

The final population data contains 1,257,501 units, with 18 binary variables, 20 categorical variables with 3 to 9 levels, and 8 continuous variables. We describe the variables in more detail in the supplementary material. We compute the population values of the estimands Q described in Section 3, including all marginal and bivariate probabilities of discrete and binned continuous variables. We vary the size of the simulated samples from 10,000 to 100,000, and simulate missing data according to either missing completely at random (MCAR) or missing at random (MAR) mechanisms in each of these scenarios.

4.2 Simulations with *n* = 10,000

We first randomly draw $H = 100$ samples of size $n = 10{,}000$, and set 30% of each sample to be missing under either MCAR or MAR. CART or RF takes around 2.8 and 9.2 hours, respectively, to create $L = 10$ imputed datasets with default parameters on a standard desktop computer with a single central processing unit (CPU). The deep learning methods are much faster because they leverage GPU computing power when implemented on the GPU-enabled TensorFlow software framework (Abadi, Agarwal, Barham, Brevdo, Chen, Citro, Corrado, Davis, Dean, Devin, Ghemawat, Goodfellow, Harp, Irving, Isard, Jia, Jozefowicz, Kaiser, Kudlur, Levenberg, Mané, Monga, Moore, Murray, Olah, Schuster, Shlens, Steiner, Sutskever, Talwar, Tucker, Vanhoucke, Vasudevan, Viégas, Vinyals, Warden, Wattenberg, Wicke, Yu and Zheng, 2015). GAIN takes roughly 1.5 minutes and MIDA takes roughly 4 minutes to create $L = 10$ completed datasets using a GeForce GTX 1660 Ti GPU. Note that it is infeasible to manually tune the hyperparameter in each of the 100 simulations in each scenario for the deep learning models. So for each scenario, we have randomly selected one simulation, and tune the hyperparameters using the procedure described in Section 2. We then apply these selected hyperparameters to all simulations.

4.2.1 MCAR scenario

To create the MCAR scenario, we randomly set 30% of the values of each variable to be missing independently. Table 4.1 displays the distributions of the estimated ASB and relative MSE of all the marginal and bivariate probabilities in the imputed data by the four imputation methods.

Overall, for the estimands of marginal and bivariate probabilities of the categorical and binned continuous variables, MICE with CART significantly outperforms all other three methods, with consistently yielding the smallest ASB and relative MSE. RF is the second best, also consistently outperforming the deep learning methods. The advantage of the MICE algorithms is particularly pronounced in the upper (e.g., 75% and 90%) quantiles, indicating that GAIN and MIDA imputations have large variations over repeated samples and variables. Indeed, MIDA and GAIN lead to ultra long tails in estimating the summary statistics of the variables. For example, for bivariate probabilities of binned continuous variables, the 90% percentile of the ASB from MIDA and GAIN is approximately 20 and 27 times, respectively, of that from CART. The discrepancy is even bigger for relative MSE. There is no consistent pattern in comparing MIDA and GAIN. Specifically, for continuous variables, MIDA generally outperforms GAIN, but the difference is small except for the upper percentiles, where GAIN tends to produce very large bias and relative MSE. For categorical variables, GAIN outperforms MIDA half of the time, but again leads to the largest variation in imputations across the variables. Moreover, an interesting and somewhat surprising observation is that MICE with CART consistently outperforms RF – sometimes by a large magnitude – regardless of the choice of estimand or metric.

Table 4.1
Distributions of absolute standardized bias $(\times 100)$ and relative mean squared error of all marginal and bivariate probabilities based on the imputations by the four MI methods, when $n = 10{,}000$ and 30% values MCAR

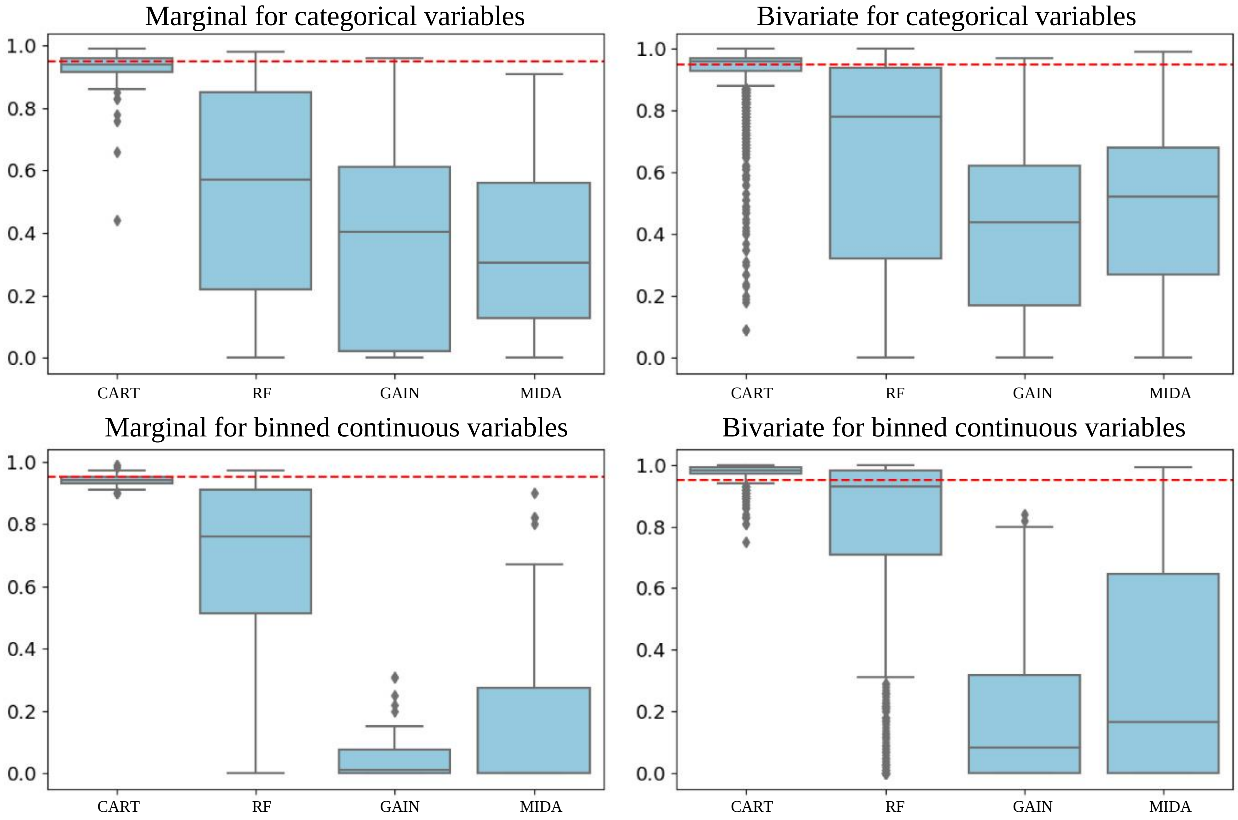
	Quantiles		Marginal				Bivariate			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
ASB $(\times 100)$	Cat.	10%	0.05	0.47	0.76	0.98	0.15	1.14	1.21	1.54
		25%	0.13	1.25	1.48	2.22	0.40	2.83	3.08	3.93
		50%	0.27	2.80	3.22	4.69	1.05	6.74	7.14	8.47
		75%	0.64	5.86	7.18	8.86	2.51	13.59	17.03	15.23
		90%	1.14	10.01	19.55	14.41	5.34	22.33	26.92	21.90
	B.Cont.	10%	0.06	0.24	7.25	2.73	0.19	1.30	6.05	4.80
		25%	0.10	1.05	12.86	8.36	0.43	3.24	17.61	12.01
		50%	0.21	3.59	27.30	18.51	1.02	6.61	34.29	24.07
		75%	0.43	5.43	30.21	26.84	1.90	11.76	49.38	39.54
		90%	0.81	8.49	46.41	31.36	3.42	20.79	90.90	64.65
Rel.MSE	Cat.	10%	1.05	1.67	2.50	3.38	0.96	1.11	2.75	2.98
		25%	1.16	2.40	4.97	9.03	1.08	1.61	4.33	4.75
		50%	1.37	5.99	10.37	14.89	1.25	3.35	7.40	8.16
		75%	1.49	10.25	27.73	26.16	1.48	9.07	14.87	15.80
		90%	1.62	16.22	97.33	40.16	1.89	23.91	36.37	27.92
	B.Cont.	10%	1.19	1.50	44.06	4.35	0.82	0.86	7.40	2.05
		25%	1.30	1.77	74.42	13.82	0.92	1.11	14.80	4.90
		50%	1.44	3.31	139.24	72.57	1.07	1.90	32.26	13.76
		75%	1.55	6.71	284.00	150.35	1.26	4.09	88.78	47.56
		90%	1.64	19.69	603.38	451.44	1.54	10.80	282.29	127.15

"Cat." means categorical variables and "B.Cont." means binned continuous variables.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

All methods generally yield less biased estimates (i.e., smaller ASB) of the marginal probabilities than the bivariate probabilities. This illustrates preserving multivariate distributional features is more challenging than univariate ones. The advantage of CART over the other methods is comparatively larger when estimating bivariate estimands than univariate ones. Interestingly, the relative MSE tends to be higher for the marginal probabilities than the bivariate probabilities. This is likely due to the fact that the denominator in the definition of relative MSE in (3.2) is the MSE from the sampled data before introducing missing data, which tends to be smaller for marginal probabilities than bivariate probabilities. CART yields MSEs that are very close to the corresponding MSEs from the sampled data before introducing missing data; i.e., the relative MSE is close to 1. On the contrary, both deep learning methods, and GAIN in particular, can result in exceedingly large relative MSE for many estimands.

Figures 4.1 displays the estimated coverage rates of the 95% confidence intervals for the marginal and bivariate probabilities. The patterns on coverage between different methods is similar to those on bias and MSE. Specifically, CART tends to result in coverage rates that are close to the nominal 95% level, with the median consistently being around 95% and tight interquartile range. In contrast, RF, GAIN and MIDA all result in coverage rates that are much farther off from the nominal 95% level. For example, the median coverage rates under both GAIN and MIDA are all under 0.60, and are even less than 0.30 for continuous variables. A closer look into the prediction accuracy of each variable reveals that GAIN and MIDA tend to generate imputations that are biased toward the most frequent levels, and GAIN in particular generally produces narrower intervals than the other methods. This once again provides evidence of significant bias under the deep learning methods. All methods tend to result in higher median coverage rates for the bivariate probabilities than the marginal probabilities, although the left tails are generally longer for the former than the latter.

Figure 4.1 Coverage rate of the 95% confidence interval for all marginal and bivariate probabilities obtained from the four imputation methods in the simulations with *n* = 10,000 and 30% values MCAR.



The red dashed line is 0.95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.2.2 MAR scenario

We also consider a MAR scenario, which is more plausible than MCAR in practice. We set six variables – age, gender, marital status, race, educational attainment and class of worker – to be fully observed. It would be cumbersome to specify different MAR mechanism for each of the remaining 40 variables, so we randomly divide them into three groups, consisting of 10, 15, and 15 variables. We then specify a separate nonresponse model by which to generate the missing data for the variables in each group. Specifically, we postulate a logistic model per group, conditional on the fully observed six variables, based on which we then generate binary missing data indicators for each variable in that group. This process results in approximately 30% missing rate for each of the 40 variables. We describe the models in more detail in the supplementary material.

Table 4.2 displays the distributions of the ASB and relative MSE of all the marginal and bivariate probabilities from the four methods. All methods yield larger ASB and relative MSE under the MAR scenario than the previous MCAR scenario. This is expected because MAR is a stronger assumption than

MCAR that requires conditioning on more information. Nonetheless, the overall patterns of relative performance between the methods remain the same as those under MCAR. Specifically, CART once again produces estimates with the least ASB and relative MSE – by an even larger margin then under MCAR – among the four methods, followed by RF, and then MIDA and GAIN. One notable observation is the deteriorating performance of the deep learning methods, particularly GAIN, in imputing continuous variables, sometimes resulting in several hundreds fold of relative MSE than CART. This indicates the huge uncertainties associated with GAIN in imputing continuous variables.

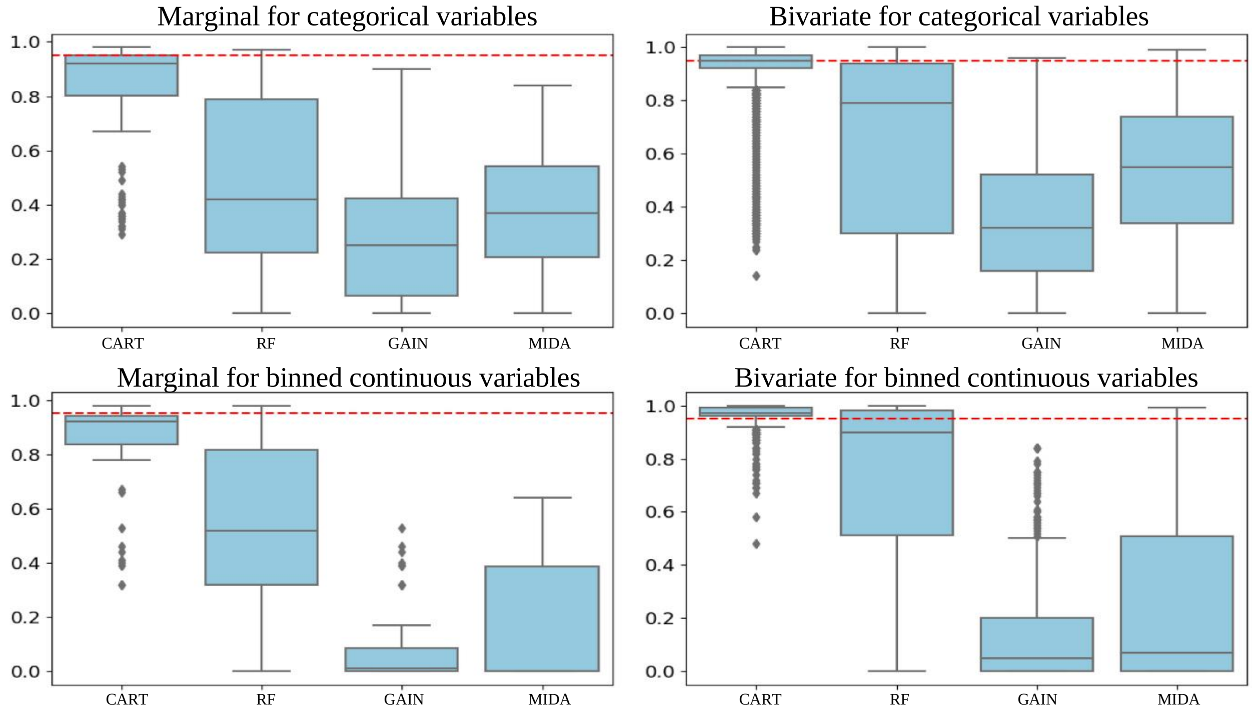
Table 4.2
Distributions of absolute standardized bias $(\times 100)$ and relative mean squared error for all methods, when *n* = 10,000 and 30% values MAR, over all possible marginal and bivariate probabilities

	Quantiles		Marginal				Bivariate			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
ASB $(\times 100)$	Cat.	10%	0.05	0.13	0.15	0.14	0.15	0.71	0.76	0.89
		25%	0.11	0.44	0.62	0.61	0.40	2.23	2.55	3.20
		50%	0.29	2.13	3.05	4.55	1.08	6.06	6.85	8.14
		75%	1.04	4.98	6.63	10.22	2.49	13.43	16.78	16.19
		90%	1.80	10.49	18.91	17.00	5.68	24.06	28.04	25.36
	B.Cont.	10%	0.07	0.29	0.33	0.33	0.27	1.17	10.87	6.18
		25%	0.17	1.07	9.64	3.13	0.69	3.49	23.67	16.26
		50%	0.67	3.14	32.86	23.85	1.58	7.83	38.52	31.17
		75%	1.20	6.95	39.57	36.09	3.40	15.20	53.59	47.34
		90%	3.40	12.39	63.45	41.99	5.94	25.16	97.47	85.44
Rel.MSE	Cat.	10%	1.00	1.00	1.00	1.00	0.97	1.00	1.53	1.93
		25%	1.08	1.82	2.56	4.75	1.04	1.39	3.78	4.03
		50%	1.33	4.33	19.03	15.13	1.25	3.00	10.42	8.38
		75%	1.72	13.08	55.07	33.36	1.59	9.56	27.45	16.95
		90%	2.27	18.70	101.91	48.44	2.23	27.44	64.01	32.85
	B.Cont.	10%	1.00	1.00	1.00	1.00	0.88	0.90	11.19	2.96
		25%	1.38	1.83	90.98	8.49	1.00	1.16	20.15	6.87
		50%	1.70	4.57	207.58	96.08	1.18	2.29	45.25	21.33
		75%	2.12	11.47	692.67	239.69	1.50	6.95	125.39	70.90
		90%	3.12	50.56	1342.23	806.43	2.12	18.07	459.78	205.14

"Cat." means categorical variables and "B.Cont." means binned continuous variables.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Figures 4.2 displays the estimated coverage rates of the 95% confidence intervals for the marginal and bivariate probabilities, under each method. Similar as the case of bias and MSE, all methods generally result in lower coverage rates under MAR than MCAR, with visibly longer left tails in some cases, but the overall patterns comparing between the methods remain the same. Specifically, CART still tends to result in coverage rates that are above 90%, while the other three methods have consistently lower coverage rate. In particular, both GAIN and MIDA result in extremely low – below 7% – median coverage rates for continuous variables. This is closely related to the previous observation of the large uncertainty of the deep learning methods in imputing continuous variables.

Figure 4.2 Coverage rates of the 95% confidence intervals for all marginal and bivariate probabilities obtained from four methods in the simulations with *n* = 10,000 and 30% values MAR.



The red dashed line is 0.95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Finally, to illustrate that evaluating only the overall RMSE and accuracy metrics may be misleading, we display the mean and empirical standard errors of the overall RMSE and accuracy over the 100 simulations in Table 4.3, where MCAR is in the top panel and MAR is in the bottom panel. Under both missing data mechanisms, for the continuous variables, MIDA leads to the smallest overall RMSE, followed by CART, and with RF and GAIN being last. For the categorical variables, CART and GAIN lead to the highest overall accuracy, with MIDA being closely behind and RF last. These patterns, not surprisingly, differ from those reported earlier based on marginal and bivariate probabilities and different metrics. As discussed in Section 3, overall RMSE and accuracy do not capture the distributional features of multivariate data or the repeated sampling properties of the imputation methods.

Table 4.3
Overall RMSE on continuous variables and overall accuracy on categorical variables averaged over 100 simulations

Mechanism	Metric	CART	RF	GAIN	MIDA
MCAR	RMSE	0.128 (0.002)	0.159 (0.003)	0.161 (0.008)	0.112 (0.002)
	Accuracy	0.785 (0.001)	0.658 (0.003)	0.782 (0.002)	0.752 (0.004)
MAR	RMSE	0.130 (0.003)	0.154 (0.004)	0.145 (0.009)	0.110 (0.002)
	Accuracy	0.819 (0.001)	0.704 (0.003)	0.820 (0.002)	0.780 (0.007)

The empirical standard errors in the parenthesis.
The top panel is under MCAR and the bottom panel is under MAR, all with 30% missing data.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.3 Simulations with *n* = 100,000 and 30% MCAR

Deep learning models usually require a large sample size to train. Therefore, to give MIDA and GAIN a more favorable setting as well as to investigate the sensitivity of our results to variations in sample size, we generate a simulation scenario of $H = 10$ samples with $n = 100{,}000$ under MCAR. That is, we randomly set 30% of the values of each variable to be missing independently. Here we only generate 10 simulations due to the huge computational cost of MICE for samples with this size. In this scenario, we omit RF because the previous results in Section 4.2 have shown that RF is consistently inferior to CART in terms of performance and computation. We use CART, GAIN, and MIDA to create $L = 10$ completed datasets.

Because it usually requires a much larger number of simulations to reliably calculate MSE and coverage, here we focus on the weighted absolute bias metric (3.4). Table 4.4 displays the distributions of the estimated weighted absolute bias, averaged over 10 simulations, of the marginal probabilities of the categorical and binned continuous variables. Overall, the patterns comparing between the four methods remain consistent with those observed in Section 4.2. Specifically, CART again results in the smallest weighted absolute difference in both categorical and continuous variables, and the advantage is particularly pronounced with continuous variables. For example, for categorical variables, MIDA and GAIN result in a median of weighted absolute bias at least 9 and 11 times, respectively, larger than CART. The advantage of CART grows to about 30 and 60 times over MIDA and GAIN, respectively, for continuous variables. Moreover, CART performs robustly across variables, evident from the small variation in the weighted absolute bias, e.g., 0.07 for 10% percentile and 0.33 for 90% percentile among the categorical variables. In contrast, both deep learning models result in much larger variation across variables; e.g., 0.57 for 10% percentile and 2.92 for 90% percentile among the categorical variables under MIDA, and even larger for GAIN. In summary, other than computational time, MICE with CART significantly outperforms MIDA and GAIN in terms of bias and variance regardless of the sample size.

Table 4.4
Distributions of the weighted absolute bias $(\times 100)$ averaged over 10 simulated samples, each with *n* = 100,000 and 30% values MCAR

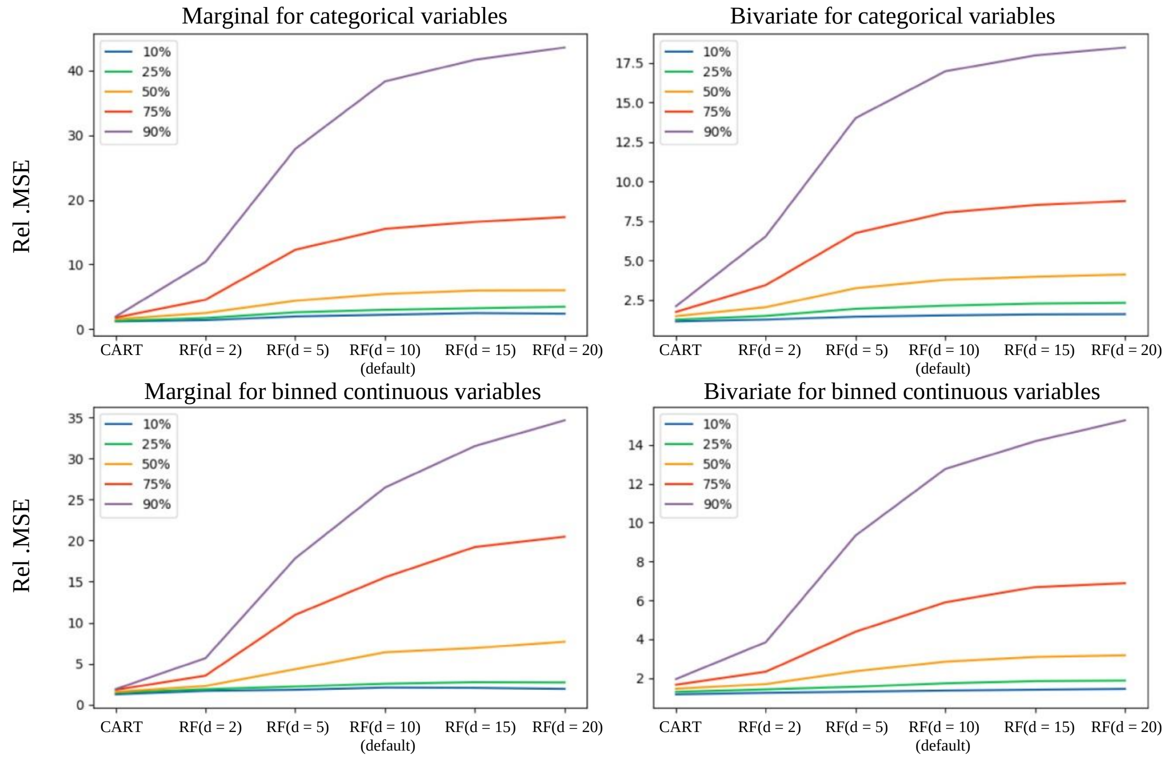
Quantiles	Categorical			Binned Continuous		
	CART	GAIN	MIDA	CART	GAIN	MIDA
10%	0.07	0.43	0.57	0.10	5.52	1.98
25%	0.11	1.11	1.02	0.11	6.65	2.78
50%	0.15	1.74	1.40	0.12	7.36	4.04
75%	0.24	3.77	2.07	0.13	9.40	6.50
90%	0.33	4.63	2.92	0.15	11.31	7.72

CART = Classification and regression trees; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.4 Role of hyperparameters in tree-based MICE

The pattern that CART outperforms RF is surprising, because the common knowledge is that ensemble methods are usually superior to single tree methods. But the same pattern was also observed in another recent study (Wongkamthong and Akande, 2021). We investigate the role of the key hyperparameter in RF – the maximum number of trees d – in the simulations. We randomly selected a simulated data of size $n =$ 10,000 and 30% of entries being MCAR. We use the `mice` package to fit RF with different number of trees: $d = 2, 5, 10, 15, 20,$ where $d = 10$ is the default setting. The relative MSE of the imputed categorical variables fitted using each d value, as well as that using CART, is shown as trajectories in Figure 4.3, which reveals a consistent pattern: the upper quantiles – particularly those above 50% – of the relative MSE deteriorates rapidly as the maximum number of trees in RF increases, while the lower quantiles, e.g., 10%, 25%, remain stable. We found a similar pattern with the standardized bias metric and continuous variables, and thus the results are omitted here. This suggests that larger number of trees in RF – at least as implemented in the `mice` package – leads to much longer tail in the distribution of the bias and MSEs. This is likely due to overfitting. We cannot exclude the possibility that a more customized hyperparameter tuning of RF may outperform CART in some applications. However, such case-specific fine-tuning of the MICE algorithm is generally not available for the vast majority of MI consumers who relies on the default setting of popular packages like `mice`.

Figure 4.3 Quantiles of the relative mean squared error over all marginal and bivariate probabilities of categorical and binned continuous variables, under CART and RF with various number of trees, for a simulation sample with *n* = 10,000 and 30% values MCAR.



5. Evaluation based on "benchmark" datasets

To verify the evaluations in the GAIN and MIDA papers (Gondara and Wang, 2018; Yoon, Jordon and Schaar, 2018; Lu et al., 2020), we also compared the two deep learning models with CART based on the five benchmark datasets and simulation procedure (different from our proposed framework) used in these papers. Details of these datasets and simulations are presented in the supplementary material. The sample sizes of these data are generally not large enough to be considered as population data from which we can repeatedly sample from without replacement, so we are unable to evaluate them in a meaningful way using absolute standardized bias, relative MSE or coverage. We therefore evaluate the methods primarily on the weighted absolute bias metric. In summary, CART again consistently and significantly outperforms MIDA and GAIN in terms of weighted absolute bias for both categorical and continuous variables, across all five benchmark datasets. The difference in performance is particularly pronounced with continuous variables. We also calculated the overall MSE and accuracy as those papers did. Except for one dataset, we could not reproduce the results reported in these papers, even with the authors' code. One possible reason is that the process of tuning and selecting model hyperparameters may not be clearly documented, which is true in the present case. More details are provided in the online supplementary material.

6. Conclusion

Recent years have seen the development of many machine learning based methods for imputing missing data, raising the hope of improving over the more traditional imputation methods such as MICE. However, efforts in evaluating these methods in real world situations remain scarce. In this paper, we adopt an evaluation framework real-data-based simulations. We conduct extensive simulation studies based on the American Community Survey to compare repeated sampling properties of two MICE methods and two deep learning imputation methods based on GAN (GAIN) and denoising autoencoders (MIDA).

We find that the deep learning models hold a vast computational advantage over MICE methods, partially because they can leverage GPU power for high-performance computing. However, our simulations as well as evaluation on several "benchmark" data suggest that MICE with CART specification of the conditional models consistently outperforms, usually by a substantial margin, the deep learning models in terms of bias, mean squared error, and coverage under a wide range of realistic settings. In particular, GAIN and MIDA tend to generate unstable imputations with enormous variations over repeated samples compared with MICE. One possible explanation is that deep neural networks excel at detecting complex sub-structures of big data, but may not suit for data with simple structure, such as the simulated data used here. Another possibility is that the sample sizes in our simulations are not adequate to train deep neural networks, which usually required much more data compared to traditional statistical models.

These results contradict previous findings based on the single performance metric of overall mean squared error in the machine learning literature (e.g., Gondara and Wang, 2018; Yoon, Jordon and Schaar, 2018; Lu et al., 2020). This discrepancy highlights the pitfalls of the common practice in the machine learning literature of evaluating imputation methods. It also demonstrates the importance of assessing repeated-sampling properties on multiple estimands of MI methods. An interesting finding is that ensemble trees (e.g., RF) do not improve over a single tree (e.g., CART) in the context of MICE, which matches the findings in another recent study (Wongkamthong and Akande, 2021). Combined with the fact that the former is more computationally intensive than the latter, we recommend using MICE with CART instead of RF in practice.

Our study has a few limitations. First, there are many deep learning methods that can be adapted to missing data imputation and all may have different operating characteristics. We choose GAIN and MIDA because both generative adversarial network and denoising autoencoders are immensely popular deep learning methods, and the imputation methods based on them have been advertised as superior to MICE. Nonetheless, it would be desirable to examine other deep learning based imputation methods in future research. Second, performance of machine learning methods is highly dependent on hyperparameter selection. So it can be argued that the inferior performance of GAIN and MIDA may be at least partially due to sub-optimal hyperparameter selection. However, practitioners would most likely rely on default hyperparameter values for any machine learning based imputation methods, which is indeed what we have adopted in our simulations and thus represents the real practice. Third, we did not consider the joint distribution between any categorical and continuous variables, but our evaluations within categorical and continuous variables have yielded consistent conclusions. Lastly, as any simulation study, one should exercise caution in generalizing the conclusions. By carefully selecting the data and metrics, we have attempted to closely mimic the settings representative of real survey data so that our conclusions are informative for practitioners who deal with similar situations. Additional evaluation studies based on different data are desired to shed more insights on the operating characteristics and comparative performances of different missing data imputation methods. Data, code, and supplementary material for the paper are available at: https://github.com/zhenhua-wang/MissingData_DL.

Acknowledgements

Poulos and Li’s research is supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org], https://www.tensorflow.org/.

Akande, O., Li, F., and Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2), 162-170.

Arnold, B.C., and Press, S.J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84, 152-156.

Barnard, J., and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1), 17-36.

Berthelot, D., Schumm, T. and Metz, L. (2017). *BEGAN: Boundary Equilibrium Generative Adversarial Networks*. CoRR, abs/1703.10717. http://arxiv.org/abs/1703.10717.

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

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworh, Inc.

Burgette, L., and Reiter, J.P. (2010). Multiple imputation via sequential regression trees. *American Journal of Epidemiology*, 172, 1070-1076.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L. and Li, Y. (2018). BRITS: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems*, 6775-6785.

Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), 1-12.

Chen, S., and Haziza, D. (2019). Recent developments in dealing with item nonresponse in surveys: A critical review. *International Statistical Review*, 87, S192-S218.

De Leeuw, E.D., Hox, J. and Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics,* Stockholm, 19(2), 153-176.

Doove, L., Van Buuren, S. and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92-104.

Dua, D., and Graff, C. (2017). *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml.

Fortuin, V., Baranchuk, D., Rätsch, G. and Mandt, S. (2020). GP-VAE: Deep probabilistic time series imputation. *International Conference on Artificial Intelligence and Statistics*, 1651-1661.

Gelman, A., and Speed, T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55, 185-188.

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Artificial Intelligence and Statistics*, 9, 249-256.

Gondara, L., and Wang, K. (2018). MIDA: Multiple imputation using denoising autoencoders. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 260-272.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672-2680.

Ham, H., Jun, T.J. and Kim, D. (2020). *Unbalanced Gans: Pre-Training the Generator of Generative Adversarial Network Using Variational Autoencoder*. arXiv preprint arXiv:2002.02112.

Harel, O., and Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16), 3057-3077.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd Ed.), Springer.

Haziza, D., and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science*, 3(2), 583-623.

Ho, T.K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, 278-282.

Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1-47.

Horton, N.J., Lipsitz, S.R. and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4), 229-232.

Huque, M.H., Carlin, J.B., Simpson, J.A. and Lee, K.J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(1), 1-16.

Kingma, D., and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980.

Li, F., Yu, Y. and Rubin, D. (2012). *Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines*. Technical report, Duke University Department of Statistical Science Discussion Paper, 11-24.

Li, F., Baccini, M., Mealli, F., Zell, E.R., Frangakis, C.E. and Rubin, D.B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *Journal of Computational and Graphical Statistics*, 23(3), 877-892.

Lipton, Z.C., Kale, D.C. and Wetzel, R. (2016). Modeling missing data in clinical time series with RNNs. *Machine Learning for Healthcare*, 56.

Little, R.J., and Rubin, D.B. (2014). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons, Inc.

Little, R.J., and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Lu, H.-M., Perrone, G. and Unpingco, J. (2020). *Multiple Imputation with Denoising Autoencoder Using Metamorphic Truth and Imputation Feedback*. arXiv preprint arXiv:2002.08338.

Maas, A.L., Hannun, A.Y. and Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML*, (1), 3.

Manrique-Vallier, D., and Reiter, J. (2014). Bayesian estimation of discrete multivariate truncated latent structure models. *Journal of Computational and Graphical Statistics*, 23, 1061-1079.

Monti, F., Bronstein, M. and Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. *Advances in Neural Information Processing Systems*, 3697-3707.

Murray, J.S., and Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466-1479.

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

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480), 1462-1471.

Royston, P., and White, I.R. (2011). Multiple imputation by chained equations (mice): Implementation in Stata. *Journal of Statistical Software*, 45(4), 1-20.

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

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, 179, 764-74.

Stekhoven, D.J., and Bühlmann, P. (2012). Missforest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118.

Su, Y.-S., Gelman, A.E., Hill, J. and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45.

Tang, L., Song, J., Belin, T.R. and Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24(14), 2111-2128.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press LLC.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 1-67.

Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096-1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

White, I.R., Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399.

Wongkamthong, C., and Akande, O. (2021). A comparative study of imputation methods for multivariate ordinal data. *Journal of Survey Statistics and Methodology*, in press.

Yoon, J., Jordon, J. and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 5689-5698.

Yoon, J., Zame, W.R. and van der Schaar, M. (2018). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5), 1477-1490.

Yuan, Y. (2011). Multiple imputation using SAS software. *Journal of Statistical Software*, 45(6), 1-25.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 401-437
Statistics Canada, Catalogue No. 12-001-X

Multilevel time series modelling of antenatal care coverage in Bangladesh at disaggregated administrative levels

Sumonkanti Das, Jan van den Brakel, Harm Jan Boonstra and Stephen Haslett[1]

Abstract

Multilevel time series (MTS) models are applied to estimate trends in time series of antenatal care coverage at several administrative levels in Bangladesh, based on repeated editions of the Bangladesh Demographic and Health Survey (BDHS) within the period 1994-2014. MTS models are expressed in an hierarchical Bayesian framework and fitted using Markov Chain Monte Carlo simulations. The models account for varying time lags of three or four years between the editions of the BDHS and provide predictions for the intervening years as well. It is proposed to apply cross-sectional Fay-Herriot models to the survey years separately at district level, which is the most detailed regional level. Time series of these small domain predictions at the district level and their variance-covariance matrices are used as input series for the MTS models. Spatial correlations among districts, random intercept and slope at the district level, and different trend models at district level and higher regional levels are examined in the MTS models to borrow strength over time and space. Trend estimates at district level are obtained directly from the model outputs, while trend estimates at higher regional and national levels are obtained by aggregation of the district level predictions, resulting in a numerically consistent set of trend estimates.

Key Words: Cross-sectional Fay-Herriot model; Hierarchical Bayesian approach; MCMC simulation; Small area estimation; Demographic and Health Surveys.

1. Introduction

Demographic and Health Surveys have been widely used in over 90 countries for estimating national and sub-national level indicators on fertility, family planning, child mortality, child health, maternal health, and nutrition of children and women (DHS, 2021). In the sampling design of the Bangladesh Demographic and Health Survey (BDHS), administrative units lower than the sub-national level (7 divisions), such as 64 districts and more than 450 sub-districts (second and third administrative hierarchies respectively), are not accounted for. Consequently sample sizes are too small to estimate any indicator under division level with standard design-based estimators. Over the time period 1994-2014 seven surveys have been conducted, providing time series of direct estimates at the national level and division level on aforementioned indicators to monitor progress in declining maternal and neonatal mortality in Bangladesh. However, for optimal allocation of resources and policy making, reliable statistical information at the more detailed regional level of districts is required. For these regions, small area estimation models are developed in this paper. Small area estimation refers to a class of model based estimation procedures that improve upon the accuracy of direct domain estimates by increasing the effective sample size in each separate domain with sample information observed in other domains or preceding reference period. This is often referred to as borrowing strength over space or time, respectively (Rao and Molina, 2015).

[1] Sumonkanti Das, School of Demography, Australian National University, Department of Quantitative Economics, Maastricht University; Jan van den Brakel, Department of Quantitative Economics, Maastricht University, Methodology Department, Statistics Netherlands. E-mail: ja.vandenbrakel@cbs.nl; Harm Jan Boonstra, Methodology Department, Statistics Netherlands; Stephen Haslett, School of Fundamental Sciences & Centre for Public Health Research, Massey University, Research School of Finance, Actuarial Studies and Statistics, Australian National University.

The BDHS is conducted repeatedly with varying time lags of 3 or 4 years between two consecutive surveys. Seven editions for the period of 1994 until 2014 are included in this study. In this paper multivariate multilevel time series (MTS) models are developed to produce reliable trend estimates of antenatal care (ANC) coverage at district level as well as division and national levels. These models are developed in an hierarchical Bayesian framework and fitted using Markov Chain Monte Carlo simulations. The advantage of a multivariate time series approach is that it takes advantage of all available information by modelling cross-sectional and temporal correlations among districts and reference periods. The models are defined at an annual frequency and therefore properly account for the varying time lags of 3 or 4 years between the subsequent survey occasions. On top of that the MTS models provide predictions for the years without survey data.

Two related response variables are considered in this paper: whether no or at least four antenatal consults have been received, abbreviated as ANC0 and ANC4. Direct estimates along with variance estimates are calculated from the cross-sectional data of the seven BDHS surveys at the most detailed regional level of districts and are used as input for the MTS model. A drawback of this approach is that additional auxiliary information, available from two censuses, cannot be included in the MTS models. The censuses are conducted with intervals of ten years. This implies that the same values of auxiliary information, available from a particular census are used in two or even three subsequent editions of the BDHS conducted after this census but preceding the next census. This creates shocks in the MTS predictions during periods in which information from a new census becomes available. This problem is circumvented by developing the following two-step approach. As a first step, cross-sectional Fay-Herriot (FH) models (Fay and Herriot, 1979) are applied to each survey occasion using the direct estimates at the district level and their smoothed standard errors. The census auxiliary information is used to improve these cross-sectional FH models. In a second step, these cross-sectional FH estimates are used as input in the MTS model. Note that the cross-sectional FH predictions for a particular survey year are correlated. The MTS models account for this correlation by using the full variance-covariance matrices of the cross-sectional FH predictions as input for the MTS model. The advantage of this two-step approach is that it removes large sampling errors from the direct estimates and stabilizes the input series for the MTS models. This relies, however, on the assumption that the input series for the MTS models are not biased due to miss-specification of the cross-sectional FH models. To avoid this, a careful model selection and evaluation process for the cross-sectional FH models in the first step is required.

The MTS models borrow strength over time and space in several ways. Cross-sectional relations are modelled using fixed effects as well as district-level random intercepts and slopes, either independent or correlated. Spatial correlations among districts are also considered. Smooth trends and local level trends at district, division and national level are used to model temporal and cross-sectional correlations. Instead of defining a full correlation matrix between the trend disturbance terms at the district level, trends are defined at the division level, which are shared by all underlying districts. Deviations from this overall trend are modelled with trends at the district level. This is a parsimonious way of modelling cross-sectional relations between districts (Boonstra and van den Brakel, 2019). Trend estimates at the district level are obtained directly from the model outputs, while trends at division and national levels are obtained by aggregation of the district level predictions. The advantage of producing estimates for higher aggregation levels by aggregating predictions from the most detailed regional level is that all publication tables are numerically consistent by definition. Estimates for districts for the non-surveyed years and districts not covered in the surveyed years are also predicted based on the estimated time series models.

The MTS models developed in this paper are extensions of the FH model. Rao and Yu (1994) extended FH model to borrow strength over time by assuming area-specific random effects follow a first-order autoregressive AR(1) model over time independently across areas. Datta, Lahiri, Maiti and Lu (1999), and You and Rao (2000) generalized a time-series extension of the FH model following Rao and Yu (1994) in hierarchical Bayes framework by considering area-specific error terms follow first-order random walk model over time instead of AR(1) process. Datta, Lahiri and Maiti (2002) also used a random walk model for the time component to estimate median income of four-person families by state using time series and cross-sectional data using empirical Bayes estimation method. Marhuenda, Molina and Morales (2013) extended Rao and Yu (1994) model to a spatio-temporal version of FH model by incorporating additional assumption that area-specific random effects follow a first order simultaneously autoregressive process (Pratesi and Salvati, 2008) to include spatial correlation among data from neighboring areas. These extensions are very specific to only area-level random effects component. In the spatio-temporal FH model considered in this study, random effects can be specified at various disaggregation levels beside the target detailed level domains to incorporate spatial, temporal and spatio-temporal correlations among the data. In this regard, the considered hierarchical Bayesian model is more flexible than the other extension of the FH model. Other relevant accounts of multilevel time-series models and state space models extending the FH model to borrow strength over both time and space, include You, Rao and Gambino (2003); You (2008); Pfeffermann and Burck (1990); Pfeffermann and Tiller (2006); Bollineni-Balabay, van den Brakel, Palm and Boonstra (2017); Boonstra and van den Brakel (2022, 2019) and Boonstra, van den Brakel and Das (2021).

The remainder of this article is organized as follows. In Section 2 the need for reliable low regional statistical information to evaluate Sustainable Development Goals related to maternal and neonatal mortality in Bangladesh is described. Section 3 briefly describes the data sources and the computation of direct estimates and variance estimates from the BDHS survey data, along with transformations of direct estimates and the Generalized Variance Function (GVF) approach for smoothing the variance estimates, which both improve model fitting. Section 4 describes the hierarchical Bayesian time series multilevel modelling framework. The models selected for ANC0 and ANC4 are presented in Section 5, along with a brief discussion of the model building process. Section 6 provides a discussion on the trend estimates based on the developed models, and some model evaluation results are illustrated in Section 7. The paper concludes with a discussion in Section 8.

2. Need for reliable regional statistics on maternal and neonatal mortality in Bangladesh

Bangladesh has made remarkable progress in reducing the maternal mortality ratio (MMR) and neonatal mortality rate (NMR) following the target of Millennium Development Goals 4 and 5. However, both the indicators MMR (170 per 100,000 live births (WHO, UNICEF and Others, 2014)) and NMR (28 per 1,000 live births (NIPORT, 2015a)) are still reasonably high compared to the Sustainable Development Goals (SDGs) of reducing MMR to 70 per 100,000 live births and NMR to 12 deaths per 1,000 live births in Bangladesh (BBS, 2020). Poor utilization of maternal health services such as antenatal care (ANC), skilled birth attendance (SBA) at delivery, and postnatal care (PNC) (NIPORT, 2016), is considered as one of the major reasons for these high mortality rates. Receiving sufficient ANC during pregnancy is important since it also increases usage of SBA and PNC (Mrisho, Obrist, Schellenberg, Haws, Mushi, Mshinda, Tanner and Schellenberg, 2009).

The most recent household survey indicates that the majority of pregnant women (75%) in Bangladesh receive ANC from medically trained providers. However, the proportion of women that receive WHO-recommended 4^+ ANC is much less at 37% (BBS and UNICEF, 2019). These data suggest that Bangladesh lags behind in reaching the national target of 50% 4^+ ANC utilization by the year 2016. To address this gap and to meet the target of the third SDG 3 of increasing 4+ ANC coverage to 98% by 2030 (NIPORT, 2015b), the country needs a comprehensive strategy and specific milestones. National level trends of ANC coverage indicate that the proportion of women having no ANC care (ANC0) improved to only 17.2% in 2019 from 85% in 1994, while the proportion of women who obtained at least four ANC (ANC4) increased to 37% in 2019 from 6% in 1994. The improvement of the indicators over this period varies by division. The most marked improvement is observed for the *Khulna* division where ANC0 and ANC4 shifted from about 70% and 5% to about 12% and 40%, respectively. The poorest development has been observed in *Sylhet* division.

The facilities for ANC services vary considerably within Bangladesh. There are community clinics and family welfare centers at the union level (also non-government organisation clinics), upazila health complexes at sub-district level and district and tertiary hospitals at district level. Moreover, the access to private doctors varies according to the level of urbanization as well as the distance between the district/sub-district and the corresponding Metropolitan cities, particularly the capital city *Dhaka*. This inequality in the access to ANC is also explicitly visible at the division level. At disaggregated administrative levels such as district and sub-district, it can be expected that inequalities are even larger. There are, however, no studies that confirm this hypothesis, mainly because sufficient detailed survey data at those levels are not available. Recent evidence from disaggregated level studies on poverty, child nutrition and morbidity indicate high levels of inequality at both district and sub-district levels (Haslett and Jones, 2004; Haslett, Jones and Isidro, 2014; Das, Kumar and Kawsar, 2020; Hossain, Das and Chandra, 2020).

3. Data sources and input estimates

3.1 Data sources

Since 1993-94 the BDHS has been conducted under the authority of the National Institute of Population Research and Training (NIPORT) of the Ministry of Health and Family Welfare (MOHFW) to evaluate existing health and social programs and to design new strategies for improving the health status of the country's women and children. Until 2018, eight BDHS surveys have been conducted: in 1993-94, 1996-97, 2000, 2004, 2007, 2011, 2014 and 2017-18. In this study, the survey data over the period 1994-2014 have been used since the district level location of the surveyed clusters is not disclosed in the most recent BDHS 2017-18. Over the period of 1994-2014, three Population and Housing Censuses have been conducted, in 1991, 2001 and 2011. Full census data are not available, but only 10% of Census 1991 data, 10% of Census 2001 data and 5% of Census 2011 data are publicly available from IPUMS-International (https://international.ipums.org). A number of district-level contextual variables have been generated and used in the development of cross-sectional FH models to produce input estimates for the MTS models.

3.2 Direct estimates

The variables analysed in this paper are ANC0 and ANC4. Bangladesh is divided into 7 sub-national regions, called divisions. These divisions are further divided into 64 districts, which is the most detailed regional level considered in this study. As a first step, estimates and variance estimates of the two target variables at the district level are obtained from each survey year's unit-level data using the standard design-based direct survey estimator (hereafter denoted by DIR), where the survey weights are used to account for the sampling design and for non-response.

In this study, reproductive age ever-married women who have given birth within the last three years before a survey year are considered as the target population. Since in the census population such pregnancy related information is not available, area-specific population size is estimated by the number of reproductive age ever-married women available in the three Censuses. This means that even though the area-specific sample sizes are based on a census, there is some uncertainty about them, which is ignored in the SAE models. See Das, van den Brakel, Boonstra and Haslett (2021) for more details about division and district specific population sizes.

The BDHS uses a two-stage stratified sample of households. The strata are formed from divisions and sub-divisions according to their urban-rural characterization. The primary sampling units (PSUs) are the enumeration areas of the Population and Housing Census created to have an average of about 120 households (slightly vary over census). In the first stage, PSUs are selected with probabilities proportional to PSU size, i.e., the number of households. In the second stage, a complete household listing is carried out in all selected PSUs and then about 30 households are selected from each PSU using systematic sampling. The response rates among eligible women have been over 95% in all BDHS years. Though the sample size of the ever-married women is greater than 10,000 in all the surveys, in this study only the ever-married women who had a child birth in the three years preceding the survey year are considered, and therefore sample sizes are smaller. At the district level, mean sample sizes vary between 60 and 114, with some districts having less than 10 or even no observed women.

Sampling weights are calculated based on selection probabilities. These weights are then adjusted for household and individual non-response. The direct estimate for the population proportion in a certain domain i for survey year t is computed as the sample mean

$$\hat{Y}_{it} = \frac{\sum_{j \in s_{it}} w_{ijt} y_{ijt}}{\sum_{j \in s_{it}} w_{ijt}}, \tag{3.1}$$

where y is the response variable of interest, s_{it} is the set of ever-married women in domain i for which y is observed in year $t,$ and w_{ijt} is the survey weight for person j living in area i in year $t.$ Note that the weights w_{ijt} are scaled such that the sum over the weights in the sample is equal to the net sample size. The corresponding variance estimates are approximated as

$$\operatorname{var}(\hat{Y}_{it}) = \frac{1}{n_{it}\left(n_{it}-1\right)} \sum_{j \in s_{it}} w_{ijt} \left(y_{ijt} - \hat{Y}_{it}\right)^2, \tag{3.2}$$

where n_{it} is the number of ever-married women observed in domain i at the survey year $t.$ Initially, the variance was approximated by calculating the variance among the estimated PSU totals as if they were selected by using stratified sampling with replacement, known as the ultimate sampling unit variance approximation. This resulted in zero variance estimates for a few domains. Variance approximation (3.2) avoids these zero variance estimates, and otherwise results in variance estimates comparable with the initial approximation where PSUs were assumed to be selected with replacement. In the first MTS model, denoted by MTS-I, these direct estimates are used as the input series.

3.3 Cross-sectional Fay-Herriot estimates

An issue with the MTS-I model is the use of census data as auxiliary variables in the MTS model. Because the time gap between two subsequent censuses is 10 years whereas the BDHS is conducted every 3 or 4 years, the census covariates remain the same until the new census data are available. Including these census data as covariates in the MTS-I models will bias estimates of trends and period-to-period changes. One way to take advantage of the census information is to model the direct estimates at the district level in separate cross-sectional FH models using relevant contextual variables extracted from the census data. It is also expected that the use of on-time available census auxiliary variables in repetitive cross-sectional FH models may affect regression coefficients and the accuracy of model predictions of the dependent variable, but not the predictions of the dependent variable itself. Compared to the direct estimates used in MTS-I, these cross-sectional FH models also provide better estimates by already borrowing some strength over districts.

The cross-sectional FH estimates and their standard errors are used as input for a second model, denoted by MTS-II. The cross-sectional FH estimates are correlated due to their common fixed effect components, which is ignored in MTS-II. Therefore a third MTS model, denoted by MTS-III, is developed using cross-sectional FH estimates and their full covariance matrix as input.

The fixed and random effect components for the survey-specific cross-sectional FH models are shown in Appendix Tables A.2 and A.3. For all the models, random effects are assumed to follow a normal distribution. Non-normal models have been considered for the random effects (Laplace and horseshoe) and the sampling error (t-distribution) as alternatives for the normal distribution. This, however, did not improve the model fit.

3.4 Generalized variance functions

In the FH and MTS models, the variance estimates of the direct estimates are largely treated as fixed given quantities. Since these variance estimates can be very noisy, they are smoothed using a GVF before using them in the FH and MTS models. It is understood that a district without sample information is considered as missing and is therefore not considered in the model development approach. The cross-sectional FH model can produce estimates and standard errors for these out-of-sample domains. These synthetic estimates are, however, not used in the development of the MTS-II and MTS-III models to allow for a better comparison with the MTS-I model.

The GVFs are regression models that relate the variance estimates to predictors such as sample size, survey design variables, and point estimates (Wolter (2007), Chapter 7). For both ANC0 and ANC4, the following GVF is used:

$$\log \mathrm{se}(\hat{Y}_{it}) = \alpha + \beta \log \tilde{Y}_{it} + \gamma \log(m_{it} + 1) + \delta \,\mathrm{Division} + \in_{it}, \tag{3.3}$$

where $\mathrm{se}(\hat{Y}_{it})$ is the standard error of $\hat{Y}_{it}$ in (3.1), m_{it} the number of sampling units contributing to district i in year t and Division is a categorical variable with 7 levels. Since we cannot trust the direct estimates for very small m_{it}, the $\tilde{Y}_{it}$ on the right hand side of (3.3) are simple smoothed estimates

$$\begin{aligned} \tilde{Y}_{it} &= \lambda_{it}\hat{Y}_{it} + (1-\lambda_{it})\,\bar{Y}_{d[i]t}, \\ \lambda_{it} &= \frac{m_{it}}{m_{it}+1}, \end{aligned} \tag{3.4}$$

where $\bar{Y}_{d[i]t}$ denotes the mean for division $d\,(d = 1$ to 7) to which district i belongs, in year t. As mentioned by a referee, a composite regression estimator can be used as an alternative for (3.4).

The regression errors $\in_{it}$, are assumed to be independent and normally distributed with a common variance parameter σ^2. The GVFs are fitted only to districts with non-zero standard errors of the direct estimates. The predicted (smoothed) standard errors based on the fitted models are

$$\mathrm{se}_{\mathrm{pred}}(\hat{Y}_{it}) = \exp\left(\hat{\alpha} + \hat{\beta}\log \tilde{Y}_{it} + \hat{\gamma}\log(m_{it}+1) + \hat{\delta}\,\mathrm{Division} + \hat{\sigma}^2/2\right), \tag{3.5}$$

where $\hat{\sigma}$ is 0.03 for ANC0 and 0.003 for ANC4, respectively. The R-squared values for both models are quite high 0.79 for ANC0 and 0.99 for ANC4. Note that the exponential back-transformation in (3.5) includes a bias correction, which in this case has only a small effect. This approach is used to get smoothed standard errors for the cross-sectional FH models and MTS-I model.

3.5 Transformations of input series

Square root, log and log-ratio transformation are considered as a variance stabilizing transformation, see Sakia (1992). The square root transformation is applied to ANC4 data (the MTS models and the cross-sectional FH models) since this transformation reduces the correlation between point estimates and their standard errors of the input series, reduces heterogeneity, improves the convergence of the MCMC simulation, and reduces the skewness of proportion data if they take values close to the lower boundary of zero. For ANC0, the square root transformation is only used for the year specific cross-sectional FH models in 2011 and 2014 only. In the other years, no transformation is applied. In all three MTS models, no transformation is applied for ANC0 since the square root transformation for the input series increases the dependency between direct estimates and standard errors.

Let $\hat{\mathcal{Y}}_{it} = \sqrt{(\hat{Y}_{it} + \varepsilon)}$ denote the square root transformed direct estimates, where ε is a small number (0.005), necessary because for some districts direct estimates equal zero. Using a first order Taylor approximation it can be shown that $\mathrm{se}(\hat{\mathcal{Y}}_{it}) \approx \mathrm{se}(\hat{Y}_{it}) \Big/ \left(2\sqrt{\hat{Y}_{it} + \varepsilon}\right)$.

If the GVF (3.3) is applied to the standard errors of the untransformed direct estimates, then the standard errors for domains with a very small number of sampling units can become unreasonably large due to the linearisation approximation. This issue is avoided by applying the GVF to the standard errors of the transformed estimates, i.e., $\mathrm{se}(\hat{\mathcal{Y}}_{it})$.

4. Time series multilevel modelling

In this study, direct estimates and their standard errors are available for the survey years 1994, 1997, 2000, 2004, 2007, 2011 and 2014. To account for the varying time-lags of 3 or 4 years between the subsequent survey years, the MTS models are defined at an annual frequency, (i.e., values refer to a reference period of one year) at the most detailed regional level of the 64 districts. With a time span of 21 years, there are 1,344 domain-year combinations. With seven available survey years, the model is fitted to the 448 domain-year observations. The years between two subsequent surveys are defined as missing in the model. In this way the period-to-period evolution of the trend is specified correctly and the model provides predictions for the missing domain-year combinations.

For convenience let us now denote by $\hat{Y}_{it}$ the input series for the time series models for either ANC0 or ANC4 in year t and domain i. This can be the untransformed direct estimates, the square root transformed direct estimates or the model predictions obtained with the cross-sectional FH models. Here domain index i runs from 1 to $M_d = 64$ and time index t from 1 to $T = 21.$ We further combine these estimates into a vector $\hat{Y} = (\hat{Y}_{11}, \ldots \hat{Y}_{M_d 1}, \ldots \hat{Y}_{1T}, \ldots \hat{Y}_{M_d T})',$ a vector of dimension $M = M_d T.$

4.1 Model structure

The multilevel models considered take the general linear additive form

$$\hat{Y} \ = \ X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} \ + e, \tag{4.1}$$

where X is a $M \times p$ design matrix for a p-vector of fixed effects $\beta,$ and the $Z^{(\alpha)}$ are $M \times q^{(\alpha)}$ design matrices for $q^{(\alpha)}$-dimensional random effect vectors $v^{(\alpha)}.$ Here the sum over α runs over several possible random effect terms at different levels, such as local level and smooth trends at district and division levels, white noise at the most detailed level of the M domains, etc. This is explained in more detail below. In formula (4.1) $e = (e_{11}, \ldots, e_{M_d 1}, \ldots e_{M_d T})'$ denotes, depending on the input series, the sampling errors of the direct estimates or the prediction errors of the cross-sectional FH model. The errors are taken to be normally distributed as $e \sim N(0, \Sigma)$ where $\Sigma = \oplus_{t=1}^{T} \Sigma_t.$ If the input series are the untransformed direct estimates, then Σ_t is the covariance matrix for the untransformed direct estimates observed in year $t.$ If the input series are transformed, then Σ_t is the covariance matrix for the transformed direct estimates, as described in Subsection 3.5. If the input series are the predictions based on the cross-sectional FH models, then Σ_t contains the estimated mean squared errors of the FH predictions. Under MTS-II, Σ_t is diagonal and ignores the correlations between the domain predictions. Under MTS-III, Σ_t is a full covariance matrix that also accommodates the correlations between domain predictions.

Based on the distribution of the sampling errors e in (4.1), the likelihood function conditional on fixed and random effects parameters can be defined as

$$p\left(\hat{Y} \middle| \eta, \Sigma\right) \ = \ N\left(\hat{Y} \middle| \eta, \Sigma\right), \tag{4.2}$$

where $\eta = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)}$ is the linear predictor. For the errors e a Student-t distribution instead of the normal distribution can be considered to give smaller weight to more outlying observations, following West (1984).

The fixed effect part of η can contain components like an intercept, a linear trend, main effects for division and district and possibly the second-order interactions for linear trends and division or district. The vector β of fixed effects is assigned a normal prior $p(\beta) = N(0, 100 I_p),$ with I_x the identity matrix of dimension $x \times x.$ This is only very weakly informative as a standard error of 10 is very large relative to the scales of the (transformed) direct estimates and the covariates used.

The second term on the right hand side of (4.1) consists of a sum of contributions to the linear predictor by random effects or varying coefficient terms. The random effect vectors $v^{(\alpha)}$ for different α are assumed to be independent, but the components within a vector $v^{(\alpha)}$ are possibly correlated to accommodate temporal or cross-sectional correlation. To describe the general model for each vector $v^{(\alpha)}$ of random effects, we suppress superscript α in what follows for notational convenience.

Each random effects vector v is assumed to be distributed as

$$v \sim N(0, A \otimes V), \tag{4.3}$$

where V and A are $d \times d$ and $l \times l$ covariance matrices, respectively, and $A \otimes V$ denotes the Kronecker product of A with V. The total length of v is $q = dl$, and these coefficients may be thought of as corresponding to d effects allowed to vary over l levels of a factor variable. If, e.g., V corresponds to division, then V defines $d = 7$ different random effects that correspond to the 7 categories of division. If subsequently A corresponds to time, then $l = 21$ years. In that case each of the 7 effects can vary over its 21 levels (years in this case). Each random effect generated for a division × year combination is shared by all districts belonging to that division in that particular year.

The covariance matrix A describes the covariance structure among the levels of the factor variable, and is assumed to be known. Instead of covariance matrices, precision matrices $Q_A = A^{-1}$ are actually used, because of computational efficiency (Rue and Held, 2005). The covariance matrix V for the d varying effects can be parameterized in one of three different ways: (i) a full parameterized covariance matrix, (ii) a diagonal matrix with unequal diagonal elements, and (iii) a diagonal matrix with equal diagonal elements. The scaled-inverse Wishart prior is used as proposed in O'Malley and Zaslavsky (2008) and recommended by Gelman and Hill (2007) when a full covariance matrix is assumed, while half-Cauchy priors are used for the standard deviations when the covariance matrix is assumed diagonal with equal or unequal elements. In case of diagonal variances, half-Cauchy priors are better default priors than the more common inverse gamma priors (Gelman, 2006).

The following random effect structures are considered in the model selection procedure:

1. Random intercepts for the M_d domains. In this case $A = I_{M_d}$ and V is a scalar variance parameter. This implies $v_{it} = v_i, \forall t$ and $v_i \sim N(0, \sigma_I^2)$.
2. First or second order random walks at different aggregation levels. A first order random walk or local level trend at district level is defined as $v_{it} = L_{it}$ with $L_{it} = L_{i,t-1} + \eta_{it}$ and $\eta_{it} \sim N(0, \sigma_{R1,i}^2)$. A second order random walk or smooth trend model at district level is defined as $v_{it} = L_{it}$ with $L_{it} = L_{i,t-1} + R_{i,t-1}$, $R_{it} = R_{i,t-1} + \eta_{it}$ and $\eta_{it} \sim N(0, \sigma_{R2,i}^2)$. Both kind of trends can be defined similarly at the division or national level. See Rue and Held (2005) for the specification of the precision matrix Q_A for first and second order random walks. A full covariance matrix for the trend innovations can be considered to allow for cross-sectional besides temporal correlations, or a diagonal matrix with different or equal variance parameters to allow for temporal correlations only. In the case of equal variances, $\sigma_{R1,i}^2 = \sigma_{R1}^2$ and $\sigma_{R2,i}^2 = \sigma_{R2}^2, \forall i$. First and second order random walk components at district level are denoted below by RW1_District and RW2_District respectively. At division level they are denoted by RW1_Division and RW2_Division.

3. The first order random walks as used in our models cannot capture an overall level as the corresponding random effects are constrained to sum to zero over time. Similarly, the second order random walks cannot capture both level and linear trend. This means that level and linear trend must be accommodated by other model terms, as either fixed or random effects. District-level intercepts have already been discussed under item 1. To also include linear trends by district, this component can be extended to random intercepts and slopes linear in time. In that case V can be either a 2×2 general covariance matrix

$$V = \begin{pmatrix} \sigma_I^2 & \rho_{IS}\,\sigma_I\,\sigma_S \\ \rho_{IS}\,\sigma_I\,\sigma_S & \sigma_S^2 \end{pmatrix},$$

   accounting for correlations between intercepts and slopes, or a diagonal matrix with diagonal elements σ_I^2 and σ_S^2 the variances of the radom intercept and slopes respectively. This model component is referred to as RIS_District below.

4. Spatial random effects: random intercepts varying over the spatial location of districts following an intrinsic conditional autoregressive (ICAR) model (Besag and Kooperberg, 1995), defined as $v_i \,|\, v_{-i} \sim N\left((\Sigma_{i' \in nb(i)} v_{i'}) \,/\, a_i \,, \sigma_{Sp}^2 \,/\, a_i\right)$ for each spatial effect conditional on the others. Here $nb(i)$ is the set of domains neighbouring domain i and a_i the number of domains neighbouring domain i. See Rue and Held (2005) for the specification of the precision matrix Q_A. This spatial component is referred to later as Spatial_District.

5. White noise: to allow for random unexplained variation, white noise at the most detailed domain-by-year level can be included. In this case $A = I_M$ and V a scalar variance parameter. This implies $v_{it} \sim N(0, \sigma_W^2)$.

We also investigated generalisations of (4.3) to non-normal distributions of random effects by implementing Student-t, horseshoe prior (Carvalho, Polson and Scott, 2010) and Laplace (Tibshirani, 1996; Park and Casella, 2008). These alternative distributions have fatter tails allowing for occasional large effects. However, these distributions did not improve results for the considered target variables in terms of model information criteria as well as the underlying trend predictions. Therefore the normal distribution is used for all random effect components. The exact lay out of the final MTS models for ANC0 and ANC4 are specified in Subsections 5.1 and 5.2 respectively.

4.2 Model estimation

The models are fitted using Markov Chain Monte Carlo (MCMC) sampling, in particular the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). See Boonstra and van den Brakel (2022) for a specification of the full conditional distributions. The models specified in Subsection 4.1 are run in R (R Core Team, 2015) using package mcmcsae (Boonstra, 2021). The Gibbs sampler is run in parallel for three independent chains with randomly generated starting values. In the model building stage 1,000 iterations are used, in addition to a “burn-in” period of 100 iterations. This was sufficient for reasonably stable Monte Carlo estimates of the model parameters and trend predictions. For the selected model we use a longer run of 1,000 burn-in plus 5,000 iterations of which the draws of every fifth iteration are stored. This leaves 3 × 1,000 = 3,000 draws to compute estimates and standard errors. The convergence of the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. For the longer simulation of the selected model all model parameters and model predictions have potential scale reduction factors below 1.01 and sufficient effective numbers of independent draws.

Many models of the form (4.1) have been fitted to the data. For the comparison of models using the same input data we use the Widely Applicable Information Criterion or Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010, 2013) and the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin and van der Linde, 2002). We also compare the models graphically by their model fits and trend predictions at three aggregation levels.

5. Selected models and model prediction

5.1 MTS model for ANC0

No transformation for the input series of the direct estimates or the FH estimates is considered. The following fixed effect components are included in the selected models for MTS-I, MTS-II, and MTS-III:

$$1 + \text{Division} + yr.c + \text{Division} * yr.c, \tag{5.1}$$

where $yr.c$ denotes the standardized quantitative year variable, which defines a fixed effect linear trend. Similarly $\text{Division} * yr.c$ defines a fixed effect linear trend for each separate division. The random effects part of the three models is shown in Table 5.1. If multiple varying effects are modeled, then there is a choice between scalar, diagonal or full covariance matrix V in (4.3). For variation over time, second order random walks RW2_Division and RW2_District were finally selected. White noise components are considered but not included in the final model since it did not further improve the model fit.

Table 5.1
Summary of the random effect components for the selected time series multilevel model for ANC0. The second and third columns refer to the varying effects with covariance matrix V in (4.3), whereas the fourth column refers to the factor variable associated with A in (4.3). The last column contains the total number of random effects for each component

Model Component	Formula V	Variance Structure	Factor A	# of Effects
RIS_District	$1 + yr.c$	full	District	128
RW2_Division	Division	scalar	RW2(yr)	147
RW2_District	District	scalar	RW2(yr)	1,344
Spatial_District	1	scalar	Spatial(District)	64

The linear predictor of the selected model can be written, element-wise for district i and year $t,$ as

$$\eta_{it} = \beta' x_{it} + v_i + z_t v_i^{(yr)} + u_{it} + u_{j[i]t}^{(div)} + s_i, \tag{5.2}$$

where β is the vector of fixed effects corresponding to the covariates x_{it} as specified in (5.1), v_i are random intercepts varying by district, z_t denotes the $yr.c$ variable for year $t,$ and $v_i^{(yr)}$ are the corresponding random slopes varying by district. These random intercepts and slopes are jointly distributed as

$$\begin{pmatrix} v_i \\ v_i^{(yr)} \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_I^2 & \rho\sigma_I\sigma_S \\ \rho\sigma_I\sigma_S & \sigma_S^2 \end{pmatrix}\right). \tag{5.3}$$

The second-order random walk effects at district and division level are distributed as

$$\begin{aligned} u_{it} - 2u_{i(t-1)} + u_{i(t-2)} &\overset{\text{iid}}{\sim} N(0, \sigma_{R2}^2) \\ u_{j[i]t}^{(div)} - 2u_{j[i](t-1)}^{(div)} + u_{j[i](t-2)}^{(div)} &\overset{\text{iid}}{\sim} N\left(0, (\sigma_{R2}^{(div)})^2\right), \end{aligned} \tag{5.4}$$

where $j[i]$ should be read as division j to which district i belongs. Finally, the spatial effects s_i are distributed as

$$s_i \,\big|\, s_{i' \neq i} \overset{\text{ind}}{\sim} N\left(\frac{1}{a_i} \sum_{i' \in nb(i)} s_{i'}, \frac{1}{a_i} \sigma_{Sp}^2\right), \tag{5.5}$$

where a_i is the size of the set $nb(i)$ of neighbouring districts of district $i.$ Priors for the covariance matrix in (5.3) and the other variance parameters are chosen as described in Section 4.1. For identifiability of the model components, the following constraints are imposed:

$$\begin{aligned} &\sum_{t=1}^{T} u_{it} = 0 \quad \text{and} \quad \sum_{t=1}^{T} t u_{it} = 0 \quad \text{for all districts } i, \\ &\sum_{t=1}^{T} u_{j[i]t}^{(div)} = 0 \quad \text{and} \quad \sum_{t=1}^{T} t u_{j[i]t}^{(div)} = 0 \quad \text{for all divisions } j, \\ &\sum_{i=1}^{M_d} s_i = 0. \end{aligned} \tag{5.6}$$

Note that RW2 trends are specified at division and district levels, both with a scalar variance structure. A division level trend is shared by all underlying districts. Deviations of each district from this division-level trend is modeled with RW2 trends at district level. This is a parsimonious alternative to borrow strength over time and space, compared to modelling RW2 trends at the district level only with a full covariance matrix (Boonstra and van den Brakel, 2019).

5.2 MTS model for ANC4

The square-root transformation is applied to the input series of the direct and FH estimates of ANC4 for models MTS-I, MTS-II, and MTS-III. For MTS-I the GVF (3.3) is applied to the transformed standard errors to obtain the variance matrix $\Sigma,$ as explained at the end of Subsection 3.5. For the fixed effect component a factor variable called "Region" has been created based on the degree of urbanization following Rahman, Mohiuddin, Kafy, Sheel and Di (2019). The variable has four levels; 1 for three big cities *Dhaka*, *Chittagong* and *Gazipur*, 2 for other nine regional big cities (*Barisal*, *Bogra*, *Comilla*, *Khulna*, *Mymensing*, *Narayanganj*, *Rajshahi*, *Rangpur*, *Sylhet*), 3 for three hilly districts (*Bandarban*, *Khagrachhari* and *Rangamati*) and 4 for the remaining districts. This variable mainly helped to adjust the estimates for the three hilly districts which have very few (even no) information in the considered seven surveys. The final model has the following fixed effects components:

$$1 + \text{Division} + yr.c + \text{Region}. \tag{5.7}$$

The interaction between "Division" and "yr.c" (like in the ANC0 model) was found to be insignificant in the ANC4 model. The random effect components for ANC4 model shown in Table 5.2 are very similar to those used for the model of ANC0 (shown in Table 5.1). A local level trend instead of smooth trend at division level (RW1_Division in Table 5.2) has been considered since the smooth trend component (RW2_Division, as in Table 5.1) resulted in some bias in the national and divisional trends. Also, the model with RW1_Division component gives better scores for the information criteria compared to the model with RW2_Division component. White noise components are considered but not included in the final model since it did not further improve the model fit.

Table 5.2
Summary of the random effect components for the selected multilevel time series model for ANC4. The second and third columns refer to the varying effects with covariance matrix V in (4.3), whereas the fourth column refers to the factor variable associated with A in (4.3). The last column contains the total number of random effects for each term

Model Component	Formula V	Variance Structure	Factor A	# of Effects
RIS_District	$1 + yr.c$	full	District	128
RW1_Division	Division	scalar	RW1(yr)	147
RW2_District	District	scalar	RW2(yr)	1,344
Spatial_District	1	scalar	Spatial(District)	64

Alternatively, the model can be expressed as in (5.2), where now β and x_{it} correspond to the fixed effects specification (5.7). The only other difference is that the division-level trends are now modelled as a first-order random walk:

$$u_{j[i]t} - u_{j[i](t-1)} \overset{\text{iid}}{\sim} N\left(0, (\sigma_{R1}^{(div)})^2\right), \tag{5.8}$$

where for identifiability reasons the constraint $\sum_{t=1}^{T} u_{jt}^{(div)} = 0$ is imposed for all division $j.$ As in the case of ANC0, RW1 trends are specified at division and RW2 trends at the district levels, both with a scalar variance structure as a parsimonious way to borrow strength over time and space.

5.3 Trend estimation

Trend estimates are computed based on the MCMC simulation results. In a first step, for each MCMC replicate, an M -dimensional vector containing predictions at the most detailed level of all year-district combinations is computed as

$$\eta^{(r)} = X\beta^{(r)} + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha,r)}, \tag{5.9}$$

where superscript (r) indexes the retained MCMC draws. Note that $\eta^{(r)}$ also includes predictions for the years without survey observations. Since a square root transformation was applied to the ANC4 series, initially the following back-transformation for the vectors $\eta^{(r)}$ was considered following Boonstra et al. (2021):

$$\theta^{(r)} = (\eta^{(r)})^2 + \left(\mathrm{se}(\hat{\mathcal{Y}}_{it})\right)^2. \tag{5.10}$$

The second term on the right hand side is a (relatively small) bias correction using the transformed and smoothed standard errors. The bias correction stems from the fact that the design expectation of the direct estimates can be written as

$$E(\hat{Y}) = E\left((\hat{\mathcal{Y}})^2\right) = E\left((\eta + \mathcal{E})^2\right) = \eta^2 + \mathrm{var}(\mathcal{E}),$$

where $\mathcal{E}$ is the vector of sampling errors after transformation, assumed to be normally distributed with standard errors $\mathrm{se}(\hat{\mathcal{Y}}_{it})$. A difficulty with the data at hand is that the bias correction can only be applied to the survey years, since standard errors are only available for those years. Applying the bias correction only for the survey years distorts the trend estimates, as illustrated in Das, van den Brakel, Boonstra and Haslett (2021). In case of MTS-I model, the impact of this bias correction is most clear for those domains with zero direct estimates particularly for *Chittagong* hilly districts. The impact of the bias correction is less in case of MTS-II and MTS-III models since the estimated standard errors of the FH estimates are already smoothed enough and consistent. However, at national and division levels this bias correction causes some overestimation in some survey years for all the trends based on the MTS models. Therefore, the bias correction for the square root transformation is not applied in the trend estimates but only used in the calculation of cross-sectional FH estimates.

Trend estimates with their standard errors at the most detailed level of districts for all years are obtained by taking the mean and the standard deviation over the MCMC replications $\eta^{(r)}$, respectively. Trends at the divisional and national levels are obtained by aggregating each MCMC replication from the most detailed regional level of districts, using the number of ever-married women as a weighting variable. Subsequently, trend estimates and their standard errors are obtained by taking the mean and the standard deviation over these aggregated MCMC replications.

6. Results

The trends of ANC0 and ANC4 shown in the figures consist of five types of estimates with their approximate 95% confidence intervals: (i) weighted direct estimates (DIR) at the surveyed year (black error-bar line), (ii) cross-sectional FH estimates at the surveyed year (green error-bar line), (iii) estimates based on MTS-I model (red line), (iv) estimates based on MTS-II (green line) and (v) estimates based on MTS-III model (blue line).

6.1 ANC0

The national level trends of ANC0 are shown in Figure 6.1. The figure shows that the DIR and cross-sectional FH estimates are very similar at the survey years with approximately equal 95% CI. This can be expected for figures at the national level, since the gain in precision obtained with a small area prediction model with respect to a direct estimator becomes smaller as the sample size increases. During the initial period 1994-2000, the national level trend based on the MTS-I model follows the DIR and cross-sectional FH estimates, while the trends based on MTS-II and MTS-III models are slightly higher. For the period 2004-2010, the trend based on MTS-I model is slightly higher than the trends based on MTS-II and MTS-III models. The differences are, however, very small.

The trends at division level, shown in Figure 6.2, indicate that the trends under MTS-I are very similar to those based on MTS-II and MTS-III models with some small exceptions in *Dhaka*, *Khulna* and *Rajshahi* divisions. The differences in *Dhaka* and *Khulna* division may cause most of the differences in the national level trends.

The trends based on the MTS-II and MTS-III models are almost identical at national and division levels. This is supported by the estimated variance components of the division-level smooth-trend random component under the developed two models ($\hat{\sigma}_{R2}^{(div)}$: about 0.020) given in Table 6.1. However, there are more substantial differences in the trends under MTS-II and MTS-III at the district level, see Figures 6.3 and 6.4. See Das, van den Brakel, Boonstra and Haslett (2021) for plots for all districts. The trends based on the MTS-III model are smoother than those based on the MTS-II model, which is a result of the smaller values of the estimated variance component $\hat{\sigma}_{R2}$ under MTS-III (see Table 6.1).

Table 6.1
Posterior means of standard deviation parameters of random components of MTS-I, MTS-II, MTS-III models for ANC0. No superscript refers to district level, superscripts (*div*) refers to division level

Model	$\hat{\sigma}_I$ (SE)	$\hat{\sigma}_S$ (SE)	$\hat{\rho}_{IS}$ (SE)	$\hat{\sigma}_{Sp}$ (SE)	$\hat{\sigma}_{R2}^{(div)}$ (SE)	$\hat{\sigma}_{R2}$ (SE)
MTS-I	0.083 (0.013)	0.054 (0.007)	0.168 (0.171)	0.068 (0.032)	0.019 (0.003)	0.024 (0.002)
MTS-II	0.069 (0.012)	0.033 (0.004)	0.254 (0.180)	0.071 (0.028)	0.020 (0.003)	0.013 (0.002)
MTS-III	0.062 (0.013)	0.027 (0.013)	0.227 (0.201)	0.067 (0.030)	0.020 (0.003)	0.009 (0.001)

The trends at the district level have a tendency to follow the pattern of their respective division level trend shown in Figure 6.2. This is particularly the case for domains with a relatively small number of observations such as districts *Bandarban*, *Khagrachnari* and *Rangamati* in Figure 6.3 that belong to *Chittagong* division in Figure 6.2. To reduce this tendency, an MTS model was developed by removing smooth trend component RW2_Division at division level in Table 5.1. This, however, resulted in highly smooth unrealistic trends at the national and divisional levels. In a similar way, to examine the need for a spatial component, MTS models were developed with and without considering the spatial component (Spatial_District in Table 5.1). It is observed that the spatial component makes the estimates more plausible for those districts with small or zero sample sizes. See for example the trends of *Bandarban* and *Rangamati* districts of *Chittagong* division.

The MTS-I model shows upward trends for some districts during the period of 1994-2000. These developments are unplausible from a subject matter point of view and are nicely corrected by the MTS-II and MTS-III models that use the FH estimates as input series. See for example *Noakhali*, *Bandarban*, *Rangamati*, *Narayanganj*, *Rajbari*, and *Narail* districts in Figure 6.3. Some districts have volatile trends according to the DIR estimates and MTS-I model during the whole period mainly due to variation in the sample size. See for example, *Bandarban*, *Bhola*, *Khagrachhari*, *Kishoreganj* and *Rangamati* in Figure 6.3, *Chapai Nababganj*, *Feni*, *Jhalokati*, *Joypurhat*, and *Pabna* districts in Figure 6.4. From a subject matter point of view a smooth decreasing trend for ANC0 coverage is expected. In particular the turning points that are visible in several districts arround 2007 and 2011 are not expected. The trends based on the MTS-II and MTS-III models ignore most of these volatilities and show reasonable smooth trends for these districts and are therefore more realistic compared to MTS-I. Nevertheless, the fits of all three models are compatible with the observed data. MTS-II appears to be a nice compromise between models I and III.

Figure 6.1 National level trends of ANC0 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).

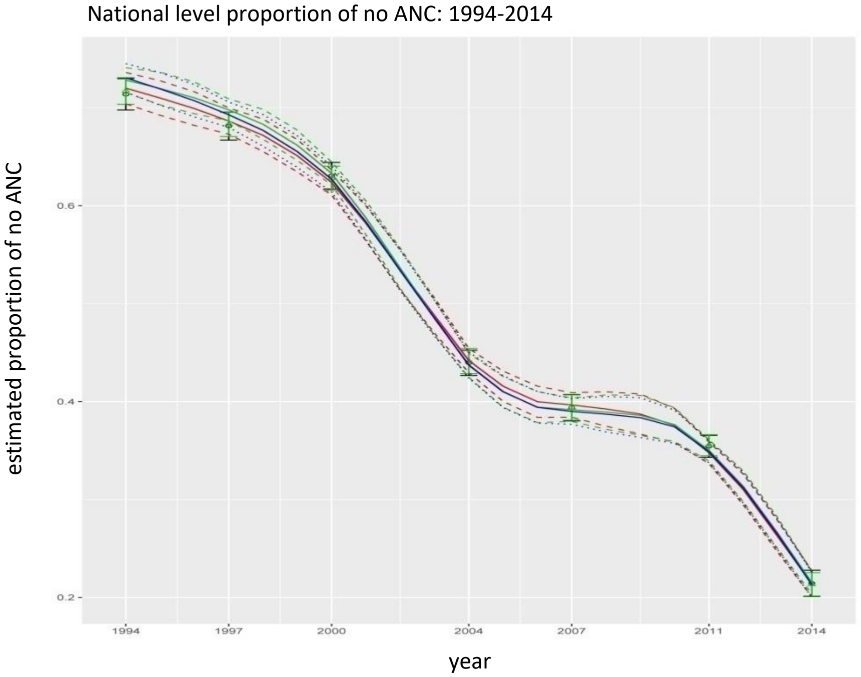


Figure 6.2 Division level trends of ANC0 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).

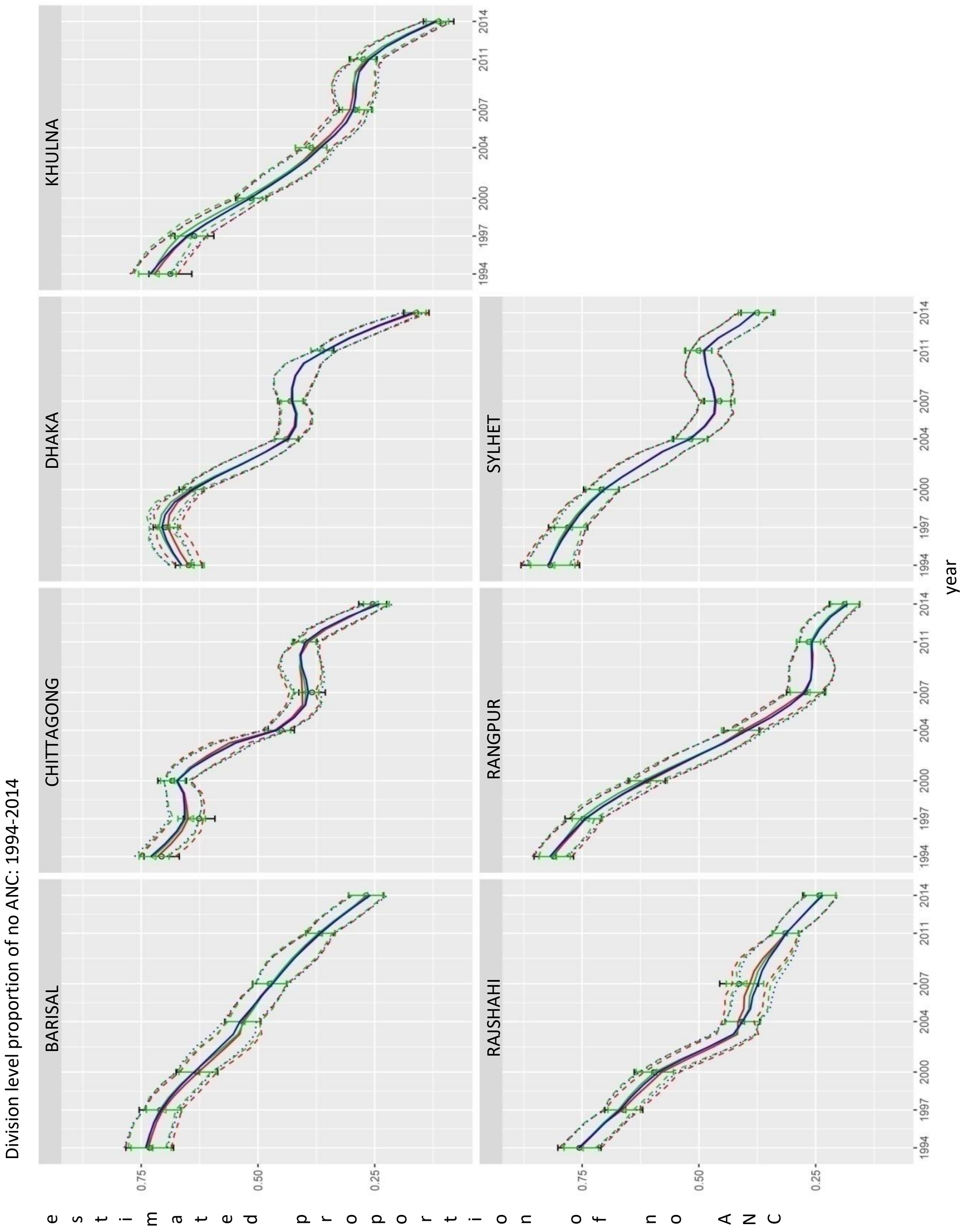


Figure 6.3 District level trends of ANC0 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).

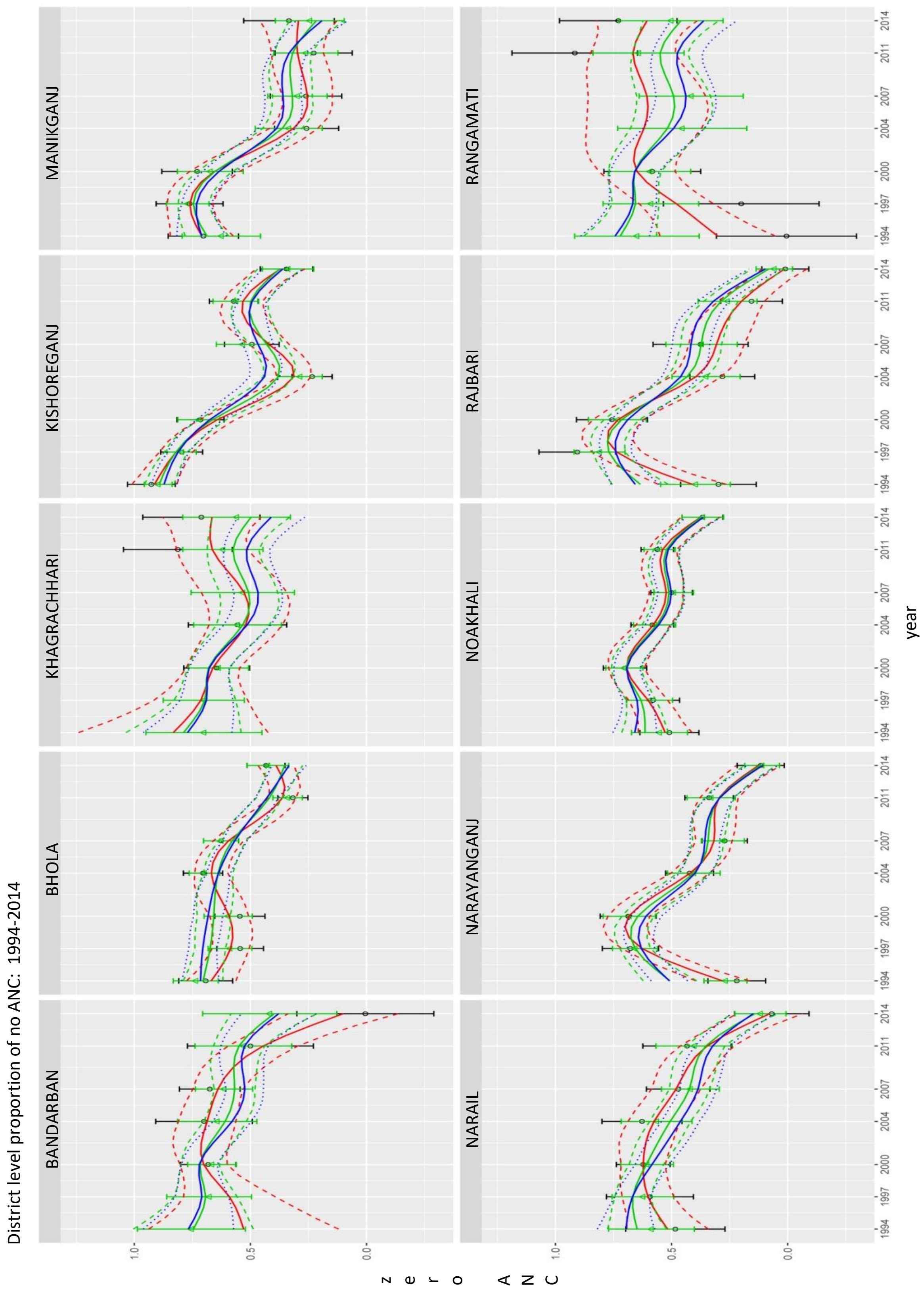
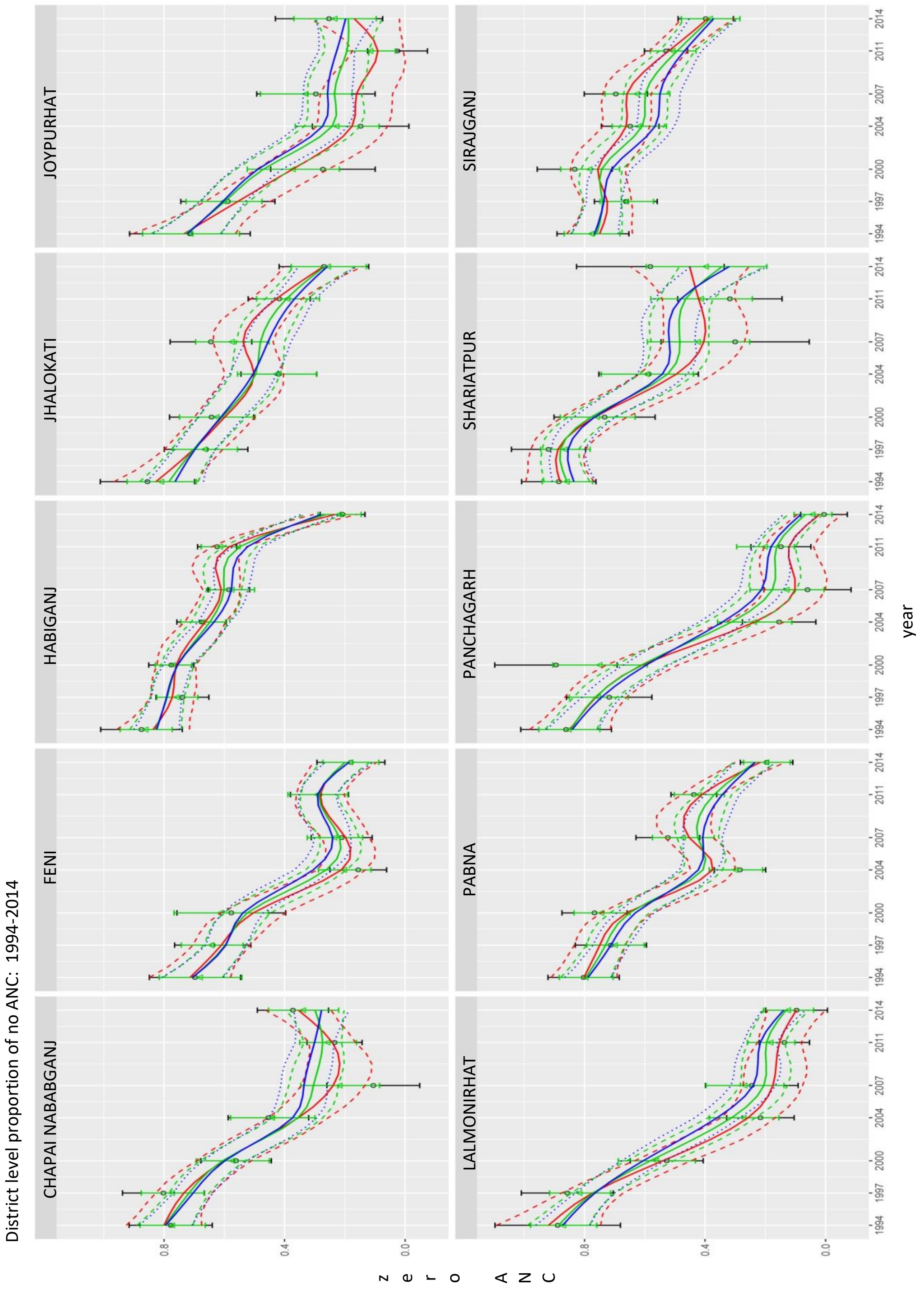


Figure 6.4 District level trends of ANC0 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).



In most cases models MTS-II and MTS-III behave similarly. However, model MTS-III, which accounts for correlation among the cross-sectional FH estimates, overestimates ANC0 for some districts (such as *Chapai Nababganj*, *Lalmonirhat* and *Shariatpur* districts in Figure 6.4) and also slightly underestimate the trend in some districts (such as *Khagrachari*, *Rangamati*, and *Shirajganj* districts in Figure 6.3) compared to the cross-sectional FH estimates. Again MTS-II seeks a compromise between smooth trends under MTS-III and more volatile trends under MTS-I in most of the districts and appears to be the preferred model for estimating trends of ANC0.

6.2 ANC4

The national level trend of ANC4 shown in Figure 6.5 shows a linear upward increase from 6% in 1994 to about 31% in 2014. Like ANC0, the DIR and cross-sectional FH estimates of ANC4 are very similar at the survey years with approximately equal 95% CI. Trends estimated from the MTS-I (red line), MTS-II (green line) and MTS-III (blue line) show very similar patterns. Compared to the DIR and cross-sectional FH estimates, the trend of MTS-I is slightly lower in 2007 and 2014. Trends under MTS-II and MTS-III in survey year 2011 are somewhat larger compared to the DIR and cross-sectional FH estimates. The trends at division level are shown in Figure 6.6. The three MTS models give very similar trend estimates. Some differences occur in *Chittagong*, *Dhaka* and *Rangpur* divisions. With MTS-I the trend is slightly higher compared to the DIR and FH estimates for *Rangpur* division over the 1994-2000 period. For MTS-II and MTS-III, the trend is somewhat higher in *Rajshahi* division during 2011-2014 period compared to the DIR and FH estimates. All three MTS models show slightly bow-shaped 95% CI bands in between two subsequent survey years, which indicates slightly higher uncertainty during the non-survey years compared to the survey years.

Figure 6.5 National level trends of ANC4 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).

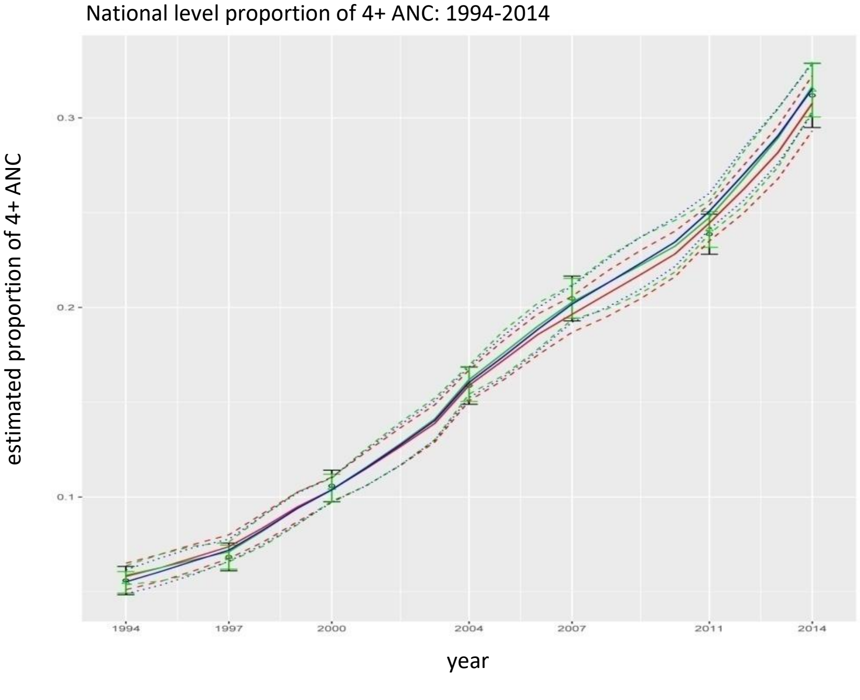
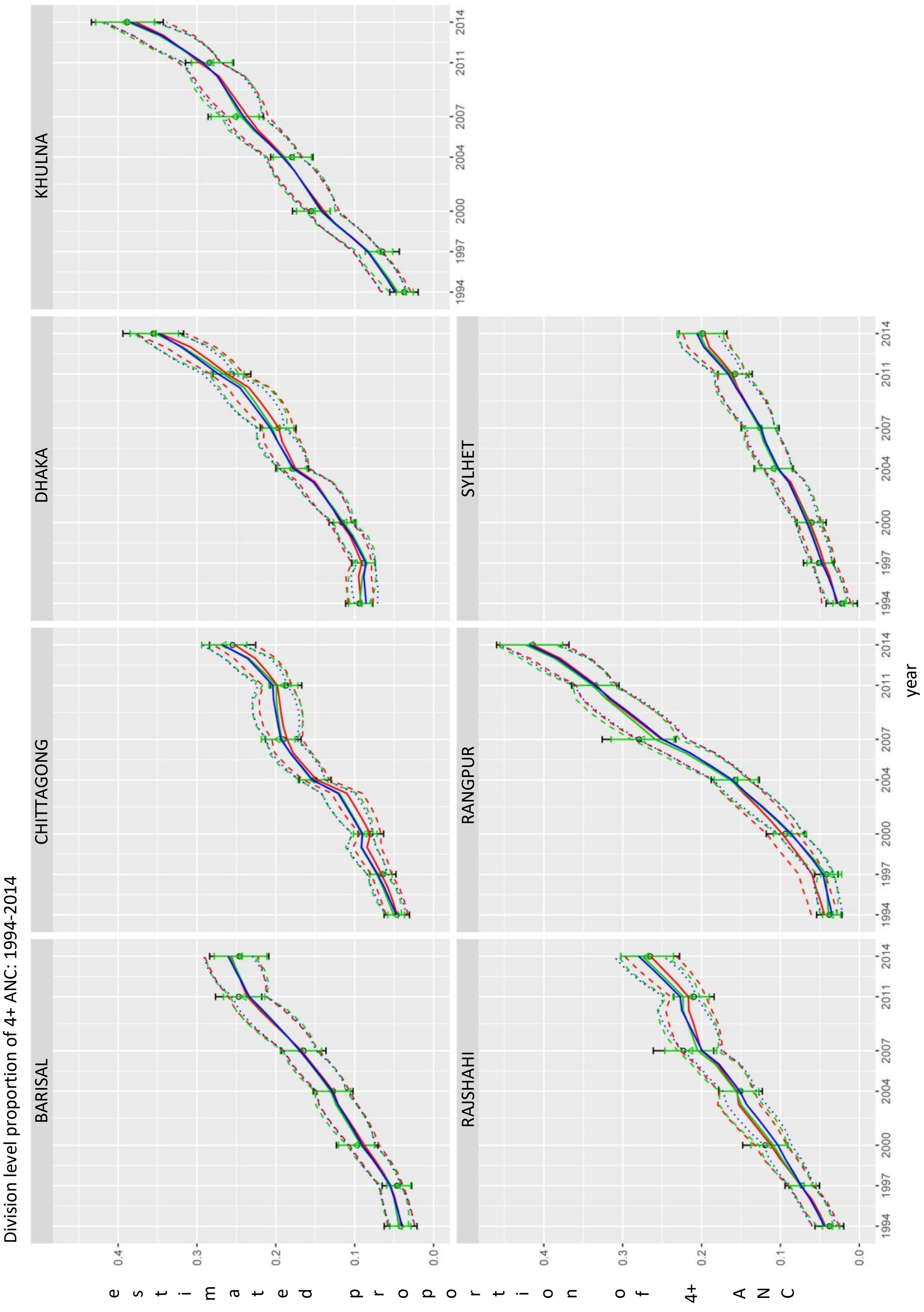


Figure 6.6 Division level trends of ANC4 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).



Although the trends based on MTS-II and MTS-III are almost identical at national and division levels, the estimated variance components of both model differ considerably as follows from Table 6.2. These differences lead to substantial differences in the trend estimates at the district level for MTS-II and MTS-III. Plots for some of the districts are provided in Figures 6.7 and 6.8. See Das, van den Brakel, Boonstra and Haslett (2021) for plots of all districts. Similar to ANC0, the trends of ANC4 under MTS-III are smoother than those under MTS-II. The smaller variance components of MTS-III also result in narrower confidence bands compared to MTS-II.

Table 6.2
Posterior means of standard deviation parameters of random components of MTS-I, MTS-II, MTS-III models for ANC4. No superscript refers to district level, superscripts (*div*) refers to division level

Model	$\hat{\sigma}_I$ **(SE)**	$\hat{\sigma}_S$ **(SE)**	$\hat{\rho}_{IS}$ **(SE)**	$\hat{\sigma}_{Sp}$ **(SE)**	$\hat{\sigma}_{R1}^{(div)}$ **(SE)**	$\hat{\sigma}_{R2}$ **(SE)**
MTS-I	0.060 (0.010)	0.033 (0.005)	0.428 (0.178)	0.047 (0.026)	0.012 (0.004)	0.009 (0.001)
MTS-II	0.046 (0.007)	0.022 (0.003)	0.501 (0.162)	0.035 (0.018)	0.016 (0.003)	0.004 (0.001)
MTS-III	0.038 (0.006)	0.018 (0.006)	0.522 (0.165)	0.027 (0.016)	0.014 (0.003)	0.002 (0.001)

The trend estimates under MTS-I are volatile and show unexpected downward trends for some districts, see for example *Bhola* and *Pirojpur* districts of Barisal division, *Gazipur*, *Kishoreganj* and *Manikganj* of *Dhaka* division, *Bogra*, *Chapai Nababganj* and *Rajshahi* districts of *Rajshahi* division, and *Habiganj* district of *Sylhet* division in Figure 6.7. From a subject matter point of view, such strong movements and turning points are not expected for ANC4 coverage. Therefore it appears that MTS-I follows the DIR estimates too strongly. The trends under MTS-II generally ignore these volatilities and show reasonably smooth trends for these districts. The trends under MTS-III are even smoother for some of these districts, as for example *Bogra* and *Habiganj* districts in Figure 6.7, and *Mymensingh* and *Sylhet* districts in Figure 6.8.

The main difficulty arises for the three hilly districts of *Chittagong* division, i.e., *Khagrachhari*, *Rangamati*, and *Lakshmipur* (the first two districts are plotted in Figure 6.8). MTS-I shows very poor trend estimates for ANC4 over the whole period mainly due to the erratic DIR estimates, which are either zero or highly inconsistent in most of the surveys. The cross-sectional FH estimates are more robust and consequently MTS-II and MTS-III show reasonable upward trends for ANC4. It is expected that women residing in urbanized and better socioeconomic areas are supposed to receive more ANC visits compared to those residing in rural and poor socioeconomic areas. MTS-I shows in some districts lower and in other districts higher than expected trend estimates over the whole time period. For example, the trend obtained with MTS-I for *Narsingdi* in Figure 6.8, which is a highly urbanized district of *Dhaka* division, is lower than expected. Similarly the trend under MTS-I *Munshiganj* in Figure 6.8, which is a less urbanized district of *Dhaka* is higher than expected. Similarly the trend estimates under MTS-I are over the whole period higher than expected in *Meherpur* district of *Khulna* division, *Lalmonirhat* and *Panchagarh* districts of *Rangpur* division. The trend estimates under MTS-II and MTS-III seem more plausible because the cross-sectional FH estimates appear to be more realistic than the DIR estimates. Overall, as in the case of ANC0, MTS-II is a good compromise between MTS-I and MTS-III.

Figure 6.7 District level trends of ANC4 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).

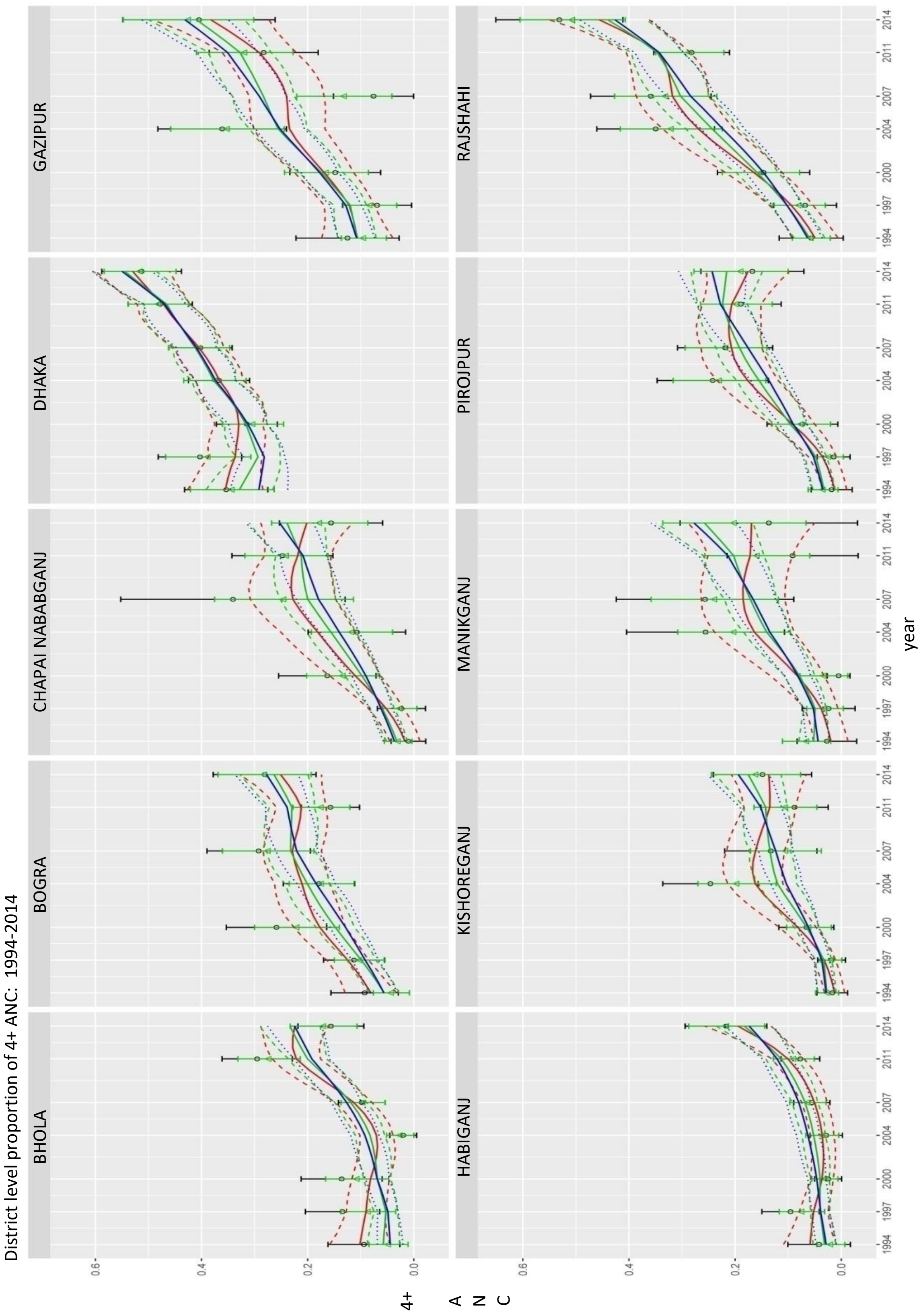
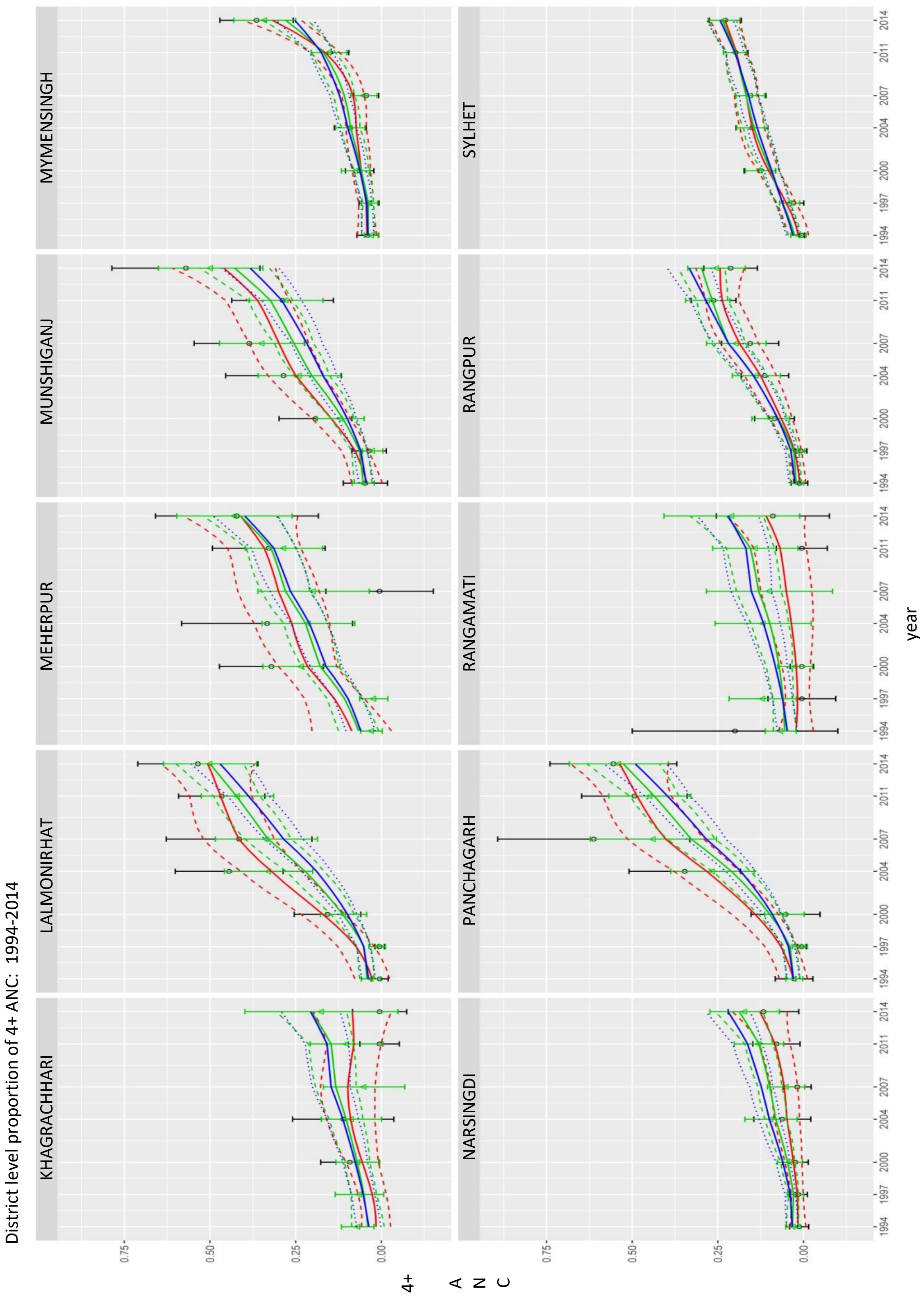


Figure 6.8 **District level trends of ANC4 in Bangladesh: (i) DIR (black error-bar line), (ii) cross-sectional FH (green error-bar line), (iii) MTS-I (red line), (iv) MTS-II (green line) and (v) MTS-III (blue line).**



7. Model assessment

In this study, models were selected based on the WAIC, DIC and graphical comparisons of their trend predictions at three hierarchical levels. In addition to these model diagnostics, three discrepancy measures are defined to evaluate and compare the time-series multilevel models. The first two measures are the Relative Bias (RB) and Absolute Relative Bias (ARB), which express the differences between model estimates and direct estimates, as percentage of the latter. For a given model, RB_{it} and ARB_{it} for domain i and (survey) year t are defined as

$$\mathrm{RB}_{it} = \frac{(\hat{\theta}_{it} - \hat{Y}_{it})}{\hat{Y}_{it}} \times 100\%, \tag{7.1}$$

$$\mathrm{ARB}_{it} = \frac{\left|\hat{\theta}_{it} - \hat{Y}_{it}\right|}{\hat{Y}_{it}} \times 100\% \tag{7.2}$$

with $\hat{\theta}_{it}$ the model prediction and $\hat{Y}_{it}$ the direct estimate. The third discrepancy measure is the Relative Reduction of the Standard Errors (RRSE), which measures the percentage of reduction in standard error of the model-based estimates compared to the direct estimates, i.e.,

$$\mathrm{RRSE}_{it} = 100\% \times \left(\mathrm{se}(\hat{Y}_{it}) - \mathrm{se}(\hat{\theta}_{it})\right) \Big/ \mathrm{se}(\hat{Y}_{it}). \tag{7.3}$$

The RRSE measure should not be interpreted too strictly, since design-based and model-based standard errors are conceptually different quantities. However, both are commonly used as measures of uncertainty, and once reasonable models that sufficiently account for variations over all levels of interest have been selected, based on other criteria, it is informative to use the RRSE as one of the comparison measures.

These three discrepancy measures are calculated at national, division and district (i.e., most detailed) levels. The distributions of these measures are presented in terms of the minimum value, 1st quartile (Q_1), median, mean, 3rd quartile (Q_3) and maximum value.

Additionally, observed coverage rate (CR expressed in %) for 95% confidence interval of the considered cross-sectional FH and MTS models are calculated at division and district levels by identifying whether the estimated 95% confidence interval (CI) of $\hat{\theta}_{it}$ contains the direct estimates $(\hat{Y}_{it})$. Coverage at the district level is the percentage of district by year combinations (about 7×64 domains) where the direct estimate is included in the CI of $\hat{\theta}_{it}$. Coverage at the division level is the percentage of division by year combinations $(7 \times 7$ domains) where the direct estimate is included in the CI of $\hat{\theta}_{it}$. Coverage rates are defined in a similar way for each survey year by averaging over all available districts in one particular survey year. Finally coverage is calculated for each division seperately by averaging over the 7 survey years.

The distributions of the RB_{it} (7.1), ARB_{it} (7.2) and RRSE_{it} (7.3) for three administrative levels are provided in Tables 7.1, 7.2, and 7.3 for ANC0 and ANC4 for the cross-sectional FH, MTS-I, MTS-II, and MTS-III models. Table 7.1 shows that FH and MTS-I models provide lower mean RB for ANC0 and

ANC4 at all three levels, while MTS-II provides slightly lower mean RB compared to MTS-III model at the district level. The ARB distributions in Table 7.2 show that the performance of MTS-II is in between MTS-I and MTS-III for all administrative levels except the national level for ANC4. The ARB values are the lowest for the cross-sectional FH model. It is also observed that the ARB increases as the domain sample size becomes smaller. Table 7.3 shows that MTS-II has the highest RRSE values at national and division levels, while at district level this model shows slightly lower RRSE than the MTS-III model for both ANC0 and ANC4. The variance reduction increases as the domain sample sizes become smaller. The reason that standard errors for the trends at national and division level under MTS-II are smaller than MTS-III is because under MTS-II the covariances between the cross-sectional FH predictions at the district level in the input series are ignored. These covariances are predominantly positive and therefore the standard errors of trends at aggregated levels are higher and more realistic under MTS-III. The higher RB, ARB and RRSE values for models MTS-II and MTS-III are a consequence of the more smooth trends obtained under both models. Small variances under smooth trends imply a larger amount of bias with respect to the direct estimates. As discussed in Section 6, these trends are more plausible compared to the cross-sectional FH model and MTS-I model, since from a subject matter point of view a smooth decreases for ANC0 and increase for ANC4 are expected.

Table 7.1
Summary statistics of relative bias (RB, in %) at different aggregation levels for the SAE estimates of ANC0 and ANC4

Parameter	Aggregation level	Model	Min.	Q_1	Median	Mean	Q_3	Max.
ANC0	Nation	FH	-0.48	-0.20	0.23	0.08	0.37	0.47
		MTS-I	-1.84	-0.68	0.54	-0.05	0.73	0.88
		MTS-II	-1.16	-0.55	0.29	0.41	1.19	2.43
		MTS-III	-1.53	-0.80	-0.57	-0.02	0.60	2.35
	Division	FH	-0.68	-0.48	-0.36	0.05	0.50	1.31
		MTS-I	-0.99	-0.50	-0.31	0.05	0.64	1.41
		MTS-II	-0.77	0.04	0.15	0.59	1.08	2.50
		MTS-III	-1.44	-0.37	0.13	0.15	0.89	1.35
	District	FH	-8.77	-1.72	0.14	0.31	1.67	12.41
		MTS-I	-10.35	-1.24	-0.49	-0.66	0.30	1.87
		MTS-II	-7.87	-1.15	0.77	1.25	2.89	18.34
		MTS-III	-10.05	-2.63	0.89	1.34	3.91	21.43
ANC4	Nation	FH	-1.65	-0.62	0.07	-0.07	0.65	1.04
		MTS-I	-4.09	-1.60	0.05	1.00	3.19	7.88
		MTS-II	-1.85	0.27	1.98	1.91	3.80	5.07
		MTS-III	-2.00	-1.35	1.06	1.11	3.10	5.23
	Division	FH	-1.33	-0.60	-0.13	0.24	0.43	3.47
		MTS-I	-1.17	-0.25	-0.04	-0.07	0.32	0.59
		MTS-II	-0.50	0.68	1.18	1.55	1.70	5.39
		MTS-III	-2.08	0.31	0.73	1.24	1.92	5.58
	District	FH	-17.83	-4.85	0.40	2.08	6.78	64.77
		MTS-I	-16.32	-3.80	-0.56	-0.42	2.98	15.57
		MTS-II	-22.00	-5.30	0.57	4.57	12.47	84.31
		MTS-III	-29.92	-8.23	0.57	6.12	14.07	124.63

This conclusion is confirmed by the CR values shown in Table 7.4. The CRs for the cross-sectional FH models are too high, indicating that the FH predictions tend too much to the direct estimates. The CR levels are reasonably good for MTS-I, substantially lower for MTS-II and the lowest for MTS-III. The lower coverage rates of MTS-II and MTS-III at the district level is reflected by the corresponding higher ARB and higher RRSE. These findings show that MTS-I model predictions are more volatile and tend to the direct estimates, MTS-III model predictions are highly smoothed, and MTS-II model predictions seem like a reasonable compromise between MTS-I and MTS-III model predictions, particularly at the district level.

Table 7.2
Summary statistics of absolute relative bias (ARB, in %) at different aggregation levels for the SAE estimates of ANC0 and ANC4

Parameter	Aggregation level	Model	Min.	Q_1	Median	Mean	Q_3	Max.
ANC0	Nation	FH	0.04	0.27	0.42	0.34	0.46	0.48
		MTS-I	0.26	0.63	0.75	0.87	0.99	1.84
		MTS-II	0.29	0.44	0.58	1.05	1.59	2.43
		MTS-III	0.49	0.61	0.96	1.18	1.61	2.35
	Division	FH	0.39	0.50	0.65	0.90	1.20	1.84
		MTS-I	0.48	0.66	0.78	1.39	2.13	2.90
		MTS-II	0.79	0.96	1.56	1.78	2.14	3.88
		MTS-III	1.00	1.14	1.41	1.88	2.43	3.61
	District	FH	1.08	2.73	4.17	5.12	5.84	15.94
		MTS-I	1.48	3.93	6.58	7.53	9.02	26.67
		MTS-II	3.15	6.46	10.31	11.32	14.50	33.01
		MTS-III	4.15	8.65	12.54	13.49	16.98	38.16
ANC4	Nation	FH	0.07	0.25	0.92	0.76	1.08	1.65
		MTS-I	0.05	1.60	2.47	3.09	4.00	7.88
		MTS-II	0.97	1.68	1.98	2.71	3.80	5.07
		MTS-III	1.06	1.19	1.46	2.46	3.53	5.23
	Division	FH	0.98	1.40	1.71	1.87	2.06	3.47
		MTS-I	1.96	3.06	4.31	4.07	4.64	6.82
		MTS-II	2.18	3.66	4.68	4.33	5.07	6.00
		MTS-III	3.66	4.60	5.36	5.27	5.63	7.46
	District	FH	1.93	7.64	12.91	14.29	17.60	64.77
		MTS-I	3.86	14.27	18.72	20.61	28.10	53.45
		MTS-II	7.07	19.47	26.22	28.51	35.88	84.31
		MTS-III	8.62	21.36	29.32	33.13	41.00	124.63

Table 7.3
Summary statistics of relative reduction of standard errors (RRSE in %) at different aggregation levels for the SAE estimates of ANC0 and ANC4

Parameter	Aggregation level	Model	Min.	Q_1	Median	Mean	Q_3	Max.
ANC0	Nation	FH	-0.65	4.03	8.10	8.00	12.72	15.01
		MTS-I	-0.03	1.35	4.01	3.67	5.82	7.33
		MTS-II	4.07	7.90	13.71	12.89	17.47	21.68
		MTS-III	-3.52	1.02	3.30	3.69	7.15	9.67
	Division	FH	2.99	5.82	7.56	7.03	8.64	9.75
		MTS-I	2.66	4.00	5.32	5.16	6.65	6.84
		MTS-II	8.47	12.74	13.70	13.12	14.34	15.53
		MTS-III	3.30	4.71	5.21	5.47	6.12	8.16
	District	FH	-1.60	7.17	10.20	9.98	12.04	21.61
		MTS-I	7.91	15.16	17.81	18.06	21.15	27.47
		MTS-II	12.60	27.84	34.08	33.80	38.46	48.53
		MTS-III	19.48	32.61	38.40	37.79	41.55	52.71
ANC4	Nation	FH	8.58	11.22	11.66	13.71	14.49	24.32
		MTS-I	6.64	12.16	14.60	14.75	18.50	20.66
		MTS-II	17.79	22.87	23.56	25.12	27.99	32.75
		MTS-III	10.33	16.58	19.45	18.15	21.04	22.04
	Division	FH	11.08	11.80	14.07	14.23	16.39	18.08
		MTS-I	11.82	14.31	14.46	15.78	18.18	19.17
		MTS-II	20.32	24.96	27.39	26.34	28.15	30.45
		MTS-III	15.49	20.37	21.75	21.72	24.51	25.05
	District	FH	0.34	11.62	16.77	17.63	22.60	38.62
		MTS-I	17.79	27.84	30.48	30.93	33.65	43.40
		MTS-II	29.58	43.37	46.86	48.10	54.96	66.75
		MTS-III	35.63	48.88	51.75	52.94	59.31	70.35

Table 7.4
Observed coverage rate (CR in %) of the model predictions for 95% confidence interval at district and division levels as well as district level by survey years for the SAE estimates of ANC0 and ANC4

Parameter	Model	Year wise CR at District Level							Overall CR by Level	
		1994	1997	2000	2004	2007	2011	2014	District	Division
ANC0	FH	100.00	98.33	100.00	100.00	100.00	98.36	100.00	99.53	100.00
	MTS-I	100.00	90.00	93.44	88.52	93.22	98.36	100.00	94.81	100.00
	MTS-II	88.33	63.33	70.49	67.21	71.19	75.41	91.53	75.10	95.92
	MTS-III	83.33	53.33	50.82	52.46	61.02	55.74	79.66	62.22	95.92
ANC4	FH	98.15	98.28	100.00	100.00	100.00	100.00	90.20	98.36	100.00
	MTS-I	87.04	84.48	68.33	76.27	81.97	96.72	100.00	84.58	95.92
	MTS-II	44.44	51.72	50.00	52.54	62.30	65.57	76.47	57.55	97.96
	MTS-III	44.44	41.38	40.00	38.98	50.82	55.74	72.55	48.70	97.96

8. Discussion

In this study, multilevel time-series (MTS) models have been developed for the percentage of women receiving no antenatal consult (ANC0) and the percentage of women receiving at least 4 antenatal consults (ANC4) in Bangladesh, using only seven editions of the Bangladesh Demographic and Health Survey (BDHS) over the period of 1994-2014. Time series models are defined at an annual frequency where years without a survey edition are treated as missing. In this way, the model accounts for the varying time gaps between the subsequent editions of the BDHS and produce predictions in the years without sample surveys. Trends are produced at three regional levels, namely the national level, a break down in 7 divisions and a breakdown in 64 districts.

In the first model (MTS-I) year-domain-specific direct estimates and their standard errors are used as input in the MTS model. Trends obtained under this model, are rather volatile since the trend estimates tend to follow the direct estimates. Another drawback of the MTS-I model is that it hampers the use of auxiliary information from two available censuses, since values for the auxiliary information from a particular census does not change in two or three subsequent editions of the survey. To use this census information, it is proposed to develop cross-sectional Fay-Herriot (FH) models for each survey year separately. In a second MTS model, MTS-II, these FH estimates and their standard errors are used as input series. In a third model, called MTS-III, the FH estimates with their full covariance matrices, are used as input series. This MTS model properly accounts for the cross-sectional correlations between the FH estimates. The overall model for MTS-II and MTS-III is then a two-step non-iterated process, for which the first stage is producing the FH estimates.

The models are developed at the most detailed regional level of districts. Division and national level trends are estimated by aggregating predictions of the district level trends. In this way, figures at different aggregation levels are numerically consistent by definition.

Compared to other time series small area estimation models proposed in the literature, our models contain more structure, since dynamic trend models are specified at different aggregation levels. This is necessary to obtain accurate aggregated predictions for the divisions and the national level and is a more parsimonious way of modelling cross-sectional correlations. Further model regularization was considered by specifying global-local priors. This, however, did not further improve the model fits.

In small area estimation, domain estimates are often benchmarked to the direct estimates at the national level for numerical consistency and as an attempt to reduce the bias in the model based domain predictions. In this application the trend estimates at the national level under the MTS models are already very close to the direct estimates. Therefore we do not consider an additional benchmark step.

All three time series models provide estimates with improved accuracy. Because MTS-II ignores the predominantly positive correlations between the cross-sectional FH input series, the standard errors of the trends at aggregated levels are actually too small. Since MTS-III accounts for these correlations, the standard errors for national and division trends are larger but also more realistic. The MTS-II model, however, seems to provide most plausible trends for both response variables, particularly at the district level, by compromising volatility in the trends under the MTS-I model and flatness in the trends under the MTS-III model. This choice is supported by the fact that these variables are likely to be relatively smooth over time. Fitting these models to the series of ANC0 and ANC4 is therefore certainly suitable in concept. This also justifies the interpolation of the trends for the years without sample surveys. Our approach can be useful also for many developing countries with repeated DHS surveys, since these are typically observed with varying time lags and mainly depend on census information that is not updated within two or three subsequent editions of the survey.

Using predictions of the cross-sectional FH models as input series for the MTS models, is proposed as a practical solution to make better use of the available census information. The additional advantage of this approach is that it stabilizes the input series for the MTS models by removing large sampling errors from the direct estimates. This requires, however, a careful model selection and evaluation process for the cross-sectional FH models, since model miss-specification of the cross-sectional FH models can result in biased input series with estimated standard errors that underestimate the real uncertainty.

One limitation of this study is related to the bias correction for the square root transformation that is applied to ANC4. The bias correction can only be applied to the trend estimates in the survey years. This results in awkward increases of point estimates if the sampling error is not smoothed enough, particularly for the domains with small sample size. This hampers estimation of period-to-period changes between survey years and non-survey years. Therefore the bias correction is only applied to the cross-sectional FH models and not to the MTS models.

The prevalence of ANC0 and ANC4 visits are negatively correlated, so a multivariate model may be an interesting alternative to the univariate models used here. The two series could be combined with the series of the remainder category in a single multivariate model. This, however, requires a multinomial model that has the advantages that it may further improve the precision of the estimates and guarantees that the predictions take values in their admissible range, and that the predictions over the categories add up to hundred percent. The multinomial model is, however, not easy to implement. Particularly in this study the variance-covariance matrix can be difficult to estimate for the districts with small number of observation. Furthermore, Datta et al. (2002) shows that univariate models may provide as good results as multivariate models proposed in Ghosh, Nangia and Kim (1996), while being simpler to implement. The extension of our univariate models to a multinomial model is therefore left for further research.

For ANC0, the national level shows a downward trend. The decline in the trend temporarily stopped during 2004-2011. The trend of ANC4 shows steady increase over the considered study period. Division level trends for ANC0 show a steady decline for all the divisions except *Dhaka*, *Chittagong* and *Sylhet* divisions. The trends for these three divisions remained stable during the period of 2004-2011 which mainly causes the flat trend at the national level of ANC0. On the other hand, at the division level ANC4 shows almost linear upward trends for most of the divisions except *Dhaka* and *Chittagong*. The greatest improvement is observed for *Khulna* and *Rangpur* divisions where the trends of ANC4 reach to more than 40% in 2014. District-level trends help to identify highly vulnerable districts in terms of the two considered response variables. Though the national level trend of ANC0 declines to about 21% in 2014, a few districts get below 10% (*Dhaka*, *Jhenaidaha*, and *Meherpur*) while a considerable number of districts still have ANC0 higher than 35% (*Bhola*, *cox's Bazar*, *Kishoregonj*, *Noakhali*, *Sunamganj*, *Sirajgonj*, and three *Chittagong* hill tract districts). For ANC4, a few districts have estimates above 50% (*Dhaka*, *Nilphamari*, and *Panchagarh*) and most of the districts with high ANC0 have ANC4 estimates less than 20%. These district level trends might help policy makers to focus on vulnerable hotspots where both ANC0 and ANC4 indicators are still poor. Obviously, detailed level trends might help policy makers to take actions for reducing disaggregated level inequalities in the race to achieve SDGs.

Acknowledgements

We wish to thank Measure Evaluation and National Institute of Population Research and Training (NIPORT) for making the BDHS data publicly available. In addition, IPUMS deserves thanks for providing the access to the sample data of Bangaldesh Census 1991, Census 2001, and Census 2011. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors are grateful to two anonymous reviewers and the Associate Editor for providing useful comments on a former draft of the paper.

Appendix

Table A.1
District level contextual variables generated from Census 1991, Census 2001, and Census 2011 data for ANC0

Variable	Definition
Division	Barishal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, Sylhet.
Region	(1) Densely populated Dhaka, Chittagong and Gazipur districts, (2) 9 regional districts with big cities, (3) 3 hilly districts (Bandarban, Khagrachhari and Rangamati), (4) 49 other districts (less urbanized areas).
Chittagong	Chittagong Division?
Dhaka	Dhaka Division?
Khulna	Khulna Division?
Rangpur	Rangpur Division?
Rajshahi	Rajshahi Division?
P_U5	Proportion of Under-5 children.
P_W	Proportion of women aged 15-49 years.
P_MW	Proportion of married women aged 15-49 years.
P_MW_Prim_Edu	Proportion of married women aged 15-49 years having primary education.
P_MW_Sec_Edu	Proportion of married women aged 15-49 years having at least secondary education.
P_HH_No_Edu_W	Proportion of household (HH) with illiterate women aged 15-49 years.
P_HH_Prim_Edu_W	Proportion of household (HH) with primary educated women aged 15-49 years.
P_HH_High_Edu_W	Proportion of household (HH) with higher educated women aged 15-49 years.
P_HH_Sec_Edu_Head	Proportion of HH with at least secondary educated HH head.
P_Ru_HH_4^+	Proportion of rural HH of size 4 and more.
P_Ru_HH_Elec	Proportion of rural HH with electricity.
P_Ru_HH_Sing_Moth	Proportion of rural HH with single mother.
P_HH_U5_Sec_Edu_W	Proportion of HH having under-5 children and women aged -49 years having at least secondary education.
P_HH_2^+_U5	Proportion of HH with 2 or more under-5 children.
P_Ru_HH_U5	Proportion of rural HH with under-5 children.
P_Ru_HH_2^+_U5	Proportion of rural HH with 2 or more under-5 children.
P_Ur_HH_2^+_U5	Proportion of urban HH with 2 or more under-5 children.

Table A.2
Fixed and Random effects of survey-year specific FH models for ANC0

Survey Year	Transformation	Fixed Effects	Random Effect	Census Data
1994	No	1 + Division + P_HH_High_Edu_W + P_Ur_HH_2$^+$_U5	RI: District level Random Intercept	1991
1997	No	1 + Division + P_MW_Sec_Edu + P_HH_Sec_Edu_Head	RI	1991
2000	No	1 + Division + P_Ru_HH_U5 + P_W	RI	1991
2004	No	1 + Division + P_HH_U5_Sec_Edu_W + P_MW	RI	2001
2007	No	1 + Division + P_U5 + P_Ru_HH_Size_4$^+$	RI	2001
2011	SQRT	1 + Division + sqrt (P_Ru_HH_U5) + sqrt (P_MW)	RI	2011
2014	SQRT	1 + Division + sqrt (P_Ur_HH_2$^+$_U5) + sqrt (P_Ru_HH_Elec)	RI	2011

Table A.3
Fixed and Random effects of survey-year specific FH models for ANC4

Survey Year	Transformation	Fixed Effects	Random Effect	Census Data
1994	SQRT	1 + Division + sqrt (P_Ru_U5) + sqrt (P_HH_2$^+$_U5) + sqrt (P_Ur_HH_2$^+$_U5)	RI: District level Random Intercept	1991
1997	SQRT	1 + Division + sqrt (P_HH_U5_Sec_Edu_W) + log (P_HH_Sec_Edu_Head)	RI	1991
2000	SQRT	1 + Khulna + Region + sqrt (P_MW$_P$rim$_E$du) + sqrt (P_HH_W$_I$lli$_E$du)	RI	1991
2004	SQRT	1 + Division + sqrt (P_HH_U5_Prim_Edu_W) + sqrt (P_W)	RI	2001
2007	SQRT	1 + Rangpur + Region + sqrt (P_U5) + sqrt (P_Ru_HH_Size_4$^+$)	RI	2001
2011	SQRT	1 + Rangpur + Chittagong + sqrt ((P_HH_U5_Sec_Edu_W) + sqrt (P_W)	RI	2011
2014	SQRT	1 + Rangpur + Chittagong + Rajshahi + Region + sqrt (P_W) + sqrt (P_Ru_HH_Sing_Mot)	RI	2011

References

BBS (2020). *SDG Tracker: Bangladesh Development Mirror*. https://www.sdg.gov.bd/page/thirty_nine_plus_one_indicator/5#1, [Online; accessed 18-December-2020].

BBS and UNICEF (2019). *Progotir Pathey Bangladesh: Bangladesh Multiple Indicator Cluster Survey 2019*. Technical report, Bangladesh Bureau of Statistics (BBS).

Besag, J., and Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biometrika*, 82(4), 733-746.

Bollineni-Balabay, O., van den Brakel, J., Palm, F. and Boonstra, H.J. (2017). Multilevel hierarchical Bayesian versus state space approach in time series small area estimation: The Dutch Travel Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 1281-1308.

Boonstra, H.J. (2021). *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. R package version 0.6.0.

Boonstra, H.J., and van den Brakel, J. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 3, 395-425. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00005-eng.pdf.

Boonstra, H.J., and van den Brakel, J. (2022). Multilevel time series models for small area estimation at different frequencies and domain levels. *Annals of Applied Statistics*, 16, 4, 2314-2338, DOI: 10.1214/21-AOAS1592.

Boonstra, H.J., van den Brakel, J. and Das, S. (2021). Multilevel time series modelling of mobility trends in the Netherlands for small domains. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3), 985-1007.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480.

Das, S., Kumar, B. and Kawsar, L.A. (2020). Disaggregated level child morbidity in Bangladesh: An application of small area estimation method. *PLoS ONE*, 15(5), e0220164.

Das, S., van den Brakel, J., Boonstra, H.J. and Haslett, S. (2021). Multilevel time series modelling of antenatal care coverage in Bangladesh at disaggregated administrative levels. Discussion paper, Statistics Netherlands.

Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102(1), 83-97.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the States of the U.S. *Journal of the American Statistical Association*, 94(448), 1074-1082.

DHS (2021). *The DHS Program: Demographic and Health Surveys*. https://dhsprogram.com/, accessed 2021-11-01.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to Census data. *Journal of the American Statistical Association*, 74(366), 269-277.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91(436), 1423-1431.

Haslett, S., and Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Technical report, Bangladesh Bureau of Statistics and United Nations World Food Programme.

Haslett, S., Jones, G. and Isidro, M. (2014). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Technical report, United Nations World Food Programme and International Fund for Agricultural Development.

Hossain, M.J., Das, S. and Chandra, H. (2020). Disaggregate level estimates and spatial mapping of food insecurity in Bangladesh by linking survey and census data. *PLoS ONE*, 15(4), e0230906.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics & Data Analysis*, 58, 308-325.

Mrisho, M., Obrist, B., Schellenberg, J.A., Haws, R.A., Mushi, A.K., Mshinda, H., Tanner, M. and Schellenberg, D. (2009). The use of antenatal and postnatal care: Perspectives and experiences of women and health care providers in rural southern Tanzania. *BMC Pregnancy and Childbirth*, 9(1), 10.

NIPORT (2015a). *Bangladesh Demographic and Health Survey 2014: Key Indicators*. Technical report, National Institute of Population Research and Training.

NIPORT (2015b). *Bangladesh Demographic and Health Survey 2014: Policy Briefs*. Technical report, National Institute of Population Research and Training.

NIPORT (2016). *Bangladesh Demographic and Health Survey 2014*. Technical report, National Institute of Population Research and Training.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484), 1405-1418.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Pratesi, M., and Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17(1), 113-141.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M.S., Mohiuddin, H., Kafy, A.-A., Sheel, P.K. and Di, L. (2019). Classification of cities in Bangladesh based on remote sensing derived spatial characteristics. *Journal of Urban Management*, 8(2), 206-224.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J.N.K, and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rue, H., and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 56.

Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.

Spiegelhalter, D.J., Best, N.G. Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64(4), 583-639.

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

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 431-439.

WHO, UNICEF and Others (2014). Trends in Maternal Mortality: 1990 to 2013: Estimates by WHO, UNICEF, UNFPA, the World Bank and the United Nations Population Division. Technical report, World Health Organization.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10614-eng.pdf.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 2, 173-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000002/article/5537-eng.pdf.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 439-461
Statistics Canada, Catalogue No. 12-001-X

Optimal linear estimation in two-phase sampling

Takis Merkouris[1]

Abstract

Two-phase sampling is a cost effective sampling design employed extensively in surveys. In this paper a method of most efficient linear estimation of totals in two-phase sampling is proposed, which exploits optimally auxiliary survey information. First, a best linear unbiased estimator (BLUE) of any total is formally derived in analytic form, and shown to be also a calibration estimator. Then, a proper reformulation of such a BLUE and estimation of its unknown coefficients leads to the construction of an "optimal" regression estimator, which can also be obtained through a suitable calibration procedure. A distinctive feature of such calibration is the alignment of estimates from the two phases in an one-step procedure involving the combined first-and-second phase samples. Optimal estimation is feasible for certain two-phase designs that are used often in large scale surveys. For general two-phase designs, an alternative calibration procedure gives a generalized regression estimator as an approximate optimal estimator. The proposed general approach to optimal estimation leads to the most effective use of the available auxiliary information in any two-phase survey. The advantages of this approach over existing methods of estimation in two-phase sampling are shown both theoretically and through a simulation study.

Key Words: Auxiliary information; Best linear unbiased estimation; Calibration; Generalized regression estimation; Double sampling.

1. Introduction

The two-phase sampling design, also called double sampling, has traditionally been used in sample surveys as a cost-effective survey method. In the first phase, a relatively large sample is drawn from the target population to provide auxiliary information that is inexpensive to collect. This sample forms a highly informative frame from which a subsample is drawn in the second phase to collect information on the items of interest. Also, two-phase sampling has been increasingly used as a mechanism for handling nonresponse. Särndal, Swensson and Wretman (1992) provide an extensive account of such uses of two-phase sampling. Groves and Heeringa (2006), and Brick and Tourangeau (2017) discuss the important role of two-phase sampling in responsive designs when costly actions are taken for reduction of non-response bias. Other applications of two-phase sampling, which have emerged in recent survey practice, involve various forms of integration of separate surveys. In one such case, a first-phase sample serves as a frame for the second-phase sample for a multitude of similar surveys (Turmelle and Beaucage, 2013). In another case, a primary large survey is used as a frame for another smaller survey with a larger set of survey items (Australian Bureau of Statistics, 2004).

Auxiliary information in two-phase sampling may be available at different levels. Some information is at the level of the whole population, and other information is at the level of the first-phase sample or the second-phase sample. Much research has been devoted to the use of such information for improved estimation of population totals or means; see Särndal et al. (1992), Hidiroglou and Särndal (1998), Hidiroglou (2001), Estevao and Särndal (2002, 2009), Wu and Luan (2003), Chen and Kim (2014), and

[1] Takis Merkouris, Department of Statistics, Athens University of Economics and Business, 2 Trias Street, Athens 11362, Greece. E-mail: merkouris@aueb.gr.

references therein. In general, two approaches are identified in this literature for incorporating auxiliary information into the estimation process. The generalized regression approach and the calibration approach; the two phases of sampling imply two regression fits or two successive calibrations. Under certain conditions the two approaches lead to identical estimators, but this is not so in general. Variance estimation of these two-phase estimators has been studied extensively; see, for example, Sitter (1997), Fuller (1998), Kim and Sitter (2003), Kim, Navarro and Fuller (2006), Hidiroglou, Rao and Haziza (2008), Kim and Yu (2011), Beaumont, Beliveau and Haziza (2015).

Irrespective of the regression or calibration formulation of the existing estimation procedures, the resulting estimators for a target variable are in effect linear combinations of Horvitz-Thompson estimators of various totals (or means), including the estimator for the target variable derived from the second-phase sample and estimators for auxiliary variables derived from both first-phase and second-phase sample. Taking a formal approach to optimal estimation, in this paper we consider the most efficient linear combination of available estimators from both phases, based on the principle of best linear unbiased estimation. We show that the derived, in analytic form, best linear unbiased estimator (BLUE) possesses a useful orthogonality property and that it can be alternatively constructed as calibration estimator, linear in the values of the associated variable and incorporating the auxiliary information into the calibrated design weigs. Estimation of the unknown coefficients of this BLUE, using all available auxiliary information from both phases of sampling, gives an “optimal” estimator, analogous to the single-phase optimal regression estimator of Montanari (1987) and Rao (1994). This estimator is a large-sample approximation of the BLUE, with the estimated coefficients minimizing its estimated approximate (large sample) variance, and preserving the orthogonality property of the BLUE. With a proper reformulation of the BLUE, the optimal estimator can also be obtained through a suitable calibration procedure. The distinctive feature of such calibration is the convenient one-step procedure of aligning estimates from the two phases using the combined first-and-second phase samples. Optimal estimation is feasible for certain two-phase designs that are used often in large scale surveys. For general designs, an alternative one-step calibration procedure gives a novel generalized regression estimator as a convenient approximation to the optimal estimator.

The proposed general method of estimation guides the construction of calibration estimators in any particular case of two-phase survey, making the most effective use of the available auxiliary information. It also provides an insig into existing less efficient estimation methods when these are placed into the framework of optimal estimation. The advantages of the proposed method over existing methods are shown both theoretically and through a simulation study.

The paper is organized as follows. The structure of the two-phase sampling design, and notation, are introduced in Section 2. The derivation of the BLUE for the standard type of auxiliary information in two-phase sampling, and its alternative construction as a calibration estimator, are described in Section 3. The two-phase optimal estimator and its calibration equivalent are presented in Section 4. The approximation of the optimal estimator by a generalized regression estimator is discussed in Section 5. Comparisons with existing methods are presented in Section 6. A simulation study is presented in Section 7. The paper concludes with a discussion in Section 8.

2. Two-phase sampling design: Structure and notation

Let $U = \{1, \ldots, k, \ldots, N\}$ denote a finite population of N units. A first-phase sample s_1 of size n_1 is drawn from the population U, using a sampling design that defines inclusion probability $\pi_{1k} = P(k \in s_1)$ for unit $k \in U$, and joint inclusion probability $\pi_{1kl} = P(k, l \in s_1)$ for units $k, l \in U$. Then, a second-phase sample s_2 of size n_2 is drawn from s_1 using a sampling design that defines conditional inclusion probability $\pi_{2k} = P(k \in s_2 \mid s_1)$ for $k \in s_1$, and joint conditional inclusion probability $\pi_{2kl} = P(k, l \in s_2 \mid s_1)$ for units $k, l \in s_1$. Assuming that $\pi_{1k} > 0$ for all $k \in U$ and $\pi_{2k} > 0$ for all $k \in s_1$, the first-phase design weight for $k \in s_1$ is $w_{1k} = 1/\pi_{1k}$, the conditional second-phase design weight for $k \in s_2$ is $w_{2k} = 1/\pi_{2k}$, and the overall design weight for $k \in s_2$ is $w_k = w_{1k} w_{2k}$.

The standard type of auxiliary variables in two-phase sampling (see, for example, Särndal et al. (1992)) involves a vector of auxiliary variables $\mathbf{x}$, partitioned as $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ by p and q components of it, with population total $\mathbf{t}_{\mathbf{x}} = \sum_U \mathbf{x}_k$ and known total $\mathbf{t}_{\mathbf{x}_1} = \sum_U \mathbf{x}_{1k}$ of $\mathbf{x}_1$. The value $\mathbf{x}_k$ is observed for every unit $k \in s_1$, whereas for a d-dimensional vector of target variables $\mathbf{y}$, with total $\mathbf{t}_{\mathbf{y}} = \sum_U \mathbf{y}_k$, the value $\mathbf{y}_k$ is observed only for the units $k \in s_2$. In some surveys, components of the vector $\mathbf{x}_2$ are also target variables. An unbiased estimator of the total $\mathbf{t}_{\mathbf{y}}$, the common Horvitz-Thompson (HT) estimator, given by $\tilde{\mathbf{t}}_{\mathbf{y}} = \sum_{s_2} w_k \mathbf{y}_k$, is obtained using the second-phase sample s_2, while two HT estimators of the total $\mathbf{t}_{\mathbf{x}}$, given by $\hat{\mathbf{t}}_{\mathbf{x}} = \sum_{s_1} w_{1k} \mathbf{x}_k$ and $\tilde{\mathbf{t}}_{\mathbf{x}} = \sum_{s_2} w_k \mathbf{x}_k$, are obtained using the samples s_1 and s_2, respectively. In the derivation of results involving these estimators we will use the vector notation $\tilde{\mathbf{t}}_{\mathbf{y}} = \mathbf{Y}_2' \mathbf{w}$, $\hat{\mathbf{t}}_{\mathbf{x}} = \mathbf{X}_1' \mathbf{w}_1$, $\tilde{\mathbf{t}}_{\mathbf{x}} = \mathbf{X}_2' \mathbf{w}$, $\hat{\mathbf{t}}_{\mathbf{x}_1} = \mathbf{X}_{11}' \mathbf{w}_1$, where $\mathbf{w}_1$ and $\mathbf{w}$ denote the vectors of design weights for samples s_1 and s_2, respectively, $\mathbf{X}_1$ and $\mathbf{X}_{11}$ denote the sample s_1 matrices of $\mathbf{x}$ and $\mathbf{x}_1$ of dimensions $n_1 \times (p+q)$ and $n_1 \times p$, respectively, and $\mathbf{Y}_2$, $\mathbf{X}_2$ denote the sample s_2 matrices of $\mathbf{y}$ and $\mathbf{x}$ of dimensions $n_2 \times d$ and $n_2 \times (p+q)$.

The primary target of estimation is the total $\mathbf{t}_{\mathbf{y}}$. However, for better understanding of the construction of the proposed estimators, and because components of the vector $\mathbf{x}_2$ may also be target variables, a unified approach to the estimation of both $\mathbf{t}_{\mathbf{y}}$ and $\mathbf{t}_{\mathbf{x}}$ will be taken.

3. Best linear unbiased estimation in two-phase sampling

3.1 An analytic form of the best linear unbiased estimator

For more efficient estimation of the totals $\mathbf{t}_{\mathbf{y}}$ and $\mathbf{t}_{\mathbf{x}}$, incorporating all the available information from both phases through the correlation of $\mathbf{y}$ and $\mathbf{x}$, we consider the best linear unbiased estimators (BLUE), denoted by $\hat{\mathbf{t}}_{\mathbf{y}}^B$ and $\hat{\mathbf{t}}_{\mathbf{x}}^B$, which are minimum-variance linear unbiased combinations of the four estimators $\tilde{\mathbf{t}}_{\mathbf{y}}$, $\hat{\mathbf{t}}_{\mathbf{x}}$, $\tilde{\mathbf{t}}_{\mathbf{x}}$, $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ and given in matrix form by

$$\left(\hat{\mathbf{t}}_{\mathbf{y}}^{B\prime}, \hat{\mathbf{t}}_{\mathbf{x}}^{B\prime}\right)' = \boldsymbol{\mathcal{P}}\left(\tilde{\mathbf{t}}'_{\mathbf{y}}, \hat{\mathbf{t}}'_{\mathbf{x}}, \tilde{\mathbf{t}}'_{\mathbf{x}}, \mathbf{t}'_{\mathbf{x}_1} - \hat{\mathbf{t}}'_{\mathbf{x}_1}\right)', \tag{3.1}$$

where $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1},$ the matrix $\mathbf{W}$ has entries 1's and 0's and satisfies $E\left[(\tilde{\mathbf{t}}'_{\mathbf{y}}, \hat{\mathbf{t}}'_{\mathbf{x}}, \tilde{\mathbf{t}}'_{\mathbf{x}}, \mathbf{t}'_{\mathbf{x}_1} - \hat{\mathbf{t}}'_{\mathbf{x}_1})'\right] = \mathbf{W}(\mathbf{t}'_{\mathbf{y}}, \mathbf{t}'_{\mathbf{x}})',$ and $\mathbf{V}$ is the covariance matrix of $(\tilde{\mathbf{t}}'_{\mathbf{y}}, \hat{\mathbf{t}}'_{\mathbf{x}}, \tilde{\mathbf{t}}'_{\mathbf{x}}, \mathbf{t}'_{\mathbf{x}_1} - \hat{\mathbf{t}}'_{\mathbf{x}_1})'.$ It follows that $\operatorname{Var}\left[(\hat{\mathbf{t}}_{\mathbf{y}}^{B\prime}, \hat{\mathbf{t}}_{\mathbf{x}}^{B\prime})'\right] = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}.$ This typical formulation of best linear unbiased estimation has been explored in two other areas of survey sampling; see Wolter (1979), Jones (1980), Fuller (1990), and Chipperfield and Steel (2009). In the present context, a more practical formulation, which leads also to the representation of the BLUE as a calibration estimator, is as follows.

Writing the two linear combinations in (3.1) in expanded form and using the condition of unbiasedness $E(\hat{\mathbf{t}}_{\mathbf{y}}^{B}) = \mathbf{t}_{\mathbf{y}}$ and $E(\hat{\mathbf{t}}_{\mathbf{x}}^{B}) = \mathbf{t}_{\mathbf{x}},$ it is easy to show that the matrix $\boldsymbol{\mathcal{P}}$ of the coefficients in these linear combinations satisfies

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{3\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{3\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{0} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix},$$

and then the two components of the BLUE in (3.1) are written in the regression form

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{B} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) + \mathbf{B}_{4\mathbf{y}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{B} &= \tilde{\mathbf{t}}_{\mathbf{x}} + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) + \mathbf{B}_{4\mathbf{x}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.2}$$

Now we can write (3.1) as

$$\begin{pmatrix} \hat{\mathbf{t}}_{\mathbf{y}}^{B} \\ \hat{\mathbf{t}}_{\mathbf{x}}^{B} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{t}}_{\mathbf{y}} \\ \tilde{\mathbf{t}}_{\mathbf{x}} \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}} \\ \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}, \tag{3.3}$$

where the matrix $\boldsymbol{\mathcal{B}}$ consists of the second and fourth columns of $\boldsymbol{\mathcal{P}},$ and has the easily derived variance-minimizing value

$$\boldsymbol{\mathcal{B}} = -\operatorname{Cov}\left[\begin{pmatrix} \tilde{\mathbf{t}}_{\mathbf{y}} \\ \tilde{\mathbf{t}}_{\mathbf{x}} \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}} \\ \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}\right] \left[\operatorname{Var}\begin{pmatrix} \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}} \\ \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}\right]^{-1}. \tag{3.4}$$

Next write

$$\mathbf{w}^{*} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w} \end{pmatrix}, \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{X}_{11} \\ \mathbf{X}_2 & \mathbf{0} \end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{Y}_2 & \mathbf{X}_2 \end{pmatrix}, \tag{3.5}$$

so that

$$\boldsymbol{\mathcal{X}}'\mathbf{w}^{*} = \begin{pmatrix} \tilde{\mathbf{t}}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}} \\ \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}, \qquad \boldsymbol{\Psi}'\mathbf{w}^{*} = \begin{pmatrix} \tilde{\mathbf{t}}_{\mathbf{y}} \\ \tilde{\mathbf{t}}_{\mathbf{x}} \end{pmatrix}, \tag{3.6}$$

and $\mathcal{B}$ may then be expressed as $\boldsymbol{\mathcal{B}} = \operatorname{Cov}(\boldsymbol{\Psi}'\mathbf{w}^{*}, \boldsymbol{\mathcal{X}}'\mathbf{w}^{*})\left[\operatorname{Var}(\boldsymbol{\mathcal{X}}'\mathbf{w}^{*})\right]^{-1}.$ For the calculation of variances and covariances we define $\mathbf{w}^{*}$ at the population level as $\mathbf{w}_{U}^{*} = (\mathbf{w}'_{1U}, \mathbf{w}'_{U})',$ where the k^{th} element of $\mathbf{w}_{1U}$ is $w_{1U_k} = (1/\pi_{1k})\, I_{1k},$ the indicator variable I_1 denoting inclusion of a population unit in

s_1, and the k^{th} element of $\mathbf{w}_U$ is $w_{U_k} = \left[1/(\pi_{1k}\pi_{2k})\right] I_{1k} I_{2k}$, the indicator variable I_2 denoting inclusion of a population unit in s_2 conditional on the selection of sample s_1. We may now write $\mathcal{X}'\mathbf{w}^* = \mathcal{X}'_U\, \mathbf{w}^*_U$ and $\boldsymbol{\Psi}'\mathbf{w}^* = \boldsymbol{\Psi}'_U\, \mathbf{w}^*_U$, where $\mathcal{X}_U$ and $\boldsymbol{\Psi}_U$ are the population counterparts of $\mathcal{X}$ and $\boldsymbol{\Psi}$, respectively; all submatrices in $\mathcal{X}$ and $\boldsymbol{\Psi}$ are expanded to population level, having N rows. Then, denoting $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^*$ and $\hat{\mathbf{t}}_{\mathcal{X}} = \mathcal{X}'\mathbf{w}^*$, we get

$$\mathcal{B} = \operatorname{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathcal{X}})\right]^{-1} = \boldsymbol{\Psi}'_U \operatorname{Var}(\mathbf{w}^*_U)\, \mathcal{X}_U \left[\mathcal{X}'_U \operatorname{Var}(\mathbf{w}^*_U)\, \mathcal{X}_U\right]^{-1}. \tag{3.7}$$

A useful more analytic expression of $\mathcal{B}$ is then obtained using the following Lemma; the proof is in the Appendix.

Lemma 1

$$\operatorname{Var}(\mathbf{w}^*_U) = \begin{pmatrix} \operatorname{Var}(\mathbf{w}_{1U}) & \operatorname{Var}(\mathbf{w}_{1U}) \\ \operatorname{Var}(\mathbf{w}_{1U}) & \operatorname{Var}(\mathbf{w}_U) \end{pmatrix}, \tag{3.8}$$

where $\operatorname{Var}(\mathbf{w}_{1U}) = \left\{(\pi_{1kl} - \pi_{1k}\pi_{1l})/\pi_{1k}\pi_{1l}\right\}$, $\operatorname{Var}(\mathbf{w}_U) = \left\{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l})/\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\right\}$.

Using (3.7) and (3.8), it is easy to show that (3.4) is expressed as

$$\mathcal{B} = \begin{bmatrix} -\operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1} & \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1} \\ \mathbf{I} & \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1} \end{bmatrix}. \tag{3.9}$$

Implicit in this representation of $\mathcal{B}$ is the property $\operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$, following from (3.8), implying that $\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \operatorname{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$, and the property $\operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$, implying $\operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ (this covariance being the off-diagonal block of $\mathcal{X}'_U \operatorname{Var}(\mathbf{w}^*_U)\, \mathcal{X}_U$). Then (3.2) can be written explicitly as

$$\begin{aligned} \hat{\mathbf{t}}^B_{\mathbf{y}} &= \tilde{\mathbf{t}}_{\mathbf{y}} - \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}^B_{\mathbf{x}} &= \hat{\mathbf{t}}_{\mathbf{x}} + \operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.10}$$

In view of the property $\operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, it follows immediately that

$$\begin{aligned} \operatorname{Var}(\hat{\mathbf{t}}^B_{\mathbf{y}}) &= \operatorname{Var}(\tilde{\mathbf{t}}_{\mathbf{y}}) - \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}\operatorname{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad - \operatorname{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\operatorname{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \operatorname{Var}(\hat{\mathbf{t}}^B_{\mathbf{x}}) &= \operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}}) - \operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\operatorname{Cov}'(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.11}$$

Remark 3.1. Every component or linear combination of components of $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ is BLUE for the corresponding total. Also, as evident from (3.11), the efficiency of $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$, relative to $\tilde{\mathbf{t}}_{\mathbf{y}}$, depends on the strength of correlation of $\mathbf{y}$ with $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, as well as on the difference in sample size (and possibly in sampling design) for the samples s_1 and s_2.

Remark 3.2. Because of the orthogonality property $\text{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, the coefficient of any of the terms $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ and $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ in (3.10) would not change if the other one would be set equal to $\mathbf{0}$ in (3.2). For instance, the BLUE for $\mathbf{t}_{\mathbf{y}}$ based on $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}})$ would be $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ as in (3.10) but without the last term. This is easily worked out as a special case of the full setup $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$. This orthogonality property explains the additive reduction of variance noticed in the first equation of (3.11).

Remark 3.3. The BLUE $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$ in (3.10) can also be produced using the reduced setup $(\hat{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ in (3.1). The same best linear estimator, for a single target variable, has been derived differently in the context of general single-phase sampling by Fuller and Isaki (1981) and Montanari (1987). In particular, for the auxiliary variable $\mathbf{x}_1$ we have $\hat{\mathbf{t}}_{\mathbf{x}_1}^{B} = \mathbf{t}_{\mathbf{x}_1}$. Next, it can be easily verified that the BLUE in (3.1) can be alternatively derived in two steps of best linear unbiased estimation using the setup $(\tilde{\mathbf{t}}_{\mathbf{y}}^{B}, \hat{\mathbf{t}}_{\mathbf{x}}^{B}, \tilde{\mathbf{t}}_{\mathbf{x}}^{B})$, where $\tilde{\mathbf{t}}_{\mathbf{y}}^{B} = \tilde{\mathbf{t}}_{\mathbf{y}} + \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ and $\tilde{\mathbf{t}}_{\mathbf{x}}^{B} = \tilde{\mathbf{t}}_{\mathbf{x}} + \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ are the BLUEs generated by the one-phase setups $(\tilde{\mathbf{t}}_{\mathbf{y}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ and $(\tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$, respectively. It can be shown through tedious algebra, that another, more explicit, BLUE setup that is equivalent to that in (3.1) is $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_2}, \tilde{\mathbf{t}}_{\mathbf{x}_2}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}, \mathbf{t}_{\mathbf{x}_1} - \tilde{\mathbf{t}}_{\mathbf{x}_1})$. This attests that the compact setup in (3.1) provides the most efficient linear estimation of $\mathbf{t}_{\mathbf{y}}$ and $\mathbf{t}_{\mathbf{x}}$ using all available relevant estimates.

3.2 The two-phase BLUE as calibration estimator

Using the notation leading to (3.7), and setting $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{B} = (\hat{\mathbf{t}}_{\mathbf{y}}^{B\prime}, \hat{\mathbf{t}}_{\mathbf{x}}^{B\prime})'$ and $\mathbf{\Delta} = \text{Var}(\mathbf{w}_U^{*})$, we may express the BLUE in (3.3) as $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{B} = \hat{\mathbf{t}}_{\mathbf{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\mathcal{X}} - \hat{\mathbf{t}}_{\mathcal{X}})$, where $\boldsymbol{\mathcal{B}} = \mathbf{\Psi}_U' \, \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U \, (\boldsymbol{\mathcal{X}}_U' \, \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}$ and $\mathbf{t}_{\mathcal{X}} = (\mathbf{0}', \mathbf{t}_{\mathbf{x}_1}')'$, or in the more suggestive form

$$\hat{\mathbf{t}}_{\mathbf{\Psi}}^{B} \;=\; \mathbf{\Psi}_U' \left[\mathbf{w}_U^{*} + \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U \, (\boldsymbol{\mathcal{X}}_U' \, \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \left(\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}_U' \, \mathbf{w}_U^{*}\right)\right]. \tag{3.12}$$

It appears from (3.12) that $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{B}$ has the form of a calibration estimator, with population vector of calibrated weigs $\mathbf{c}_U^{*} = \mathbf{w}_U^{*} + \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U \, (\boldsymbol{\mathcal{X}}_U' \, \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \left(\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}_U' \, \mathbf{w}_U^{*}\right)$ and vector of calibration totals $\mathbf{t}_{\mathcal{X}}$. This is formalized in the following theorem; the proof is in the Appendix.

Theorem 1. *The vector* $\mathbf{c}_U^{*} = \mathbf{w}_U^{*} + \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U \, (\boldsymbol{\mathcal{X}}_U' \, \mathbf{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \left(\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}_U' \, \mathbf{w}_U^{*}\right)$ *minimizes the generalized least-squares distance* $(\mathbf{c}_U^{*} - \mathbf{w}_U^{*})' \mathbf{\Delta}^{-1} (\mathbf{c}_U^{*} - \mathbf{w}_U^{*})$ *subject to the constraints* $\boldsymbol{\mathcal{X}}_U' \, \mathbf{c}_U^{*} = \mathbf{t}_{\mathcal{X}}$, *i.e.,* $\mathbf{X}_U' \, \mathbf{c}_{1U} = \mathbf{X}_U' \, \mathbf{c}_U$ *and* $\mathbf{X}_{1U}' \, \mathbf{c}_{1U} = \mathbf{t}_{\mathbf{x}_1}$, *where* $(\mathbf{c}_{1U}, \mathbf{c}_U)$ *corresponds to* $(\mathbf{w}_{1U}, \mathbf{w}_U)$.

Theorem 1 states that best linear unbiased estimation using the setup $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ is essentially a calibration procedure whereby the two estimates $\hat{\mathbf{t}}_{\mathbf{x}}$ and $\tilde{\mathbf{t}}_{\mathbf{x}}$ of $\mathbf{t}_{\mathbf{x}}$ are calibrated to each other, i.e., they are aligned, and the estimate $\hat{\mathbf{t}}_{\mathbf{x}_1}$ is calibrated to the total $\mathbf{t}_{\mathbf{x}_1}$. We may now write formally the BLUE $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{B}$

as a calibration estimator $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \boldsymbol{\Psi}_U' \mathbf{c}_U^{*}$, with its two components given in the simple linear forms $\hat{\mathbf{t}}_{\mathbf{y}}^{B} = \mathbf{Y}_U' \mathbf{c}_U$ and $\hat{\mathbf{t}}_{\mathbf{x}}^{B} = \mathbf{X}_U' \mathbf{c}_U$.

The alternative two-step construction of the BLUE noted in Remark 3.3 above can also be carried out through a two-step calibration procedure involving $\mathbf{w}_U^{*}$ in both steps. Indeed, partitioning $\mathcal{X}_U$ by its two column submatrices as $\mathcal{X}_U = (\mathcal{X}_{12U}, \mathcal{X}_{1U})$, and noting that $\mathcal{X}_{12U}' \boldsymbol{\Delta} \mathcal{X}_{1U} = \operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, it is easy to decompose the vector $\mathbf{c}_U^{*}$ as

$$\begin{aligned} \mathbf{c}_U^{*} &= \mathbf{w}_U^{*} + \boldsymbol{\Delta} \mathcal{X}_{12U} (\mathcal{X}_{12U}' \boldsymbol{\Delta} \mathcal{X}_{12U})^{-1} \left(\mathbf{0} - \mathcal{X}_{12U}' \mathbf{w}_U^{*}\right) \\ &\quad + \boldsymbol{\Delta} \mathcal{X}_{1U} (\mathcal{X}_{1U}' \boldsymbol{\Delta} \mathcal{X}_{1U})^{-1} \left(\mathbf{t}_{\mathbf{x}_1} - \mathcal{X}_{1U}' \mathbf{w}_U^{*}\right). \end{aligned} \tag{3.13}$$

In the rig hand side of (3.13), the sum of the first and second terms results from calibration with constraint $\mathcal{X}_{12U}' \mathbf{c}_U^{*} = \mathbf{X}_U' \mathbf{c}_{1U} - \mathbf{X}_U' \mathbf{c}_U = \mathbf{0}$ only, while the sum of the first and third terms results from calibration with constraint $\mathcal{X}_{1U}' \mathbf{c}_U^{*} = \mathbf{t}_{\mathbf{x}_1}$ only.

Now setting $\boldsymbol{\Delta}_1 = \operatorname{Var}(\mathbf{w}_{1U})$ and $\boldsymbol{\Delta}_2 = \operatorname{Var}(\mathbf{w}_U)$, these variances being specified by (3.8), it follows easily from (3.13) that the optimal calibration estimators $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ and $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$ in (3.10) can be written in the explicit form, which will be recalled later,

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{B} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U\right] \left[\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{B} &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.14}$$

4. Optimal linear estimation in two-phase sampling

4.1 The two-phase optimal estimator

The matrix $\mathcal{B}$ in (3.7) comprises variances and covariances which need to be estimated. In view of $\operatorname{Var}(\hat{\mathbf{t}}_{\mathcal{X}}) = \mathcal{X}_U' \boldsymbol{\Delta} \mathcal{X}_U$ and $\operatorname{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \mathcal{X}_U$, and recalling (3.8), the obvious unbiased estimates are $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathcal{X}}) = \mathcal{X}' \hat{\boldsymbol{\Delta}} \mathcal{X}$ and $\widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \mathcal{X}$, where the $(n_1 + n_2) \times (n_1 + n_2)$ matrix $\hat{\boldsymbol{\Delta}} = \widehat{\operatorname{Var}}(\mathbf{w}_U^{*})$ has diagonal blocks $\hat{\boldsymbol{\Delta}}_1 = \left\{(\pi_{1kl} - \pi_{1k}\pi_{1l}) / \pi_{1k}\pi_{1l}\pi_{1kl}\right\}$, $\hat{\boldsymbol{\Delta}}_2 = \left\{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}) / \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\pi_{1kl}\pi_{2kl}\right\}$, and off-diagonal blocks $\hat{\boldsymbol{\Delta}}_{12}$, $\hat{\boldsymbol{\Delta}}_{21} = \hat{\boldsymbol{\Delta}}_{12}'$ with $\hat{\boldsymbol{\Delta}}_{12} = \left\{(\pi_{1kl} - \pi_{1k}\pi_{1l}) / \pi_{1k}\pi_{1l}\pi_{1kl}\pi_{2l}\right\}$, and $\mathcal{X}$, $\boldsymbol{\Psi}$ are the sample matrices in (3.5).

We now obtain, as elements of the matrices $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathcal{X}})$ and $\widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}})$, the unbiased estimates of all variances and covariances in (3.9), i.e, $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$, $\widehat{\operatorname{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$, $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$, $\widehat{\operatorname{Cov}}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\operatorname{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$, $\widehat{\operatorname{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\operatorname{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$. However, the matrix $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathcal{X}})$ includes also the elements $\widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$, and $\widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1 - \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$, which clearly do not retain the properties $\operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ and $\operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, respectively. Unbiased estimates for the variances and covariances in (3.9) could be directly used, but then the estimate of the simple form $\mathcal{B}$ in (3.9) could not be expressed as $\boldsymbol{\Psi}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$, and thus the resulting estimator would not retain the calibration form of the BLUE in (3.12). This complication is circumvented using the following reformulation. Reset $\mathbf{w}^*$, $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\Psi}$ as

$$\mathbf{w}^* = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w} \\ \mathbf{w}_1 \end{pmatrix}, \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{11} \end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix} -\mathbf{Y}_1 & -\mathbf{X}_1 \\ \mathbf{Y}_2 & \mathbf{X}_2 \\ \mathbf{Y}_1 & \mathbf{X}_1 \end{pmatrix}, \tag{4.1}$$

where the sample matrices $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_{11}$ and $\mathbf{Y}_2$ are as before, and $\mathbf{Y}_1$ is the matrix of $\mathbf{y}$ for sample s_1 with dummy values $\mathbf{y}_k$ for $k \notin s_2$. Clearly, $\boldsymbol{\mathcal{X}}'\mathbf{w}^*$ and $\boldsymbol{\Psi}'\mathbf{w}^*$ are exactly as in (3.6). Then, having as before $\hat{\mathbf{t}}_{\mathcal{X}} = \boldsymbol{\mathcal{X}}'\mathbf{w}^*$ and $\hat{\mathbf{t}}_{\Psi} = \boldsymbol{\Psi}'\mathbf{w}^*$, we obtain again $\mathcal{B} = \mathrm{Cov}(\hat{\mathbf{t}}_{\Psi}, \hat{\mathbf{t}}_{\mathcal{X}})\left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}})\right]^{-1}$, where $\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\mathcal{X}}_U'\,\mathrm{Var}(\mathbf{w}_U^*)\boldsymbol{\mathcal{X}}_U$ and $\mathrm{Cov}(\hat{\mathbf{t}}_{\Psi}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}_U'\,\mathrm{Var}(\mathbf{w}_U^*)\,\boldsymbol{\mathcal{X}}$, as in (3.7) but with $\mathbf{w}_U^*$, $\boldsymbol{\mathcal{X}}_U$ and $\boldsymbol{\Psi}_U$ being the population counterparts of the redefined $\mathbf{w}^*$, $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$. An extension of Lemma 1 to the redefined $\mathbf{w}^*$ gives

$$\mathrm{Var}(\mathbf{w}_U^*) = \begin{pmatrix} \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_U) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \end{pmatrix},$$

where $\mathrm{Var}(\mathbf{w}_{1U})$ and $\mathrm{Var}(\mathbf{w}_U)$ are the same as in Lemma 1. It is easy now to verify that again $\mathcal{B}$ may be expressed analytically as in (3.9), and the two components of the BLUE are identical to those given by (3.10). More importantly, it follows from this special form of $\mathrm{Var}(\mathbf{w}_U^*)$ that we have again $\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\mathcal{X}}_U'\,\boldsymbol{\Delta}\boldsymbol{\mathcal{X}}_U$ and $\mathrm{Cov}(\hat{\mathbf{t}}_{\Psi}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}_U'\,\boldsymbol{\Delta}\boldsymbol{\mathcal{X}}_U$, where now $\boldsymbol{\Delta} = \mathrm{diag}(-\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_1)$ and $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ as already defined. Thus we obtain again the BLUE in the calibration form of (3.12), and the retained orthogonal decomposition of the vector of calibrated weigs in (3.13) leads readily to the expression (3.14). Now the orthogonality property $\boldsymbol{\mathcal{X}}_{12U}'\,\boldsymbol{\Delta}\boldsymbol{\mathcal{X}}_{1U} = \mathbf{0}$ is induced by the block-diagonal structure of the redefined $\boldsymbol{\mathcal{X}}_U$, rather than by the special structure of the initial matrix $\boldsymbol{\Delta}$ used in (3.12).

For the reconstructed BLUE we now have the unbiased estimates $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}$ and $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\Psi}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$ are the sample matrices in (4.1), and $\hat{\boldsymbol{\Delta}} = \mathrm{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$ with $\hat{\boldsymbol{\Delta}}_1$, $\hat{\boldsymbol{\Delta}}_2$ as defined at the beginning of the section. From these we rederive easily the unbiased estimates of the variances and covariances in (3.9), but two of the elements of the sample matrix $\boldsymbol{\Psi}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}$ which involve $\mathbf{Y}_1$, namely $\mathbf{Y}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$ and $\mathbf{Y}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}$, require special consideration. The dummy (unobserved) values $\mathbf{y}_k$ for $k \notin s_2$, necessary for expanding $\mathbf{Y}_1$ to the population matrix $\mathbf{Y}_U$ in the reconstructed BLUE, are set equal to zero, and the values $\mathbf{y}_k$ for $k \in s_2$ are then necessarily weiged by $1/\pi_{2k}$. Then $\mathbf{Y}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$ and $\mathbf{Y}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}$ reduce to $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_1$ and $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_{11}$, which are the unbiased estimates $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}})$ and $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$, respectively. The estimated $\mathcal{B}$ in (3.9) is now given by

$$\hat{\mathcal{B}} = \begin{bmatrix} \left[\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_1\right]\left[\mathbf{X}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1\right]^{-1} & \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_{11}\left[\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\right]^{-1} \\ \mathbf{I} & \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\left[\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\right]^{-1} \end{bmatrix}.$$

The BLUE $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\mathcal{X}} - \hat{\mathbf{t}}_{\mathcal{X}})$ with estimated $\boldsymbol{\mathcal{B}}$ will be called optimal linear unbiased estimator, optimal estimator in short, denoted by $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\mathcal{X}} - \hat{\mathbf{t}}_{\mathcal{X}}),$ with its two components given by

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{O} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1\right] \left[\mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{O} &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{4.2}$$

This is the sample version of the BLUEs in (3.14), with estimated coefficients. In particular, $\hat{\mathbf{t}}_{\mathbf{x}}^{O}$ is the customary single-phase optimal estimator of $\mathbf{t}_{\mathbf{x}}$ using $\mathbf{x}_1$ as auxiliary variable, and data from the full first-phase sample s_1; see Montanari (1987) and Rao (1994).

Remark 4.1. When n_2 is very close to n_1, the optimal estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ can be quite unstable because of the near singularity of the inverted matrix in the coefficient of $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}},$ and thus can become very inefficient; see, though, later Remark 6.1 on two-phase designs in which this is not an issue. Generally this is not a realistic setting in two-phase sampling, where n_2 is typically much smaller than n_1.

Following the construction of $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$ and $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$ as two of the estimates in $\hat{\boldsymbol{\mathcal{B}}},$ it transpires that these two bilinear forms can be written alternatively as $\breve{\mathbf{Y}}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$ and $\breve{\mathbf{Y}}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11},$ respectively, where $\breve{\mathbf{Y}}_1$ is a weiged version of $\mathbf{Y}_1$ in which $\breve{\mathbf{y}}_k = \mathbf{y}_k / \pi_{2k}$ if $k \in s_2$ and $\breve{\mathbf{y}}_k = 0$ if $k \notin s_2$. Then $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}' \mathbf{w}^* = \breve{\boldsymbol{\Psi}}' \mathbf{w}^*,$ where $\breve{\boldsymbol{\Psi}}$ is $\boldsymbol{\Psi}$ in (4.1) with $\breve{\mathbf{Y}}_1$ in place of $\mathbf{Y}_1,$ and $\hat{\boldsymbol{\mathcal{B}}}$ can be written compactly as $\hat{\boldsymbol{\mathcal{B}}} \equiv \breve{\boldsymbol{\Psi}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1},$ where $\hat{\boldsymbol{\Delta}} = \operatorname{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1).$ Henceforth, $\hat{\boldsymbol{\Delta}}$ will be meant to be the matrix $\operatorname{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1).$

As in Montanari (1987) and Rao (1994) for the single-phase optimal estimator, for large samples s_1 and s_2 the optimal estimator $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\mathcal{X}} - \hat{\mathbf{t}}_{\mathcal{X}})$ approximates the BLUE $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B},$ and thus it is approximately unbiased. Furthermore, the variance of $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$ approximates that of $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B},$ which works out easily to be $\operatorname{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B}) = \operatorname{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \operatorname{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}})\left[\operatorname{Var}(\hat{\mathbf{t}}_{\mathcal{X}})\right]^{-1} \operatorname{Cov}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}),$ i.e., the compact form of (3.11). Then, using the estimates $\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}})$ and $\widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}),$ derived earlier, we obtain the estimated approximate variance of $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$ as $\widehat{\operatorname{AV}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}) = \widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \widehat{\operatorname{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) \left[\widehat{\operatorname{Var}}(\hat{\mathbf{t}}_{\mathcal{X}})\right]^{-1} \widehat{\operatorname{Cov}}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}).$ From this we derive the computationally convenient expressions $\widehat{\operatorname{AV}}(\hat{\mathbf{t}}_{\mathbf{y}}^{O}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{Y}_2 - \breve{\boldsymbol{\Psi}}_1' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \breve{\boldsymbol{\Psi}}_1,$ where $\breve{\boldsymbol{\Psi}}_1$ is the first column submatrix of $\breve{\boldsymbol{\Psi}},$ and $\widehat{\operatorname{AV}}(\hat{\mathbf{t}}_{\mathbf{x}}^{O}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1.$

4.2 The two-phase optimal estimator as calibration estimator

The optimal estimator $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\mathcal{X}} - \hat{\mathbf{t}}_{\mathcal{X}}),$ with $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1},$ takes the form

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O} = \breve{\boldsymbol{\Psi}}' \left[\mathbf{w}^* + \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1} (\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}' \mathbf{w}^*)\right],$$

of a calibration estimator, with vector of calibration totals $\mathbf{t}_{\mathcal{X}}$ and sample vector of calibrated weigs $\mathbf{c}^* = \mathbf{w}^* + \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1} \left(\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}' \mathbf{w}^*\right)$ satisfying $\boldsymbol{\mathcal{X}}' \mathbf{c}^* = \mathbf{t}_{\mathcal{X}}.$ This is established formally by the following theorem; the proof is similar to that of Theorem 1, and is omitted.

Theorem 2. *The vector* $\mathbf{c}^* = \mathbf{w}^* + \hat{\mathbf{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\mathbf{\Delta}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)$ *minimizes the generalized least-squares distance* $(\mathbf{c}^* - \mathbf{w}^*)'\hat{\mathbf{\Delta}}^{-1}(\mathbf{c}^* - \mathbf{w}^*)$ *subject to the constraints* $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$*, i.e.,* $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_2'\mathbf{c}$ *and* $\mathbf{X}_{11}'\mathbf{c}_1 = \mathbf{t}_{\mathbf{x}_1}$*, where* $(\mathbf{c}_1, \mathbf{c})$ *corresponds to* $(\mathbf{w}_1, \mathbf{w})$.

The sample vector $\mathbf{c}^*$ admits the same orthogonal decomposition as its population counterpart $\mathbf{c}_U^*$ in (3.13). We may now write formally the optimal estimator $\hat{\mathbf{t}}_{\mathbf{\Psi}}^O$ as a calibration estimator $\hat{\mathbf{t}}_{\mathbf{\Psi}}^O = \breve{\mathbf{\Psi}}'\mathbf{c}^*$, which in view of $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$ is generated by the simultaneous calibration of the two estimates $\hat{\mathbf{t}}_{\mathbf{x}}$ and $\tilde{\mathbf{t}}_{\mathbf{x}}$ of $\mathbf{t}_{\mathbf{x}}$ to each other, and of the estimate $\hat{\mathbf{t}}_{\mathbf{x}_1}$ to the total $\mathbf{t}_{\mathbf{x}_1}$.

Now, in expanded form the vector $\mathbf{c}^*$ is

$$\mathbf{c}^* = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{w}_1 + \hat{\mathbf{\Delta}}_1\mathbf{X}_1[\mathbf{X}_2'\hat{\mathbf{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\mathbf{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ \mathbf{w} + \hat{\mathbf{\Delta}}_2\mathbf{X}_2[\mathbf{X}_2'\hat{\mathbf{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\mathbf{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ \mathbf{w}_1 + \hat{\mathbf{\Delta}}_1\mathbf{X}_{11}(\mathbf{X}_{11}'\hat{\mathbf{\Delta}}_1\mathbf{X}_{11})^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \end{pmatrix}. \tag{4.3}$$

Then, using the partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, where $\boldsymbol{\mathcal{X}}_{12}$ and $\boldsymbol{\mathcal{X}}_1$ are the two orthogonal column submatrices of $\boldsymbol{\mathcal{X}}$ shown in (4.1), the two constraints are written as $\boldsymbol{\mathcal{X}}_{12}'\mathbf{c}^* = \mathbf{X}_2'\mathbf{c}_2 - \mathbf{X}_1'\mathbf{c}_1 = \mathbf{0}$ and $\boldsymbol{\mathcal{X}}_1'\mathbf{c}^* = \mathbf{X}_{11}'\mathbf{c}_3 = \mathbf{t}_{\mathbf{x}_1}$. It also follows from (4.3) that $\hat{\mathbf{t}}_{\mathbf{\Psi}}^O = \breve{\mathbf{\Psi}}'\mathbf{c}^*$ implies (4.2). Regarding the two components of $\hat{\mathbf{t}}_{\mathbf{\Psi}}^O$ we observe that $\hat{\mathbf{t}}_{\mathbf{x}}^O = -\mathbf{X}_1'\mathbf{c}_1 + \mathbf{X}_2'\mathbf{c}_2 + \mathbf{X}_1'\mathbf{c}_3 = \mathbf{X}_1'\mathbf{c}_3$, and that

$$\hat{\mathbf{t}}_{\mathbf{y}}^O = \breve{\mathbf{Y}}_1'(\mathbf{c}_3 - \mathbf{c}_1) + \mathbf{Y}_2'\mathbf{c}_2 = \sum\nolimits_{s_2}\left[(c_{3k} - c_{1k})/\pi_{2k} + c_{2k}\right]\mathbf{y}_k.$$

The explicit expression of $\hat{\mathbf{t}}_{\mathbf{y}}^O$, in terms of sample units, is

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^O &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\sum\nolimits_{s_2}\sum\nolimits_{s_2}\hat{\Delta}_{2kl}\,\mathbf{y}_k\,\mathbf{x}_l' - \sum\nolimits_{s_2}\sum\nolimits_{s_1}\hat{\Delta}_{1kl}\,\breve{\mathbf{y}}_k\,\mathbf{x}_l'\right] \times \\ &\quad \left[\sum\nolimits_{s_2}\sum\nolimits_{s_2}\hat{\Delta}_{2kl}\,\mathbf{x}_k\,\mathbf{x}_l' - \sum\nolimits_{s_1}\sum\nolimits_{s_1}\hat{\Delta}_{1kl}\,\mathbf{x}_k\,\mathbf{x}_l'\right]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \left(\sum\nolimits_{s_2}\sum\nolimits_{s_1}\hat{\Delta}_{1kl}\,\breve{\mathbf{y}}_k\,\mathbf{x}_{1l}'\right)\left(\sum\nolimits_{s_1}\sum\nolimits_{s_1}\hat{\Delta}_{1kl}\,\mathbf{x}_{1k}\,\mathbf{x}_{1l}'\right)^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}), \end{aligned} \tag{4.4}$$

where $\hat{\Delta}_{1kl}$ and $\hat{\Delta}_{2kl}$ are the kl^{th} elements of $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$, respectively. Formula (4.4) is simplified in certain two-phase designs employed in important large scale surveys; examples of such surveys are presented in Hidiroglou and Särndal (1998) and Turmelle and Beaucage (2013). Specifically, this is the case when independent sampling (Poisson, or stratified Poisson) is used in one of the two phases, that is, when $\pi_{1kl} = \pi_{1k}\pi_{1l}$ or $\pi_{2kl} = \pi_{2k}\pi_{2l}$. The simplification is considerable in the case of independent sampling in both phases. Then, both $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$ are diagonal, with diagonal elements $\hat{\Delta}_{1kk} = (1/\pi_{1k})((1/\pi_{1k}) - 1)$ and $\hat{\Delta}_{2kk} = (1/\pi_{1k}\pi_{2k})((1/\pi_{1k}\pi_{2k}) - 1)$, respectively, and (4.4) involves only single summations. Other two-phase designs in which (4.4) involves single summations only, although $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$ are not diagonal, involve simple random sampling or stratified simple random sampling in either phase; for an example of a survey with such two-phase design see Hidiroglou (2001). In general, however, the optimal estimator may not be practical because it requires the use of first-phase and second-phase joint inclusion probabilities π_{1kl} and π_{2kl}, which are not known for some complex sampling designs. Even when these joint probabilities are known, but the matrices $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$ are not diagonal, the estimated coefficient $\hat{\boldsymbol{\mathcal{B}}}$ and, hence, the optimal estimator may be unstable in very small samples – especially if the dimension of the auxiliary vector $\mathbf{x}$ is large. These difficulties may be overcome, at some loss of optimality, by employing simple approximations of the variances and covariances in $\hat{\boldsymbol{\mathcal{B}}}$; for approximate variance estimates based only on first order inclusion probabilities see, for example, Haziza, Mecatti and Rao (2008) and references therein. A computationally very convenient approximation of $\hat{\boldsymbol{\mathcal{B}}}$ leading to a two-phase estimator that belongs to the class of generalized regression estimators is described in the next section.

5. A two-phase generalized regression estimator

A variant of $\hat{\boldsymbol{\mathcal{B}}} = \breve{\mathbf{\Psi}}'\hat{\mathbf{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\mathbf{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$ that is computationally efficient, but generally suboptimal, is the generalized regression (GREG) coefficient $\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}} = \mathbf{\Psi}'\mathbf{\Lambda}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}\boldsymbol{\mathcal{X}})^{-1}$, where $\mathbf{\Psi}$ is as in (4.1) and with $\mathbf{y}_k = 0$ if $k \notin s_2$, and $\mathbf{\Lambda}$ is the "weiging" matrix $\operatorname{diag}(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{\Lambda}_1)$, with $\mathbf{\Lambda}_1 = \operatorname{diag}\{w_{1k} / q_{1k}\}$ and $\mathbf{\Lambda}_2 = \operatorname{diag}\{w_k / q_{2k}\}$, and with q_{1k}, q_{2k} being positive constants. This gives the GREG estimator

$$\hat{\mathbf{t}}_{\mathbf{\Psi}}^{\mathrm{GR}} = \hat{\mathbf{t}}_{\mathbf{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}\mathbf{t}_{\boldsymbol{\mathcal{X}}} + (\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})'\mathbf{w}^{*}. \tag{5.1}$$

Note that $\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}$ is optimal in the sense of least squares, i.e., it minimizes the quadratic distance $(\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})'\,\mathbf{\Lambda}(\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime})$, involving the residuals $\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}\prime}$ in $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{\mathrm{GR}}$, whereas the coefficient $\hat{\boldsymbol{\mathcal{B}}}$ minimizes $(\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}})'\hat{\mathbf{\Delta}}(\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}})'$, the estimated approximate variance of the optimal estimator $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{O}$. In this sense $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{\mathrm{GR}}$ is an approximation to $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{O}$. The two components of $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{\mathrm{GR}}$, similar in structure to the components of $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{O}$ in (4.2), are

$$\begin{aligned}
\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2'\mathbf{\Lambda}_2\mathbf{X}_2 + \mathbf{Y}_1'\mathbf{\Lambda}_1\mathbf{X}_1\right]\left[\mathbf{X}_2'\mathbf{\Lambda}_2\mathbf{X}_2 + \mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1\right]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\
&\quad + \mathbf{Y}_1'\mathbf{\Lambda}_1\mathbf{X}_{11}\left[\mathbf{X}_{11}'\mathbf{\Lambda}_1\mathbf{X}_{11}\right]^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\
\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}} &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_{11}\left[\mathbf{X}_{11}'\mathbf{\Lambda}_1\mathbf{X}_{11}\right]^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}).
\end{aligned} \tag{5.2}$$

The GREG estimator $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}}$ is the standard single-phase GREG estimator based on s_1 and the auxiliary variable $\mathbf{x}_1$. The GREG estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$, with the two orthogonal regression terms shown in (5.2), is expressed explicitly in terms of sample units as

$$\begin{aligned}
\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\textstyle\sum_{s_2}(\Lambda_{1k} + \Lambda_{2k})\,\mathbf{y}_k\,\mathbf{x}_k'\right]\left[\textstyle\sum_{s_2}\Lambda_{2k}\,\mathbf{x}_k\,\mathbf{x}_k' + \sum_{s_1}\Lambda_{1k}\,\mathbf{x}_k\,\mathbf{x}_k'\right]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\
&\quad + \left(\textstyle\sum_{s_2}\Lambda_{1k}\,\mathbf{y}_k\,\mathbf{x}_{1k}'\right)\left(\textstyle\sum_{s_1}\Lambda_{1k}\,\mathbf{x}_{1k}\,\mathbf{x}_{1k}'\right)^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}),
\end{aligned}$$

where $\Lambda_{1k} = w_{1k} / q_{1k}$ and $\Lambda_{2k} = w_k / q_{2k}$ are the k^{th} element of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, respectively. The constants q_{ik} should be specified as $q_{ik} = n_i$, to account for the differential in the sample size of s_i; see Merkouris (2004) for a justification in the context of calibrating combined samples. An equivalent adjustment of the weigs in Λ_{1k} and Λ_{2k} can be made through the multiplication of w_{1k} in Λ_{1k} by $\phi = n_2 / n_1$. Values of q_{ik}

that convert the GREG estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ to the optimal estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ can be specified for two-phase sampling designs for which optimal estimation is possible, as in the similar context of matrix sampling (Merkouris, 2015). For the simple example involving Poisson sampling in both phases, this specification is $q_{1k} = \pi_{1k} \,/\, (1 - \pi_{1k})$ and $q_{2k} = \pi_{1k}\, \pi_{2k} \,/\, (1 - \pi_{1k}\, \pi_{2k})$, rendering $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ identical to $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$.

The vector of calibrated weigs associated with the GREG estimator $\hat{\mathbf{t}}_{\mathbf{\Psi}}^{\mathrm{GR}}$ is $\mathbf{c}^{\mathrm{GR}} = \mathbf{w}^* + \mathbf{\Lambda}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\mathcal{X}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)$. It has the same form as $\mathbf{c}^*$ in (4.3), but with $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ in place of $-\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$, and minimizes the generalized least-squares distance $(\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^*)'\mathbf{\Lambda}^{-1}(\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^*)$ subject to the constraints $\boldsymbol{\mathcal{X}}'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\mathcal{X}}$. The partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, defined after (4.3), allows the orthogonal decomposition of the vector $\mathbf{c}^*$

$$
\begin{aligned}
\mathbf{c}^{\mathrm{GR}} \;=\; & \mathbf{w}^* + \mathbf{\Lambda}\boldsymbol{\mathcal{X}}_{12}(\boldsymbol{\mathcal{X}}_{12}'\,\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}_{12}'\,\mathbf{w}^*) \\
& + \mathbf{\Lambda}\boldsymbol{\mathcal{X}}_1(\boldsymbol{\mathcal{X}}_1'\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}(\mathbf{t}_{\mathbf{x}_1} - \boldsymbol{\mathcal{X}}_1'\mathbf{w}^*).
\end{aligned}
\tag{5.3}
$$

In the rig hand side of (5.3), the sum of the first and second terms would result from calibration with constraint $\boldsymbol{\mathcal{X}}_{12}'\,\mathbf{c}^{\mathrm{GR}} = \mathbf{0}$ only, while the sum of the first and third terms would result from calibration with constraint $\boldsymbol{\mathcal{X}}_1'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\mathbf{x}_1}$ only. The practical implication of this is that the vector $\mathbf{c}^*$ could be formed by concatenating the weig vectors generated by two separate calibrations, i.e., calibration of $(\mathbf{w}_1', \mathbf{w}')'$ using $(-\mathbf{X}_1', \mathbf{X}_2')'$ followed by calibration of $\mathbf{w}_1$ using $\mathbf{X}_{11}$. However, the one-step calibration procedure generating $\mathbf{c}^{\mathrm{GR}}$ is more convenient.

On the basis of its Taylor linearization, the GREG estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ in (5.1) is approximately (for large samples) unbiased. Furthermore, denoting by $\mathbf{e}$ the matrix of sample residuals $\mathbf{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}'}$, the estimated approximate variance of $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} = \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}\mathbf{t}_{\mathcal{X}} + \mathbf{e}'\mathbf{w}^*$ is the estimated variance of $\mathbf{e}'\mathbf{w}^*$, i.e., $\widehat{\mathrm{AV}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) = \widehat{\mathrm{Var}}$ $(\mathbf{e}'\mathbf{w}^*) = \mathbf{e}'\hat{\mathbf{\Delta}}\mathbf{e}$, whereas the estimated variance of the HT estimator $\tilde{\mathbf{t}}_{\mathbf{y}}$ is $\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{Var}}(\mathbf{\Psi}_1'\,\mathbf{w}^*) = \mathbf{Y}_2'\,\hat{\mathbf{\Delta}}_2\,\mathbf{Y}_2$, with $\mathbf{\Psi}_1$ being the first column submatrix of $\mathbf{\Psi}$.

Now using the calibration form $\mathbf{\Psi}_1'\,\mathbf{c}^{\mathrm{GR}}$ of $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ and the orthogonal decomposition (5.3) of $\mathbf{c}^{\mathrm{GR}}$, we easily obtain the decomposition $\mathbf{e} = \mathbf{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\,\hat{\boldsymbol{\beta}}_{\mathbf{x}}' - \boldsymbol{\mathcal{X}}_1\,\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$, where $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \mathbf{\Psi}_1'\,\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_{12}\,(\boldsymbol{\mathcal{X}}_{12}'\,\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{x}_1} = \mathbf{\Psi}_1'\,\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_1\,(\boldsymbol{\mathcal{X}}_1'\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}$ are the coefficients of $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ and $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$, respectively. Note that $\mathbf{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\,\hat{\boldsymbol{\beta}}_{\mathbf{x}}'$ is the matrix of residuals in the GREG estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}}\,(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})$ resulting from calibration involving only $\boldsymbol{\mathcal{X}}_{12}$, and $\mathbf{\Psi}_1 - \boldsymbol{\mathcal{X}}_1\,\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$ is the matrix of residuals in the GREG estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}_1}\,(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ resulting from calibration involving only $\boldsymbol{\mathcal{X}}_1$. Then, using the orthogonality of $\boldsymbol{\mathcal{X}}_1$ and $\boldsymbol{\mathcal{X}}_{12}$, it is shown without difficulty that

$$
\widehat{\mathrm{AV}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{AV}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) + \widehat{\mathrm{AV}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}),
$$

which implies that the reduction of variance due to using the two auxiliary variables $\mathbf{x}_1$ and $\mathbf{x}$ in the regression (also calibration) procedure is additive. Thus, recalling Remark 3.2, the generalized regression estimator retains this additivity property of the BLUE of $\mathbf{t}_{\mathbf{y}}$.

6. Comparisons with existing methods

An earlier approach to optimal linear estimation in two-phase sampling designs, involving the standard type of auxiliary information considered in Sections 2 to 5, is described in Hidiroglou (2001). The formulation starts with postulating a regression form for the estimator of $\mathbf{t}_{\mathbf{y}}$, for univariate y, which is identical to the form of the estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ in the first line of (3.2), and then the two unknown coefficients are determined so as to minimize the variance of this estimator. In estimating the two coefficients, the identities $\operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \operatorname{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \operatorname{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ and $\operatorname{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ were ignored in the first and the second coefficient, respectively, and variances and covariances in both coefficients involving first-phase estimators were estimated using the second-phase sample only, thereby ignoring relevant information from the larger part of the first-phase sample. The resulting estimator was not shown to be a calibration estimator. In fact, this version of optimal estimator cannot be constructed as a calibration estimator. As a practicable variant of this, Hidiroglou (2001) considered a GREG estimator whose two coefficients (of $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ and $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$) can be justified either by assuming different regression models for each phase or by two successive calibrations. The same GREG estimator had been proposed earlier by Hidiroglou and Särndal (1998), but with no reference to optimal estimation. In the calibration approach of Hidiroglou and Särndal (1998), the estimator $\hat{\mathbf{t}}_{\mathbf{x}_1}$ is first calibrated to its total $\mathbf{t}_{\mathbf{x}_1}$, using s_1, and the GREG estimator $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{\mathrm{GR}}$ of $\mathbf{t}_{\mathbf{x}}$ is then generated using the calibrated weigs, denoted by $\tilde{w}_{1k}$. Then the overall weig for $k \in s_2$ is formed as $\tilde{w}_k = \tilde{w}_{1k}\, w_{2k}$. In a second calibration involving s_2 and $\tilde{w}_k$, the estimator $\tilde{\mathbf{t}}_{\mathbf{x}}$ is calibrated to $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{\mathrm{GR}}$. The resulting calibrated weigs of s_2 are then used to generate the GREG estimator of $\mathbf{t}_{\mathbf{y}}$, denoted here by $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{HS}}$.

Estevao and Särndal (2002, 2009) proposed a simpler version of the estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{HS}}$, in which the overall design weigs $w_k = w_{1k}\, w_{2k}$ for $k \in s_2$ are used in the second calibration. Using current notation, this estimator, denoted here by $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$, can be expressed in regression form as

$$\begin{aligned}\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}).\end{aligned} \tag{6.1}$$

Here the standard weiging matrices $\boldsymbol{\Lambda}_1 = \operatorname{diag}\{w_{1k}\}$ and $\boldsymbol{\Lambda}_2 = \operatorname{diag}\{w_k\}$ are used. Estevao and Särndal (2009) showed that this estimator is asymptotically equivalent to the estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{HS}}$. For the estimator $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$ in (6.1), Estevao and Särndal (2002) provide two linear regression representations corresponding to the two calibration steps. Replacing $\mathbf{y}$ by $\mathbf{x}$ in (6.1) gives $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{ES}}$, which is identical to $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}}$ in (5.2).

In comparison, the regression estimator proposed in Section 5 is motivated by the single-step calibration structure of the optimal two-phase estimator, of which it serves as practical approximation. It derives its statistical and computational efficiency, relative to competing regression estimators assessed in this section, from a single-step calibration procedure involving the combined first-phase and second-phase samples, and in which first-phase and second-phase estimated totals are calibrated to each other. As a consequence, the regression coefficients of the terms $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ and $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ incorporate information from the full sample s_1, as in the optimal estimator, and because of that they are more stable estimates of their population counterparts. An empirical comparison of the proposed regression estimator with the competing regression estimators is included in the simulation study in Section 7.

Replacing $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ by $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$ in (6.1) converts the coefficient of the GREG estimator $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}|\mathbf{x}_1}$ generated by the first step calibration into the coefficient $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})[\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1})]^{-1}$ of the single-phase optimal regression estimator $\hat{\mathbf{t}}^{O}_{\mathbf{x}|\mathbf{x}_1}$, and the coefficient of the GREG estimator $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$ generated by the second step calibration into the coefficient $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})[\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}})]^{-1}$. This latter coefficient may be viewed as pseudo-optimal since $\hat{\mathbf{t}}^{O}_{\mathbf{x}|\mathbf{x}_1}$ is treated as constant in the second step calibration, generating a pseudo-optimal estimator $\hat{\mathbf{t}}^{\mathrm{PSO}}_{\mathbf{y}}$. In turn, if in place of the sample matrices $\hat{\mathbf{\Delta}}_1$ and $\hat{\mathbf{\Delta}}_2$ in (6.1) we use the population matrices $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ we construct the pseudo-BLUE

$$\begin{aligned}\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{y}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}'_U \mathbf{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}'_U \mathbf{\Delta}_2 \mathbf{X}_U\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}'_U \mathbf{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}'_U \mathbf{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}'_U \mathbf{\Delta}_1 \mathbf{X}_{1U} \left(\mathbf{X}'_{1U} \mathbf{\Delta}_1 \mathbf{X}_{1U}\right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}),\end{aligned} \tag{6.2}$$

where the coefficients of $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ and $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ are, respectively, $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}$ and $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}$. Thus, the GREG estimator (6.1) may be viewed as an approximation of $\hat{\mathbf{t}}^{\mathrm{PSO}}_{\mathbf{y}}$, which is the estimator $\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{y}}$ with estimated coefficients (in analogy with the relationship of the optimal estimator $\hat{\mathbf{t}}^{O}_{\mathbf{y}}$ and the BLUE $\hat{\mathbf{t}}^{B}_{\mathbf{y}}$, in (4.2) and (3.14)). The pseudo-BLUE estimator $\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{x}}$, obtained from (6.2) by replacing $\mathbf{y}$ with $\mathbf{x}$, is identical to the BLUE $\hat{\mathbf{t}}^{B}_{\mathbf{x}}$, in (3.14). On the other hand, the estimators $\hat{\mathbf{t}}^{B}_{\mathbf{y}}$ and $\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{y}}$ are identical only under the condition of the following proposition; see proof in the Appendix.

Proposition 1. The estimators $\hat{\mathbf{t}}^{B}_{\mathbf{y}}$ and $\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{y}}$ are identical only if $\mathbf{\Delta}_1 = \delta \mathbf{\Delta}_2$, for a constant δ.

Remark 6.1. The condition of Proposition 1 holds if the same equal-inclusion probability design is used in both phases; the constant δ is then a function of the sample inclusion probabilities. Two-phase designs that satisfy this condition are SRS and Bernoulli in both phases, as well as their stratification versions with identical stratification and proportional sample allocation in both phases. The practical importance of this is that for these designs the sample counterparts of $\hat{\mathbf{t}}^{B}_{\mathbf{y}}$ and $\hat{\mathbf{t}}^{\mathrm{PSB}}_{\mathbf{y}}$, i.e., $\hat{\mathbf{t}}^{O}_{\mathbf{y}}$ and $\hat{\mathbf{t}}^{\mathrm{PSO}}_{\mathbf{y}}$, will be for large samples almost identical. Furthermore, $\mathbf{\Delta}_1 = \delta \mathbf{\Delta}_2$ implies that the minus sign in the coefficient of $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ in (3.14) and (4.2) could change to plus sign, with $1 - \delta$ factoring out, and thus the singularity problem identified in Remark 4.1 will not exist.

Remark 6.2. It is simple to verify that $\hat{\mathbf{t}}^{\mathrm{PSO}}_{\mathbf{y}}$ is a calibration estimator, constructed by a two-step calibration procedure (as with the $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$ estimator). Also, like $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$, the estimator $\hat{\mathbf{t}}^{\mathrm{PSO}}_{\mathbf{y}}$ is formed using the calibrated weigs of the second-phase sample only.

Finally, it should be mentioned that a "design-optimal regression estimator", not having the calibration property, has been proposed by Chen and Kim (2014) for a specific application and auxiliary variable setup. Also, a calibration estimator that is optimal under a model-assisted framework, and with $\mathbf{x}_2$ as the only auxiliary vector, has been proposed by Wu and Luan (2003).

7. Simulation study

We have conducted a simulation study to assess the performance of the proposed two-phase estimators of the total t_y, for scalar variables y, x_1 and x_2, and compare them with the competing regression estimators considered above. Distributions of these variables were specified as follows. The distribution of x_1 is the lognormal with mean and variance parameters $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$. The distribution of x_2 is specified by the linear model $x_2 = 5 + x_1 + \epsilon$, where $\epsilon \sim N(0, \sigma^2_\epsilon)$, and the distribution of y is specified by the linear model $y = 10 + 2x_1 + 3x_2 + \eta$, where $\eta \sim N(0, \sigma^2_\eta)$. The value of σ^2_ϵ determines the linear relationship between x_2 and x_1, as defined by the population square correlation coefficient $r^2_{x_1, x_2}$, and the value of σ^2_η determines the linear relationship of y with x_1 and x_2, as defined by the coefficient of determination $r^2 = [r^2_{y,x_1} + r^2_{y,x_2} - 2r_{y,x_1} r_{y,x_2} r_{x_1,x_2}] \big/ (1 - r^2_{x_1,x_2})$.

Three values, 0, 0.25, 0.75, were specified for $r^2_{x_1,x_2}$, and two values, 0.25 and 0.75, for r^2, giving six combinations of values $(r^2_{x_1,x_2}, r^2)$. For the value $r^2_{x_1,x_2} = 0$, in particular, the bivariate lognormal distribution for (x_1, x_2) with parameters $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$, $(\mu_{x_2} = 9, \sigma^2_{x_2} = 9)$ and zero correlation was used. The required values of σ^2_ϵ and σ^2_η are readily determined, while values for r^2_{y,x_1} and r^2_{y,x_2} are implicitly specified. For each of these six combinations, a population of size N = 50,000 was simulated by generating values from the distributions of the components of the vector (y, x_1, x_2). Four combinations of first-phase and second-phase sample sizes (n_1, n_2) with fixed n_1 and varying n_2 were specified, i.e., (3,000; 2,000), (3,000; 1,500), (3,000; 1,000), (3,000; 500), thus creating a total of 24 simulation settings.

Three different two-phase sampling designs were considered. Simple random sampling (SRS) without replacement was first used in both phases. For this sampling design, denoted by (SRS, SRS), the BLUE $\hat{t}^B_y$ in (3.10) and its exact variance in (3.11) can be calculated. Using the fact that under SRS the correlation of the HT estimators for two totals is identical to the correlation coefficient of the associated variables, tedious but straigforward algebra gives the relative difference (RDV) of variances of the estimators $\hat{t}^B_y$ and $\tilde{t}_y$ as

$$\frac{\mathrm{Var}(\tilde{t}_y) - \mathrm{Var}(\hat{t}^B_y)}{\mathrm{Var}(\tilde{t}_y)} = \frac{N(n_1 - n_2)}{n_1(N - n_2)} r^2 + \frac{n_2(N - n_1)}{n_1(N - n_2)} r^2_{y,x_1}.$$

The percent RDV is the measure of the efficiency of the BLUE $\hat{t}^B_y$ relative to the HT estimator $\tilde{t}_y$. This exact maximum efficiency will serve to measure the closeness of the approximation of the BLUE by the optimal estimator, for the different sample sizes, as well as the efficiency of the other competing estimators relative to the HT estimator. Notice that as n_2 tends to zero, the RDV tends to r^2, and as n_2 tends to n_1, the RDV tends to r^2_{y,x_1} (the efficiency of the BLUE based on s_1 and x_1). The second two-phase design, denoted by (STRSRS, SRS), was stratified simple random sampling (STRSRS) and SRS in the first and second phase, respectively. The simulated populations were stratified by the size of the variable y, with three strata of sizes N_1 = 30,000, N_2 = 15,000, N_3 = 5,000 and proportional allocation of the sample s_1 to the three strata – giving equal inclusion probabilities in each of the two phases. For this design too, the BLUE $\hat{t}^B_y$ and its exact variance can be calculated. The third two-phase design, denoted by

(SRS, PPSS), involved SRS in the first phase and probability proportional to size systematic sampling (PPSS) in the second phase, using as size measure the simple transformation $z_2 = 15 + 0.5x_2$ of the variable x_2; using x_2 as size would result in $\hat{t}_{x_2} = \tilde{t}_{x_2}$. In this case the BLUE $\hat{t}_y^B$ (and the optimal estimator $\hat{t}_y^O$) cannot be calculated, because of the unknown probabilities π_{2kl}. However, GREG estimators can be calculated.

For each of these three two-phase designs, and all the 24 simulation settings, sampling was repeated 30,000 times, and each time we computed the estimators $\tilde{t}_y$, $\hat{t}_y^O$, $\hat{t}_y^{\text{GR}}$, $\hat{t}_y^{\text{ES}}$ and $\hat{t}_y^{\text{HS}}$, to obtain their empirical bias and variance. In all these cases, the simulation showed that the bias of all estimators was negligible, even for the smaller subsample sizes n_2. Thus their comparison is based on their variances relative to the benchmark variance of the HT estimator $\tilde{t}_y$. Specifically, the efficiency of each of the competing estimators $\hat{t}_y^O$, $\hat{t}_y^{\text{GR}}$, $\hat{t}_y^{\text{ES}}$ and $\hat{t}_y^{\text{HS}}$ is assessed through the percent relative difference between its empirical variance and the empirical variance of the estimator $\tilde{t}_y$; for example, for $\hat{t}_y^{\text{GR}}$ the relative difference is $\left[\text{Var}(\tilde{t}_y) - \text{Var}(\hat{t}_y^{\text{GR}})\right] \big/ \text{Var}(\tilde{t}_y)$. The relative difference shows the reduction of the variance of the particular estimator relative to the variance of the basic estimator $\tilde{t}_y$.

In the (SRS, SRS) design, the exact efficiency of the BLUE $\hat{t}_y^B$, relative to the HT estimator $\tilde{t}_y$, increases as n_2 decreases and as we move to higher values of (r_{y,x_2}^2, r^2), confirming Remark 3.1; see column 2 of Table 7.1. It is also confirmed that the efficiency of $\hat{t}_y^B$ tends to r^2 as n_2 decreases, faster for higher r_{x_1,x_2}^2. This maximum efficiency is closely approximated by the empirical efficiency of $\hat{t}_y^O$, even for the smaller subsample sizes n_2; see column 3 of Table 7.1. For the (STRSRS, SRS) design the exact efficiency of $\hat{t}_y^B$ is shown in column 6 of Table 7.1, exhibiting a pattern similar to that in the (SRS, SRS). This efficiency is closely approximated by the empirical efficiency of $\hat{t}_y^O$; see column 7 of Table 7.1. In both (SRS, SRS) and (STRSRS, SRS) the approximation of $\hat{t}_y^B$ by $\hat{t}_y^O$ is a little weaker in some settings involving the largest value of n_2, for the reason given in Remark 4.1.

Although the estimator $\hat{t}_y^O$ can be calculated in the (SRS, SRS) and (STRSRS, SRS) designs, the performance of the more practical, and of general applicability, calibration (GREG) estimators $\hat{t}_y^{\text{GR}}$ and $\hat{t}_y^{\text{ES}}$ is of great interest. For (SRS, SRS), the empirical efficiencies of these estimators are shown in columns 4 and 5 of Table 7.1. The negative sign indicates loss of efficiency with respect to the HT estimator. The efficiency of $\hat{t}_y^{\text{GR}}$ approximates closely the efficiency of $\hat{t}_y^O$, except for the four settings specified by $r_{x_1,x_2}^2 = 0$, $r^2 = 0.25$, 0.75 and $n_2 = 2{,}000$; 1,500; in particular, when $n_2 = 2{,}000$ the estimator $\hat{t}_y^{\text{GR}}$ is a little less efficient than the estimator $\tilde{t}_y$. In contrast, the estimator $\hat{t}_y^{\text{ES}}$ is less efficient than the estimator $\tilde{t}_y$ in six settings, when $r_{x_1,x_2}^2 = 0$, $r^2 = 0.25$, 0.75 and $n_2 = 2{,}000$; 1,500; 1,000; substantially less efficient when $n_2 = 2{,}000$; 1,500. The highlig in columns 4 and 5 is that the estimator $\hat{t}_y^{\text{GR}}$ is much more efficient than the estimator $\hat{t}_y^{\text{ES}}$ in all settings, more so for higher values of n_2 and for the higher values of (r_{y,x_2}^2, r^2); this indicates that $\hat{t}_y^{\text{GR}}$ is more effective in using information from the complement of s_2 and in exploiting higher correlations of y with x_1 and x_2. The efficiency of the estimator $\hat{t}_y^{\text{HS}}$ was virtually identical with that of $\hat{t}_y^{\text{ES}}$, in all three designs, and hence is not reported in Table 7.1. For (STRSRS, SRS), the empirical efficiencies of the calibration estimators $\hat{t}_y^{\text{GR}}$ and $\hat{t}_y^{\text{ES}}$ are shown in columns 8 and 9 of Table 7.1. It should be noted that the correlations within the strata are much weaker than the correlations for the whole population (shown in Table 7.1). Also, the HT estimator $\tilde{t}_y$ is highly efficient because of the stratification, especially for the larger values of n_2. The estimator $\hat{t}_y^{\mathrm{GR}}$ is less efficient than the estimator $\tilde{t}_y$ in 3 of the 24 settings, involving $n_2 = 2{,}000$, while for the rest its efficiency increases greatly as n_2 decreases, approaching the efficiency of $\hat{t}_y^{O}$. The estimator $\hat{t}_y^{\mathrm{ES}}$ is less efficient than the estimator $\tilde{t}_y$ in 12 settings. The estimator $\hat{t}_y^{\mathrm{GR}}$ is much more efficient than the estimator $\hat{t}_y^{\mathrm{ES}}$ in all settings, more so for higher values of n_2 and as we move from $r^2 = 0.25$ to $r^2 = 0.75$, and considerably more than in the (SRS, SRS) design.

Table 7.1
Percent efficiency of $\hat{t}_y^{B}$, $\hat{t}_y^{O}$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ relative to $\tilde{t}_y$

	(SRS, SRS)				(STRSRS, SRS)				(SRS, PPSS)	
n_2	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$
$\sigma_\eta^2 = 292.41,\ r^2_{x_1,x_2} = 0.00,\ r^2_{y,x_1} = 0.04,\ r^2_{y,x_2} = 0.21,\ r^2 = 0.25$										
2,000	11.54	10.09	-1.66	-26.88	16.74	13.69	-23.49	-79.94	-5.51	-30.38
1,500	15.00	13.84	10.91	-13.61	20.01	17.80	7.22	-38.46	4.57	-20.69
1,000	18.41	17.68	17.79	-0.71	22.22	21.20	19.34	-11.76	9.69	-10.22
500	21.74	20.62	20.77	11.29	23.81	22.34	22.75	7.84	11.24	0.67
$\sigma_\eta^2 = 32.15,\ r^2_{x_1,x_2} = 0.00,\ r^2_{y,x_1} = 0.13,\ r^2_{y,x_2} = 0.62,\ r^2 = 0.75$										
2,000	34.35	31.21	-4.23	-80.22	52.31	48.02	-66.06	-232.15	-21.68	-107.41
1,500	44.84	42.34	33.09	-41.49	61.53	59.02	24.66	-108.30	15.95	-74.74
1,000	55.10	53.80	53.83	-2.25	67.58	65.48	59.17	-30.84	36.49	-39.03
500	65.16	63.87	63.90	34.31	71.85	70.53	70.41	27.83	45.55	2.00
$\sigma_\epsilon^2 = 12.11,\ \sigma_\eta^2 = 632.52,\ r^2_{x_1,x_2} = 0.25,\ r^2_{y,x_1} = 0.12,\ r^2_{y,x_2} = 0.24,\ r^2 = 0.25$										
2,000	16.70	16.89	16.69	9.88	17.14	15.56	2.08	-23.18	10.24	3.03
1,500	18.85	19.02	19.52	13.57	20.26	19.45	16.46	-3.21	13.83	7.08
1,000	20.97	20.79	20.52	16.50	22.38	21.07	21.04	6.84	13.03	8.04
500	23.04	22.28	21.67	19.64	23.91	22.84	23.41	16.40	11.57	9.38
$\sigma_\epsilon^2 = 12.11,\ \sigma_\eta^2 = 70.68,\ r^2_{x_1,x_2} = 0.25,\ r^2_{y,x_1} = 0.36,\ r^2_{y,x_2} = 0.71,\ r^2 = 0.75$										
2,000	49.70	48.33	46.89	25.33	53.11	50.78	20.11	-38.15	35.49	11.64
1,500	56.23	55.48	56.46	38.08	61.86	60.63	53.81	8.36	46.33	22.98
1,000	62.63	62.20	61.35	48.69	67.71	66.38	66.10	34.90	47.91	30.19
500	68.90	68.09	65.58	59.65	71.90	70.99	71.00	55.27	47.68	38.93
$\sigma_\epsilon^2 = 1.33,\ \sigma_\eta^2 = 340.40,\ r^2_{x_1,x_2} = 0.75,\ r^2_{y,x_1} = 0.22,\ r^2_{y,x_2} = 0.24,\ r^2 = 0.25$										
2,000	23.36	23.67	23.01	10.81	18.09	15.47	-6.07	-46.54	16.62	3.38
1,500	23.78	23.63	24.19	13.85	20.83	19.60	14.52	-17.17	19.77	7.95
1,000	24.20	23.83	23.15	16.97	22.68	21.67	21.58	0.86	17.04	10.02
500	24.61	23.54	22.24	19.92	24.01	22.39	22.98	13.24	14.52	11.77
$\sigma_\epsilon^2 = 1.33,\ \sigma_\eta^2 = 37.82,\ r^2_{x_1,x_2} = 0.75,\ r^2_{y,x_1} = 0.67,\ r^2_{y,x_2} = 0.72,\ r^2 = 0.75$										
2,000	69.84	67.98	65.10	26.96	60.26	56.75	32.34	-27.50	56.65	13.24
1,500	71.17	69.57	70.70	38.91	66.25	64.49	59.73	14.39	65.49	26.11
1,000	72.47	71.26	69.17	49.62	70.17	68.80	68.69	40.44	61.00	35.28
500	73.74	72.19	67.58	60.66	72.94	71.12	71.10	56.90	54.67	44.54

SRS = Simple random sampling; STRSRS = stratified simple random sampling; PPSS = probability proportional to size systematic.

For (SRS, PPSS), the empirical efficiencies of the calibration estimators $\hat{t}_y^{\mathrm{GR}}$ and $\hat{t}_y^{\mathrm{ES}}$ are shown in columns 10 and 11 of Table 7.1. The pattern of these efficiencies is very similar to that in the (SRS, SRS) design. This is particularly so for the efficiency of $\hat{t}_y^{\mathrm{GR}}$ relative to $\hat{t}_y^{\mathrm{ES}}$, which is not included in Table 7.1

but can be easily derived using the displayed efficiencies of $\hat{t}_y^{\mathrm{GR}}$ and $\hat{t}_y^{\mathrm{ES}}$ relative to $\tilde{t}_y$. The HT estimator $\tilde{t}_y$ itself is more efficient with this two-phase design, which explains why the efficiency of the two calibration estimators $\hat{t}_y^{\mathrm{GR}}$ and $\hat{t}_y^{\mathrm{ES}}$ relative to $\tilde{t}_y$ is somewhat lower than in the (SRS, SRS) and (STRSRS, SRS) designs.

The whole simulation study was repeated with the simulated population for the vector (y, x_1, x_2) generated from a trivariate lognormal distribution with the specified correlation structures. For all three designs (SRS, SRS), (SRS, PPSS) and (STRSRS, SRS), the results (not shown here) were very similar to those based on the linear model for y used above.

It is of interest to consider the setup of auxiliary variables in which the scalar variable x_1 is augmented to $(1, x_1)$, with known totals (N, t_{x_1}). Then in the (SRS, SRS) design, in which construction of the BLUE $\hat{t}_y^B$ and the optimal estimator $\hat{t}_y^O$ is feasible, using the complete setup $(1, x_1, x_2)$ in calibration gives the same $\hat{t}_y^B$ and practically the same $\hat{t}_y^B$ as when using (x_1, x_2). It would also convert the regression estimator $\hat{t}_y^{\mathrm{GR}}$ to $\hat{t}_y^O$ (using the same adjustment $1/\pi_{2k}$ of $\mathbf{Y}_1$ as in $\hat{t}_y^O$), and the regression estimator $\hat{t}_y^{\mathrm{ES}}$ estimator to the pseudo-optimal estimator $\hat{t}_y^{\mathrm{PSO}}$ (defined in Section 6). These properties are derived from known theory, see for example Merkouris (2004, 2015), more directly for $\hat{t}_y^{\mathrm{ES}}$ and the second regression term of $\hat{t}_y^{\mathrm{GR}}$ and the optimal $\hat{t}_y^O$, irrespective of any specific functional relationship of y with $(1, x_1, x_2)$. Then, the three sample-based estimators would show virtually identical empirical behavior. This follows from Proposition 1, which gives the condition (satisfied by specific designs, including (SRS, SRS)) under which the pseudo-optimal regression estimator $\hat{t}_y^{\mathrm{PSO}}$ is asymptotically equivalent to the proposed optimal estimator $\hat{t}_y^O$. Experimental calculations have confirmed this equivalence. In the (STRSRS, SRS) design too, using $(1, x_1, x_2)$ gives the same $\hat{t}_y^B$ and $\hat{t}_y^O$ as when using (x_1, x_2), and converts the $\hat{t}_y^{\mathrm{GR}}$ and $\hat{t}_y^{\mathrm{ES}}$ estimators to the $\hat{t}_y^O$ and $\hat{t}_y^{\mathrm{PSO}}$ estimators, respectively. However, by Proposition 1 the equivalence of the latter two estimators, and hence of $\hat{t}_y^{\mathrm{GR}}$ and $\hat{t}_y^{\mathrm{ES}}$, does not hold in this sampling design.

8. Discussion

The described method of optimal and regression estimation for two-phase sampling involves a single-step calibration of the weigs of the combined first-and-second phase samples. Thus, using a single set of calibrated weigs that incorporate all the available information from the two phases, a substantially improved estimate of the total of a target variable can be obtained, as shown by the simulation study. These weigs could be used to calculate other weiged statistics, including means, ratios, quantiles and regression coefficients. The framework of the method is general enough to encompass complex designs with multiple stages and different stratification at the two phases, as well as various types of auxiliary variables known at the population or sample level – ten different cases of auxiliary information are identified in Estevao and Särndal (2002). Furthermore, the method may be extended to multi-phase sampling designs through the appropriate calibration setup.

Estimation of a total for any domain (subpopulation) of interest, $U_d \subset U$, can be carried out readily using the calibrated weigs and summing the weiged sample values of the variable of interest over U_d. For the resulting domain estimator to be optimal linear estimator, the domain estimates of $\mathbf{t}_{\mathbf{y}}$, $\mathbf{t}_{\mathbf{x}_1}$ and $\mathbf{t}_{\mathbf{x}_2}$ need to be combined linearly, by carrying out optimal calibration at the domain level with domain calibration totals and with the appropriate modification of the matrix $\boldsymbol{\mathcal{X}}$. A number of calibration options, regarding the use of the available auxiliary information at the population, domain and two-phase sample levels, could be considered for the most efficient estimation of domain totals in any particular application. Related work in Merkouris (2010) would be helpful in this context.

The estimated approximate variances of the two-phase optimal estimator and the two-phase regression estimator, based on Taylor linearization, were given in Sections 4.1 and Section 5, respectively. For the two-phase regression estimator, replication methods of variance estimation, such as the jackknife method or the bootstrap method, could be alternatively applied, or would be the only option when first-phase or second-phase joint inclusion probabilities are not known. There is extensive literature on such replication methods for existing regression estimators in two-phase sampling. The single-step calibration feature of the proposed regression estimation method may be helpful in this direction; detailed study of this is beyond the scope of this paper.

Acknowledgements

The author thanks the Editor, the Associate Editor and the two referees for their comments and suggestions, which have led to a considerable improvement of the manuscript.

Appendix

Proof of Lemma 1

The symmetric matrix $\mathrm{Var}(\mathbf{w}_U^*)$ has the form of (3.8) but with $\mathrm{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ as off-diagonal block. The kl^{th} element of the matrix $\mathrm{Var}(\mathbf{w}_{1U})$ is

$$\mathrm{Cov}(w_{1U_k}, w_{1U_l}) = [E(I_{1k}\, I_{1l}) - E(I_{1k})\, E(I_{1l})] \,/\, \pi_{1k}\, \pi_{1l} = (\pi_{1kl} - \pi_{1k}\pi_{1l}) \,/\, \pi_{1k}\, \pi_{1l}\,.$$

The kl^{th} element of the matrix $\mathrm{Var}(\mathbf{w}_U)$ is

$$\begin{aligned}\mathrm{Cov}(w_{U_k}, w_{U_l}) &= [E(I_{1k}\, I_{2k}\, I_{1l}\, I_{2l}) - E(I_{1k}\, I_{2k})\, E(I_{1l}\, I_{2l})] \,/\, \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\ &= [E_1(I_{1k}\, I_{1l}\, E_2(I_{2k}\, I_{2l})) - E_1(I_{1k}\, E_2(I_{2k}))\, E_1(I_{1l}\, E_2(I_{2l}))] \,/\, \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\ &= [\pi_{1kl}\, \pi_{2kl} - \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}] \,/\, \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}\,,\end{aligned}$$

where E_1 and E_2 denote expectation under first and second phase of sampling, respectively. Using similar arguments it follows that the kl^{th} element of the matrix $\mathrm{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ is

$$\mathrm{Cov}(w_{1U_k}, w_{U_l}) = [E(I_{1k} I_{1l} I_{2l}) - E(I_{1k}) E(I_{1l} I_{2l})] / \pi_{1k} \pi_{1l} \pi_{2l} = (\pi_{1kl} - \pi_{1k} \pi_{1l}) / \pi_{1k} \pi_{1l}.$$

This shows that $\mathrm{Cov}(w_{1U_k}, w_{U_l}) = \mathrm{Cov}(w_{1U_k}, w_{1U_l})$ and thus $\mathrm{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \mathrm{Var}(\mathbf{w}_{1U})$, which completes the proof.

Proof of Theorem 1

Matrix $\boldsymbol{\Delta} = \mathrm{Var}(\mathbf{w}_U^*)$ is nonsingular if and only if $\mathrm{Var}(\mathbf{w}_U) - \mathrm{Var}(\mathbf{w}_{1U})$ is nonsingular. This follows from a general result on inverses of partitioned matrices (see Harville, 2008, page 98). But $\mathrm{Var}(\mathbf{w}_U) - \mathrm{Var}(\mathbf{w}_{1U}) = \mathrm{Var}(\mathbf{w}_{1U} - \mathbf{w}_U)$, because $\mathrm{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \mathrm{Var}(\mathbf{w}_{1U})$, and therefore $\mathrm{Var}(\mathbf{w}_U) -$ $\mathrm{Var}(\mathbf{w}_{1U})$ is nonsingular, being a variance-covariance matrix. Next, to find the vector $\mathbf{c}_U^*$ that minimizes $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*)$ subject to the constraints $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, consider the function $\mathbf{F} =$ $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\lambda}' \boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^*$ where $\boldsymbol{\lambda}$ is a vector of Langrange multipliers. We then get the system of equations

$$\begin{aligned} \frac{\partial \mathbf{F}}{\partial \mathbf{c}_U^*} \; = \; 2\boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\mathcal{X}}_U \boldsymbol{\lambda} \; &= \; \mathbf{0} \\ \boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* - \mathbf{t}_{\boldsymbol{\mathcal{X}}} \; &= \; \mathbf{0}. \end{aligned}$$

Multiplying the first equation by $\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta}$, using $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$ and solving for $\boldsymbol{\lambda}$ gives $\boldsymbol{\lambda} = 2(\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \right)$. Inserting this into the first equation and solving for $\mathbf{c}_U^*$ gives $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \right)$.

Proof of Proposition 1

Clearly, the coefficients of $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ in (3.14) and (6.2) are identical if $\boldsymbol{\Delta}_1 = \delta \boldsymbol{\Delta}_2$. Next, using the partition $\mathbf{X}_U = (\mathbf{X}_{1U}, \mathbf{X}_{2U})$, the coefficient of $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ in (6.2) is expressed as follows. First we obtain

$$\begin{aligned} \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \; &= \; \mathbf{Y}_U' \boldsymbol{\Delta}_2 (\mathbf{X}_{1U}, \mathbf{X}_{2U}) \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \end{pmatrix} \\ &= \; \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} \left[A_{11} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) \right] \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \left[A_{21} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) \right], \end{aligned}$$

where A_{11}, A_{12}, A_{21}, A_{22} are derived by algebra of inverses of partitioned matrices. In particular,

$$A_{11} = (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} - A_{12} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U}) (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$$

and $A_{21} = -A_{22} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U}) \times (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$. Then,

$$\begin{aligned} A_{11} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) \; &= \; (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} + A_{12} \mathbf{B} \\ A_{21} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22} (\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) \; &= \; A_{22} \mathbf{B}, \end{aligned}$$

where

$$\mathbf{B} = \mathbf{X}'_{2U}\,\mathbf{\Delta}_1\,\mathbf{X}_{1U} - \mathbf{X}'_{2U}\,\mathbf{\Delta}_2\,\mathbf{X}_{1U}\,(\mathbf{X}'_{1U}\,\mathbf{\Delta}_2\,\mathbf{X}_{1U})^{-1}(\mathbf{X}'_{1U}\,\mathbf{\Delta}_1\,\mathbf{X}_{1U}).$$

It is then easy to verify that if $\mathbf{\Delta}_1 = \delta\mathbf{\Delta}_2$, we have $(\mathbf{X}'_{1U}\,\mathbf{\Delta}_2\,\mathbf{X}_{1U})^{-1}\,\mathbf{X}'_{1U}\,\mathbf{\Delta}_1\,\mathbf{X}_{1U} = \delta\mathbf{I}$ and $\mathbf{B} = \mathbf{0}$. It follows that $\mathbf{Y}'_U\,\mathbf{\Delta}_2\,\mathbf{X}_U\,\left(\mathbf{X}'_U\,\mathbf{\Delta}_2\,\mathbf{X}_U\right)^{-1}\mathbf{X}'_U\,\mathbf{\Delta}_1\,\mathbf{X}_{1U} = \mathbf{Y}'_U\,\mathbf{\Delta}_1\,\mathbf{X}_{1U}$, and thus the coefficients of $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ in (3.14) and (6.2) are also identical.

References

Australian Bureau of Statistics (2004). Estimation for the household income and expenditure survey. Research paper 1352.0.55.063.

Beaumont, J.-F., Beliveau, A. and Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

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

Chen, S., and Kim, J.K. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples. *The Annals of Applied Statistics*, 3, 1492-1515.

Chipperfield, J.O., and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 227-244.

Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., and Särndal, C.-E. (2009). A new face on two-phase sampling with calibration estimators. *Survey Methodology*, 35, 1, 3-14. Paper available at tps://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10880-eng.pdf.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 2, 167-180. Paper available at tps://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14537-eng.pdf.

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

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

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.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *METRON-International Journal of Statistics*, vol LXVI, 91-108.

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

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 1, 11-20. Paper available at tps://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998001/article/3905-eng.pdf.

Hidiroglou, M.A., Rao, J.N.K. and Haziza, D. (2008). Variance estimation in two-phase sampling. *Australian and New Zealand Journal of Statistics*, 51, 127-141.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.

Kim, J.K., and Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kim, J.K., and Yu, C.L. (2011). Replication variance estimation under two-phase sampling. *Survey Methodology*, 37, 1, 67-74. Paper available at tps://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11448-eng.pdf.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 311-320.

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

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

Merkouris, T. (2015). An efficient estimation method for matrix survey sampling. *Survey Methodology*, 41, 1, 237-262. Paper available at tps://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14174-eng.pdf.

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

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

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

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

Turmelle, C., and Beaucage, Y. (2013). The integrated business statistics program: Using a two-phase design to produce reliable estimates. *Proceedings: Symposium 2013, Producing reliable estimates from imperfect frames.*

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Wu, C., and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 2, 119-131.

Bayesian spatial models for estimating means of sampled and non-sampled small areas

Hee Cheol Chung and Gauri S. Datta[1]

Abstract

In many applications, the population means of geographically adjacent small areas exhibit a spatial variation. If available auxiliary variables do not adequately account for the spatial pattern, the residual variation will be included in the random effects. As a result, the independent and identical distribution assumption on random effects of the Fay-Herriot model will fail. Furthermore, limited resources often prevent numerous sub-populations from being included in the sample, resulting in non-sampled small areas. The problem can be exacerbated for predicting means of non-sampled small areas using the above Fay-Herriot model as the predictions will be made based solely on the auxiliary variables. To address such inadequacy, we consider Bayesian spatial random-effect models that can accommodate multiple non-sampled areas. Under mild conditions, we establish the propriety of the posterior distributions for various spatial models for a useful class of improper prior densities on model parameters. The effectiveness of these spatial models is assessed based on simulated and real data. Specifically, we examine predictions of statewide four-person family median incomes based on the 1990 Current Population Survey and the 1980 Census for the United States of America.

Key Words: Conditional autoregression; Current Population Survey; Fay-Herriot model; Intrinsic autoregression; Mixed effects; Simultaneous autoregression; non-sampled small areas.

1. Introduction

Sample surveys provide useful data in estimating various characteristics of a population of interest. Surveys are generally designed so that design-based estimators have adequate accuracy. However, when it comes to estimating a sub-population characteristic, a design-based direct estimate, based solely on data from that sub-population alone, is usually inaccurate as the accessible sample size is small and sometimes nonexistent. Sub-populations that lack a reasonable sample size to produce reliable direct estimates are known as small areas. Also, limited resources often preclude many sub-populations from selecting in the sample, creating non-sampled small areas. For example, the American Community Survey (ACS) is conducted to produce reliable statistics for the U.S. counties. However, the ACS usually samples about one-third of the counties resulting in many non-sampled small areas.

To enhance the accuracy of direct estimates of small areas, a model-based approach has been widely used to facilitate borrowing information from direct estimates of other domains and other auxiliary data. In many applications, supplementary information from other surveys and administrative data provide useful covariates. A model-based estimate of an area is produced by suitably shrinking its direct estimate (if available) to a synthetic regression estimate based on auxiliary variables. The improvement in prediction greatly depends on to what extent the sub-population means of the characteristic are related to the auxiliary variables. If a small area has no direct estimate, the traditional independent random-effects model of Fay and Herriot (1979) estimates estimates the mean by a synthetic regression estimate alone.

[1] Hee Cheol Chung, Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC. E-mail: hchung13@uncc.edu; Gauri S. Datta, Department of Statistics, University of Georgia, Athens, GA and Center for Statistical Research and Methodology, U.S. Census Bureau, Suitland, MD.

Fay and Herriot (1979) proposed a useful model for developing estimates of small area means based on direct survey estimates (if available) and computed synthetic regression estimates from auxiliary variables. This model, which is essentially a mixed linear model, is popularly known as the Fay-Herriot (FH) model in small area estimation. For $i = 1, \ldots, m,$ let Y_i be the direct estimate of the small area characteristic θ_i obtained from a survey. Also let $\mathbf{x}_i$ and $\boldsymbol{\beta}$ be the p-component vectors of covariates and corresponding regression coefficients, respectively. Denoting the sampling error of Y_i as $e_i,$ the independent FH model can be written as

$$Y_i = \theta_i + e_i, \quad \theta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \tag{1.1}$$

where e_i's and random effects v_i's $i = 1, \ldots, m,$ are all independently distributed with $e_i \sim N(0, D_i),$ and $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2).$ Sampling variances $D_i,$ $i = 1, \ldots, m$ are taken as known, whereas the regression parameter β and model error variance $\sigma_v^2,$ called model parameters, are unknown quantities. For non-sampled areas with auxiliary variables, only the second part of (1.1) holds for $\theta_i.$

There has been extensive research on the independent FH model and its many variants. While Fay and Herriot (1979) used an empirical Bayes (EB) approach, subsequently, Prasad and Rao (1990), Datta and Lahiri (2000) and Datta, Rao and Smith (2005) used the frequentist approach and derived the second-order mean squared error (MSE) of empirical best linear unbiased predictor (EBLUP) of θ_i and various second-order approximate unbiased estimators of the MSE's (see Datta and Lahiri, 2000). However, Ghosh (1992) proposed a hierarchical Bayesian (HB) approach for the Fay-Herriot model (see also Datta et al. (2005)). In the Bayesian framework, the FH model in (1.1) can be expressed as the following HB model:

$$Y_i \,|\, \theta_1, \ldots, \theta_m, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\theta_i, D_i), \quad i = 1, \ldots, m, \tag{1.2}$$

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma_v^2), \quad i = 1, \ldots, m, \tag{1.3}$$

$$\pi(\boldsymbol{\beta}, \sigma_v^2) \propto g(\boldsymbol{\beta}, \sigma_v^2), \tag{1.4}$$

where $g(\cdot)$ is a suitably chosen function of $\boldsymbol{\beta}$ and $\sigma_v^2,$ which expresses a prior probability density function (pdf) for these parameter. An EB predictor for $\theta_i,$ which does not require a prior pdf as in (1.4), was originally developed by Fay and Herriot (1979). While a standard EB approach usually underestimates the measure of uncertainty of the EB estimator of $\theta_i,$ the HB approach facilitates quantification of uncertainty due to estimation of unknown model parameters, $\boldsymbol{\beta}$ and $\sigma_v^2.$ The uncertainty is fully captured by the posterior distribution of the model parameters.

In model-based estimations, random effects are of great importance in capturing the remaining variability of the θ_i's that is not explained by the regression model. In real applications, small areas generally involve features such as population size, ethnicity, age-group, and education level, which might affect the variability of small area effects. Furthermore, when disease prevalence rates are of interest, it is reasonable to assume that random effects of adjacent small areas are correlated in a certain way. In such cases, the FH model given in (1.1), which we refer to as the independent FH random-effects model, oversimplifies and misspecifies the distribution of random effects by assuming a common and independent distribution. Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008) and Rao, Sinha and Dumitrescu (2014) proposed nonparametric small area estimation models, which capture spatial proximity effect using the P-spline function. However, these approaches require additional computational cost for model inference and uncertainty quantification.

In this work, we propose spatial FH models which effectively account for heteroscedasticity and spatial dependence of the small area effects. We take a fully Bayesian approach by specifying a class of noninformative priors on the model parameters and model spatial dependence of small area random effects by four widely used autocorrelation structures. These include simultaneous autoregressive and three types of conditional autoregressive models. There is an abundance of literature on spatial models under the Bayesian framework. Sun, Tsutakawa and Speckman (1999) studied an HB model with the conditional and intrinsic autoregressive models on the random effects. The same models were considered by Speckman and Sun (2003) in the context of Bayesian spline smoothing. For small area estimation, You and Zhou (2011) modeled small area effects using a conditional autoregressive model. As an extension of the time series FH model (Datta, Lahiri, Maiti and Lu, 1999), Torabi (2012) proposed a spatio-temporal model with intrinsic autoregressive random effects. Porter, Holan, Wikle and Cressie (2014) proposed an extension of the FH model with functional covariates and intrinsic autoregressive random effects. Porter, Wikle and Holan (2015) incorporated the conditional autoregressive random effects on the multivariate FH model.

The existing Bayesian spatial small area estimation models consider a proper prior on σ_v^2 even though the specification of a proper prior will require subject matter expertise. Furthermore, all existing models assume a conditional autoregressive structure on the random effects. The main contributions of this paper are as follows. First, to the best of our knowledge, the proposed models in Section 2 (Section 2.1) include most of the popularly used spatial structures. Second, in Section 2.2, we further extend the spatial models to estimate means of several non-sampled small areas with no direct estimates. The non-sampled area mean θ_i is estimated by borrowing strength from the auxiliary variables of the area and, for spatial models, from the regression residuals of its neighboring areas. Third, for all proposed models, we provide, in Section 2.3, sufficient conditions for posterior propriety for a class of improper noninformative priors on model parameters. Interestingly, the sufficient conditions do not depend on the assumed spatial model, provided that the model yields a positive definite covariance matrix for the random effects. We provide rejection sampling steps for simulating from the posterior of the proposed models in Section 3. The effectiveness of the proposed spatial models is demonstrated in Sections 4 and 5. We apply the spatial models to simulated datasets and real survey data from the Current Population Survey (CPS). We compare various spatial models in Section 5 to estimate four-person family median incomes for the forty-nine contiguous states of the U.S. based on the CPS data and appropriate covariates from the previous Census and administrative data. Our data analysis and simulation studies reveal that proposed spatial models significantly improve prediction accuracy and reduce the measure of uncertainty, posterior standard deviation. We provide concluding remarks in Section 6. All technical details are provided in Appendix.

2. Some spatial alternatives to the independent FH model

2.1 Incorporating spatial random effects

Let $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\top}$ be the m-component vector with the direct estimates of m small areas, and $\mathbf{D} = \operatorname{diag}\{D_i\}_{i=1}^{m}$ be the $m \times m$ diagonal matrix with the sampling variances of the direct estimates. We denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\top}$ the m-component vector of small area means. Also, let $\mathbf{x}_i \in \mathbb{R}^p$ be the p-component vector of auxiliary variables (including the intercept term) for the i^{th} small area, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^{\top}$. A special case of the HB model given in (1.2)-(1.4) can be expressed as

$$\mathbf{Y} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2 \sim N_m(\boldsymbol{\theta}, \mathbf{D}), \tag{2.1}$$

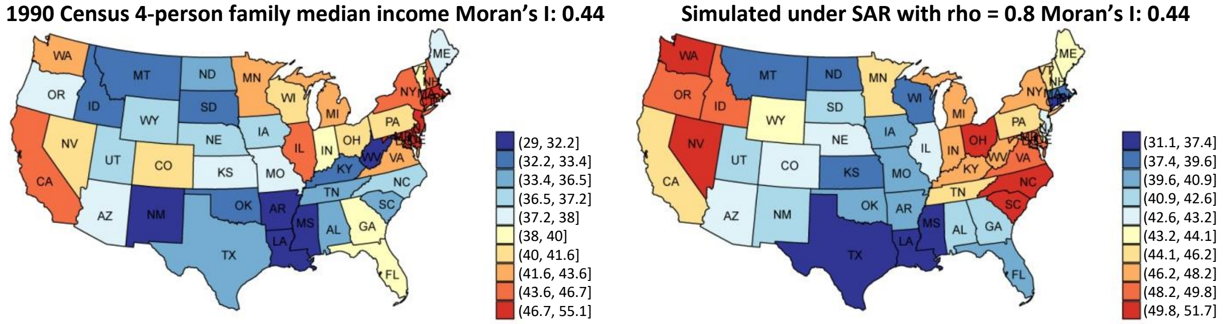
$$\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma_v^2 \sim N_m(\mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \mathbf{I}_m), \tag{2.2}$$

$$\pi(\boldsymbol{\beta}, \sigma_v^2) \propto 1, \tag{2.3}$$

where $\boldsymbol{\beta}$ is the p-component regression coefficient vector, σ_v^2 is the model error variance, and $\mathbf{I}_m$ is the identity matrix of order m. The uniform prior (2.3) on the model parameters is a popularly used noninformative prior, and the resulting posterior pdf is proper provided that $m > p + 2.$ See Berger (1985) and Datta and Smith (2003) for detailed discussion.

The model (2.2) assumes that $\theta_i,\ i = 1, \ldots, m,$ are independently distributed over the small areas with common random effects variance σ_v^2. In many small area estimation problems, however, the area characteristic of interest is closely related to geographical factors such as population size, ethnicity, age-group and education level. When available covariates do not fully explain such spatial association, the independence and equal variance assumptions of random effects fail, and inference based on the hierarchical model (2.1)-(2.3) may generate unreliable estimates, consequently resulting in erroneous decisions. Figure 2.1 illustrates spatial associations of the 1990 Census 4-person family median incomes of (scaled by \$1,000) the $m = 49$ states of the U.S., including the District of Columbia. Simulated data with the same Moran's I value are also displayed for comparison, where Moran's I is a measure for spatial autocorrelation. The simulated data are generated under the SAR model (defined below) with $\rho = 0.8$ matching the location and scale with the Census data. The two panels demonstrate the existence of spatial dependence in the 1990 Census 4-person family median incomes. In practice, covariates capable of fully capturing existing spatial variation are not always available, and the problem can be exacerbated if lurking variables exist, as they introduce additional variability that cannot be explained by the independently and identically distributed (i.i.d.) random effects.

Figure 2.1 Graphical illustrations of the 1990 Census 4-person family median incomes of the $m =$ 49 states of U.S. and simulated data.



1990 Census 4-person family median income Moran's I: 0.44

Simulated under SAR with rho = 0.8 Moran's I: 0.44

- The simulated data are generated under the SAR model (defined below) with $\rho =$ 0.8 and have the same Moran's I estimate.
- AL = Alabama, AZ = Arizona, AR = Arkansas, CA = California, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District of Columbia, FL = Florida, GA = Georgia, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiana, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = New Mexico, NY = New York, NC = North Carolina, ND = North Dakota, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvania, RI = Rhode Island, SC = South Carolina, SD = South Dakota, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginia, WA = Washington, WV = West Virginia, WI = Wisconsin, WY = Wyoming.

To address this issue, we propose to use spatially correlated random effects. Let $\mathbf{W} = \{w_{ij}\}_{ij}$, $1 \le i,\, j \le m,$ be the adjacency matrix which plays an important role in capturing spatial dependency. In particular, $w_{ij} = 1$ if the i^{th} and j^{th} small areas are connected via geographical boundary or through other mechanisms (for example, air traffic), and $w_{ij} = 0,$ otherwise. Also, $w_{ii} = 0$ for $i = 1, \ldots, m.$ The off-diagonal entries, w_{ij}'s, need not be binary; they can take other positive values, such as the "length" of the geographical border or volumes of air traffic between the two areas. Since the adjacency matrix $\mathbf{W}$ is symmetric, its eigenvalues are real. We denote the i^{th} largest eigenvalue of $\mathbf{W}$ by $\lambda_i(\mathbf{W}),$ such that $\lambda_m(\mathbf{W}) \le \ldots \le \lambda_1(\mathbf{W}).$ Since $\mathbf{W}$ is non-null and $\sum_{i=1}^m w_{ii} = 0,$ we get as a result that $\lambda_m(\mathbf{W}) < 0 < \lambda_1(\mathbf{W}).$ Let $w_{i.} = \sum_{j=1}^m w_{ij}$ be the sum of the i^{th} row of $\mathbf{W}$ and $\mathbf{L} = \operatorname{diag}\{w_{i.}\}_{i=1}^m.$ Assuming that diagonal elements of $\mathbf{L}$ are positive, i.e., all small areas have at least 1 neighboring area, we define $\tilde{\mathbf{W}} = \mathbf{L}^{-1}\mathbf{W}.$ Since $\tilde{\mathbf{W}}$ is a row stochastic matrix, all of its eigenvalues are between -1 and 1, with at least one of them is 1. Consequently, $\lambda_1(\tilde{\mathbf{W}}) = 1.$ Moreover, $\tilde{\mathbf{W}}$ and $\operatorname{diag}\{w_{i.}^{-1/2}\}_{i=1}^m \mathbf{W} \operatorname{diag}\{w_{i.}^{-1/2}\}_{i=1}^m$ have the same set of eigenvalues, and the latter matrix is symmetric. So all the eigenvalues of $\tilde{\mathbf{W}}$ are real, and $\lambda_m(\tilde{\mathbf{W}})$ will be negative. We consider four alternative spatial dependencies associated with random effects, which are represented by the following positive definite precision matrices (excluding the scale parameter σ_v^2):

$$\text{SAR}: \quad \mathbf{\Omega}_2(\rho) = (\mathbf{I}_m - \rho\tilde{\mathbf{W}})^{\top}(\mathbf{I}_m - \rho\tilde{\mathbf{W}}), \quad \rho \in (-1, 1), \tag{2.4}$$

$$\text{SCAR}: \quad \mathbf{\Omega}_3(\rho) = \mathbf{I}_m - \rho\mathbf{W}, \quad \rho \in (\lambda_m(\mathbf{W})^{-1}, \lambda_1(\mathbf{W})^{-1}), \tag{2.5}$$

$$\text{CAR}: \quad \mathbf{\Omega}_4(\rho) = \mathbf{L} - \rho\mathbf{W}, \quad \rho \in (\lambda_m(\tilde{\mathbf{W}})^{-1}, \lambda_1(\tilde{\mathbf{W}})^{-1}), \tag{2.6}$$

$$\text{LCAR}: \quad \mathbf{\Omega}_5(\rho) = \rho\mathbf{R} + (1-\rho)\,\mathbf{I}_m, \quad \rho \in (0, 1), \tag{2.7}$$

where ρ is the spatial dependence parameter that represents the strength of spatial dependence (Hodges, 2019, Chapter 5.2) and $\mathbf{R}$ is defined as $\mathbf{R} = \mathbf{\Omega}_4(1) = \mathbf{L} - \mathbf{W}.$ Since the eigenvalues of $\mathbf{I}_m - \tilde{\mathbf{W}}$ are between 0 (the smallest eigenvalue) and $1-\lambda_m(\tilde{\mathbf{W}})$ (the largest eigenvalue, >1), the matrix $\mathbf{R}$ is nonnegative definite. Each precision matrix is guaranteed to be positive definite as long as ρ is in the range specified in the respective definition.

The adjacency matrix $\tilde{\mathbf{W}}$ of the simultaneous autoregressive (SAR) model (Whittle, 1954) is row-normalized so that ρ can vary from -1 to 1 while preserving the positive definiteness (Banerjee, Carlin and Gelfand, 2003, Chapter 4.4). The model (2.5) is a simple version of conditional autoregressive (CAR) model (Rao and Molina, 2015, Chapter 9.6.2), where diagonal entries of the precision matrix are all equal to one. Even though the diagonal elements of a precision matrix are all equal, the diagonal elements of the inverse may not be all equal, leading to heteroscedasticity of random effects. We call this model the simple conditional autoregressive (SCAR) model. The model (2.6) is widely used conditional autoregressive model (CAR; Banerjee et al., 2003; Besag and Kooperberg, 1995; You and Zhou, 2011), where diagonal entries of the precision matrix are the number of neighborhoods of the corresponding area. The upper limit of ρ is $\lambda_1(\tilde{\mathbf{W}})^{-1}=1,$ and in the case of $\rho=1,$ the model with $\mathbf{\Omega}_4(1)$ is referred to as the intrinsic autoregressive (IAR) model (Banerjee et al., 2003, Chapter 4.3). The model (2.7) is a conditional autoregressive model, which we call Leroux's conditional autoregressive (LCAR), whose precision matrix is given by the convex combination of $\mathbf{R}=\mathbf{\Omega}_4(1)$ and $\mathbf{I}_m.$ This model has been considered by Leroux, Lei and Breslow (2000); MacNab (2003); You and Zhou (2011), where the i^{th} diagonal element of $\mathbf{R}$ is the number of neighborhoods of the i^{th} small area, and the $(i,j)^{\text{th}}$ off-diagonal element is -1 if the i^{th} and the j^{th} small areas are connected and 0 otherwise.

The conditional autoregressive models, SCAR, CAR, and LCAR, assume that θ_i depends only on neighboring small area means. In other words, θ_i is correlated with θ_j's, $j\neq i,$ only through the means of surrounding areas. On the contrary, the SAR model assumes that θ_i is dependent on all other θ_j concurrently, $j\neq i,$ but has stronger (weaker) correlations for neighboring (remote) areas. The independent FH model can be viewed as a special case of the SAR, SCAR, or LCAR model with $\rho=0.$ For notational convenience, we include the independent FH model as part of our model by taking its precision matrix $\mathbf{\Omega}_1(\rho)=\mathbf{I}_m,$ although it is free from $\rho.$

We consider the following HB spatial models incorporating the five spatial dependencies defined in (2.4)-(2.7):

$$\mathbf{Y}\,\big|\,\boldsymbol{\theta},\boldsymbol{\beta},\sigma_v^2,\rho \sim N_m(\boldsymbol{\theta},\mathbf{D}), \tag{2.8}$$

$$\boldsymbol{\theta}\,\big|\,\boldsymbol{\beta},\sigma_v^2,\rho \sim N_m(\mathbf{X}\boldsymbol{\beta},\sigma_v^2\{\mathbf{\Omega}_k(\rho)\}^{-1}),\quad k=1,\ldots,5, \tag{2.9}$$

$$\pi(\boldsymbol{\beta},\sigma_v^2,\rho) \propto g(\sigma_v^2)\,h(\rho),\quad \boldsymbol{\beta}\in\mathbb{R}^p,\sigma_v^2>0,\ l_k<\rho<u_k, \tag{2.10}$$

where σ_v^2 is the model scale parameter, $g(\sigma_v^2)$ and $h(\rho)$ are suitable functions of σ_v^2 and $\rho,$ l_k and u_k are the lower and upper bounds of ρ under the k^{th} model. We avoid the term "model error variance" for σ_v^2 as diagonal entries of $\mathbf{\Omega}_k(\rho)$ vary across small areas and are not necessarily all one.

2.2 Estimation of population means of non-sampled small areas

In this section, we consider the case when, in the survey, there are several non-sampled small areas that have no direct estimates. In many applications, limited resources frequently preclude the inclusion of many subpopulations in the sample, resulting in non-sampled small areas. In this section, we consider the case when, in the survey, there are several non-sampled small areas that have no direct estimates. In many applications, limited resources frequently preclude the inclusion of many subpopulations in the sample, resulting in non-sampled small areas. Non-sampled small areas are sometimes referred to as misaligned areas (Trevisani and Gelfand, 2013) when they arise from domain misalignment between the direct estimate and auxiliary variables. For any of these non-sampled areas, the prediction of its mean from any non-spatial model is only based on its synthetic estimator. We propose to exploit spatial dependencies in predicting area means of non-sampled small areas. The predictions of the proposed models are obtained by modifying its synthetic estimator, using the vector of regression residuals, with more emphasis on the regression residuals of the neighboring areas.

Without loss of generality, let there be m_1 non-sampled small areas and $Y_{m_1+1}, \ldots, Y_m$ be the direct estimates of the $m_2 = m - m_1$ sampled small areas. Based on the direct estimates of m_2 sampled areas, we consider the following HB models:

$$\mathbf{Y}_{(2)} \left| \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho \right. \sim N_{m_2}(\boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}), \tag{2.11}$$

$$\boldsymbol{\theta} \left| \boldsymbol{\beta}, \sigma_v^2, \rho \right. \sim N_m\left(\mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \left\{\boldsymbol{\Omega}_k(\rho)\right\}^{-1}\right), \quad k = 1, \ldots, 5, \tag{2.12}$$

$$\pi(\boldsymbol{\beta}, \sigma_v^2, \rho) \propto g(\sigma_v^2)\, h(\rho), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma_v^2 > 0,\ l_k < \rho < u_k, \tag{2.13}$$

where $\mathbf{Y}_{(2)} = (Y_{m_1+1}, \ldots, Y_m)^\top$, $\mathbf{D}_{(2)} = \mathrm{diag}\{D_i\}_{i=m_1+1}^m$, and $\boldsymbol{\theta}_{(2)} = (\theta_{m_1+1}, \ldots, \theta_m)^\top$, which is the subvector of $\boldsymbol{\theta}$ corresponding to the sampled areas.

2.3 Propriety of the posterior distributions

In this section, we establish propriety of the posterior distributions of spatial small area models given in (2.8)-(2.10) and (2.11)-(2.13). Let $I(\cdot)$ be the indicator function taking the value 1 when its argument is true and 0 otherwise. We first provide general conditions for the posterior propriety of the proposed models.

Theorem 1. For all the HB spatial models given in (2.8)-(2.10) and (2.11)-(2.13), the posterior probability density functions are proper if the following conditions hold for some positive constant $c > 0$:

(a) $\int_0^\infty g(\sigma_v^2)\, I(\sigma_v^2 \le c)\, d\sigma_v^2 < \infty.$

(b) $\int_0^\infty (\sigma_v^2)^{-(m^* - p)/2}\, g(\sigma_v^2)\, I(\sigma_v^2 > c)\, d\sigma_v^2 < \infty.$

(c) $\int_{l_k}^{u_k} h(\rho)\, d\rho < \infty,$

where $m^* = m$ for (2.8)-(2.10), and $m^* = m - m_1$ for (2.11)-(2.13).

If $g(\cdot)$ is a proper pdf, then (a) holds true automatically, and (b) is satisfied if $m^* \geq p.$ The condition $m^* \geq p$ is obvious since at least p observations are needed to estimate p components of $\boldsymbol{\beta}$ when no substantive information about it is available. Also, any bounded function of ρ satisfies (c) in Theorem 1 as their supports are all bounded. In particular, under the popular family of noninformative priors

$$\pi(\boldsymbol{\beta}, \sigma_v^2, \rho) \propto (\sigma_v^2)^{-\alpha} I(l_k < \rho < u_k), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma_v^2 > 0, \tag{2.14}$$

the posterior pdfs are proper under the following conditions.

Corollary 1. For any of the HB spatial models given in (2.8)-(2.9) and (2.11)-(2.12) with the prior in (2.14), the posterior pdf is proper as long as $\alpha < 1$ and $m^* > p + 2 - 2\alpha.$

For the uniform prior with $\alpha = 0$ (which will be used in this paper), the propriety of the posterior distributions for models (2.8)-(2.9) are guaranteed as long as the number of small areas is greater than $p + 2.$ For the models incorporating non-sampled areas given in (2.11)-(2.12), the second condition of Corollary 1.1 becomes $m - m_1 > p + 2,$ and thus, the posterior pdfs are proper as long as the number of non-sampled areas is fewer than $m - p - 2,$ or at least $p + 3$ areas have sample.

3. Simulating posterior distributions

In this section, we illustrate the rejection sampling steps to obtain independent posterior samples from the posterior distributions of proposed models. We assume that the components of the small area mean vector $\boldsymbol{\theta}$ are arranged so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top)^\top,$ where $\boldsymbol{\theta}_{(1)} \in \mathbb{R}^{m_1}$ and $\boldsymbol{\theta}_{(2)} \in \mathbb{R}^{m_2}$ are the small area mean vectors corresponding to the non-sampled and sampled areas, respectively. For notational convenience, we denote the precision matrix of the k^{th} spatial model by $\boldsymbol{\Omega} = (\sigma_v^2)^{-1} \boldsymbol{\Omega}_k(\rho)$ and the permissible range of ρ by (l, u) suppressing the model index $k.$

We first derive the marginal posterior density of (σ_v^2, ρ) and provide subsequent sampling procedures. Let $\mathbf{0}_{m_2 \times m_1}$ be the $m_2 \times m_1$ null matrix and $\mathbf{M} = [\mathbf{0}_{m_2 \times m_1}, \mathbf{I}_{m_2}]$ such that $\boldsymbol{\theta}_{(2)} = \mathbf{M}\boldsymbol{\theta}.$ We also let $\mathbf{X}_{(2)} = \mathbf{MX}.$ Integrating out $\boldsymbol{\theta}$ from the model (2.11)-(2.12), we have $\mathbf{Y}_{(2)} \,|\, \boldsymbol{\beta}, \sigma_v^2, \rho \sim N_{m_2}(\mathbf{X}_{(2)} \boldsymbol{\beta}, \boldsymbol{\Delta}),$ where $\boldsymbol{\Delta} = \mathbf{D}_{(2)} + \mathbf{M}\boldsymbol{\Omega}^{-1}\mathbf{M}^\top.$ Subsequent marginalization of $\boldsymbol{\beta}$ gives the marginal posterior density $p(\sigma_v^2, \rho \,|\, \mathbf{y}_{(2)})$ as

$$p(\sigma_v^2, \rho \,|\, \mathbf{y}_{(2)}) \propto \frac{\exp\left[-\frac{1}{2}\mathbf{y}_{(2)}^\top \boldsymbol{\Delta}^{-1} \left\{\boldsymbol{\Delta} - \mathbf{X}_{(2)} (\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1} \mathbf{X}_{(2)})^{-1} \mathbf{X}_{(2)}^\top\right\} \boldsymbol{\Delta}^{-1} \mathbf{y}_{(2)}\right]}{|\boldsymbol{\Delta}|^{1/2} \left|\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1} \mathbf{X}_{(2)}\right|^{1/2}} I(l < \rho < u). \tag{3.1}$$

Furthermore, we have conditional posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as

$$\boldsymbol{\beta} \,|\, \sigma_v^2, \rho, \mathbf{y} \sim N_p(\boldsymbol{\gamma}, \boldsymbol{\Gamma}), \tag{3.2}$$

$$\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma_v^2, \rho, \mathbf{y} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Psi}), \tag{3.3}$$

where $\boldsymbol{\Gamma} = (\mathbf{X}_{(2)}^{\top} \boldsymbol{\Delta}^{-1} \mathbf{X}_{(2)})^{-1}$, $\boldsymbol{\gamma} = \boldsymbol{\Gamma} \mathbf{X}_{(2)}^{\top} \boldsymbol{\Delta}^{-1} \mathbf{y}_{(2)}$, $\boldsymbol{\mu} = \mathbf{y}_{*} - \boldsymbol{\Psi}\boldsymbol{\Omega}(\mathbf{y}_{*} - \mathbf{X}\boldsymbol{\beta})$, $\mathbf{y}_{*} = (\mathbf{0}_{m_1}^{\top}, \mathbf{y}_{(2)}^{\top})^{\top}$, and $\boldsymbol{\Psi}^{-1} = \mathbf{M}^{\top} \mathbf{D}_{(2)}^{-1} \mathbf{M} + \boldsymbol{\Omega}$. Accordingly, we can obtain a independent posterior sample via rejection sampling from (3.1) and subsequent samplings from (3.2) and (3.3). For the data with no non-sampled area, we have desired sampling procedures by setting $\mathbf{M} = \mathbf{I}_m$ and $\mathbf{y}_{*} = \mathbf{y}$.

4. A simulation study

In this section, we compare prediction performances of the independent FH model and the four spatial models in the absence of informative covariates with multiple non-sampled areas. Excluding Hawaii and Alaska, we consider contiguous $m = 49$ states of the U.S., including the District of Columbia. To evaluate the quality of prediction in the absence of direct estimates, we do not simulate direct estimates of randomly chosen $m_1 = \lfloor 0.15m \rfloor = 7$ states. These areas are Delaware, Massachusetts, Michigan, Nebraska, Rhode Island, South Dakota, and Texas. This results in $m_2 = 42$ areas which have direct estimates.

To make simulation settings realistic, we mimic the 1989 4-person family median income (median income) data described in Section 5. We generate replicated datasets so that Moran's I values of each replicated small area means, $\theta_1, \ldots, \theta_m$, are approximately centered around 0.44, the Moran's I value of the 1990 Census median income. Direct estimates are generated with the sampling variances $D_1, \ldots, D_{m_2}$ of the 1990 Current Population Survey (CPS) estimates. These sampling variances of sampled small areas range from 1.95 to 25.03 with the mean of 9.08, where dollar amounts are scaled by $1,000. For each setting, we consider $S = 100$ replicated datasets.

Data generation: Let $\bar{D} = m_2^{-1} \sum_{i=1}^{m_2} D_i$ and $\mathbf{D}_{(2)} = \operatorname{diag}\{D_i\}_{i=1}^{m_2}$. We set $\rho = 0.85$ and $\sigma_v^2 = \bar{D}/2$ and consider two independent covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ with SAR spatial dependence, i.e., $\mathbf{x}_1, \mathbf{x}_2 \sim N_m\left(\mathbf{0}_m, \{\boldsymbol{\Omega}_3(\rho)\}^{-1}\right)$. Then, letting $(\beta_1, \beta_2)^{\top} = (2, 1)^{\top}$ and $\boldsymbol{\mu} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$, we generate small area means and direct estimates from the following independent FH model:

$$\boldsymbol{\theta} \sim N_m(\boldsymbol{\mu}, \sigma_v^2 \mathbf{I}_m), \quad \mathbf{Y}_{(2)} \,\big|\, \boldsymbol{\theta}_{(2)} \sim N_{m_2}(\boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}),$$

where the components of $\mathbf{Y}_{(2)}$ and $\boldsymbol{\theta}_{(2)}$ correspond to the m_2 sampled small areas as defined below equation (2.13). The covariate $\mathbf{x}_1 (\mathbf{x}_2)$ introduces stronger (weaker) spatial pattern to the θ_i's, and accordingly, we call $\mathbf{x}_1$ $(\mathbf{x}_2)$ as the strong (weak) covariate. Moran's I values of 100 replicated small area means range from 0.115 to 0.713 with mean 0.449.

We consider two different covariate settings to examine how the spatial models can capture extra variability introduced by the spatial dependence from a missing covariate, i.e., $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_2]$ and $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1]$, where $\mathbf{1}_m$ represents the m-component vector of ones. Excluding any of the covariates from the fitted model will leave the spatial variation of that covariate to the residual. We do not consider the full model involving both the covariates since that model will fully capture $\boldsymbol{\mu}$ and leave no spatial variability unexplained; consequently, the independent FH model will be sufficient to capture the variability of the i.i.d. random effects.

Posterior simulations: For all models proposed, independent posterior samples can be obtained by the rejection sampling scheme outlined in Section 3. The sampling procedure begins with the rejection sampling from the marginal posterior distribution of (σ_v^2, ρ) and continues with successive samplings of the rest of the parameters from the conditional posterior distributions. However, when the marginal posterior density of (σ_v^2, ρ) is concentrated at or around the boundaries of $\rho,$ a proposal distribution must be carefully specified to have a sufficiently high acceptance rate, which may require adaptive specification a proposal distribution for each replicated dataset. To avoid such difficulties, we use Hamiltonian Monte Carlo algorithm with the `R` package `rstan` (Stan Development Team, 2018). We fit the HB model (2.11)-(2.13) with each combination of covariates for $k = 1, \ldots, 5.$ For each model, we run four parallel Hamiltonian Monte Carlo chains (No-U-Turn Sampler) for 2,500 iterations after 5,000 burn-in iterations. We keep every 10th iteration and concatenate the four chains to obtain a posterior sample of size 10,000. The `R` codes implementing the rejection sampling step described in Section 3 and the `stan` models are available at https://github.com/heech31/spatial_sae.

Measures of performance: With the posterior sample under each model, we predict the true small area mean vector, $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \ldots, \theta_m^{(s)})^\top,$ of the s^{th} replicated dataset using the posterior mean, which we denote by $\hat{\boldsymbol{\theta}}^{(s)} = (\hat{\theta}_1^{(s)}, \ldots, \hat{\theta}_m^{(s)})^\top.$ Let $\mathcal{A}$ be a subset of $\{1, \ldots, m\},$ which is determined by indices only of the sampled or non-sampled small areas. For a given subset $\mathcal{A},$ we calculate the mean squared prediction error, $\text{MSPE}^{(s)} = \sum_{i \in \mathcal{A}} (\hat{\theta}_i^{(s)} - \theta_i^{(s)})^2 \big/ m_{\mathcal{A}},$ where $m_{\mathcal{A}} = |\mathcal{A}|$ is the number of areas in $\mathcal{A}.$ We then average $\text{MSPE}^{(s)}$ over S replications to compute the average empirical mean squared prediction error (AeMSPE), where

$$\text{AeMSPE} = \frac{1}{S}\sum_{s=1}^{S} \text{MSPE}^{(s)} = \frac{1}{S}\sum_{s=1}^{S} \frac{1}{m_{\mathcal{A}}} \sum_{i \in \mathcal{A}} (\hat{\theta}_i^{(s)} - \theta_i^{(s)})^2. \tag{4.1}$$

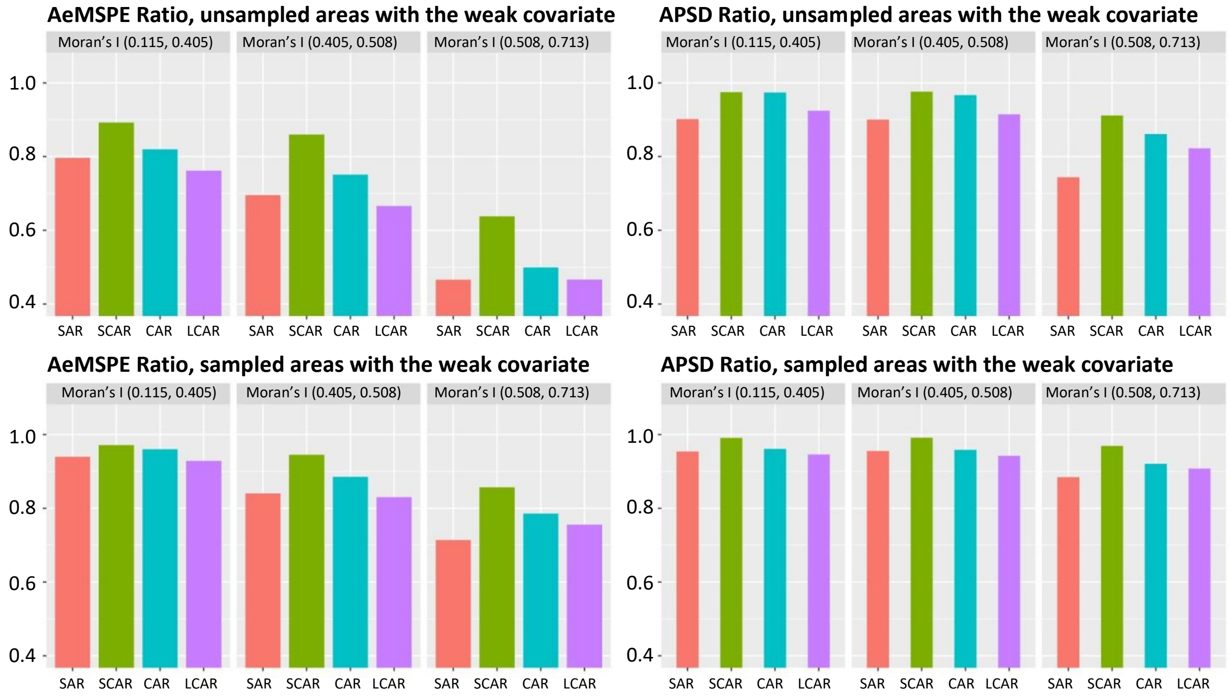
We also evaluate the uncertainty of the predictions using the average posterior standard deviation (APSD) defined as $S^{-1} \sum_{s=1}^{S} m_{\mathcal{A}}^{-1} \sum_{i \in \mathcal{A}} \text{sd}(\theta_i^{(s)}),$ where $\text{sd}(\theta_i^{(s)})$ is the posterior standard deviation of $\theta_i.$ By setting the independent FH model as a reference model, we consider the following ratios:

$$\text{AeMSPE} - \text{Ratio}_k = \frac{\text{AeMSPE}_k}{\text{AeMSPE}_1}, \quad \text{APSD} - \text{Ratio}_k = \frac{\text{APSD}_k}{\text{APSD}_1}, \tag{4.2}$$

where the subscript k indicates that the quantity is calculated from the posterior sample under the k^{th} model. These ratios measure improvements in AeMSPE and APSD achieved by fitting a spatial model over the independent FH model. A ratio less than 1 indicates superiority of the spatial model, otherwise the independent FH model is better. The smaller the ratio is, the more superior a spatial model is.

Model comparison: Various plots of Figure 4.1 summarize the ratios when the strong covariate $\mathbf{x}_1$ is excluded from the fitted models. We categorize AeMSPE-Ratios and APSD-Ratios under each replication into three groups based on the Moran's I values of θ_i's, namely the lowest, middle, and biggest thirds, respectively. The first row summarizes the prediction results for the seven non-sampled areas. As expected, spatial models show remarkable improvement in AeMSPE and APSD, and the improvements are greater when Moran's I values are larger. In terms of AeMSPE, the SAR and LCAR models produce at least 20%, 30%, and 50% more accurate predictions when Moran's I values are in the first, second, and third groups, respectively. In terms of the uncertainty of prediction, predictions of the SAR and LCAR models have 10% smaller APSD for the first and second groups. In the third group, the SAR model shows more than 25% reduction in APSD.

Figure 4.1 AeMSPE-Ratios and ASPDE-Ratios for predictions with the *weak* covariate.

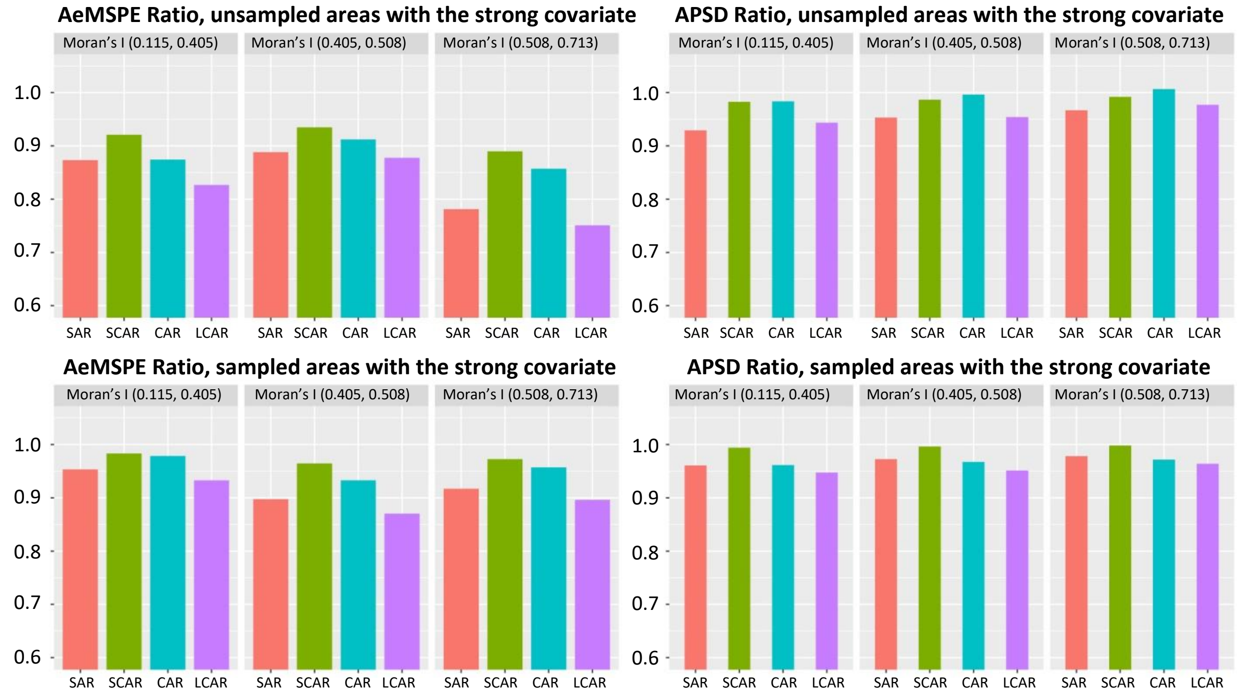


- Vertical bars are drawn for the ratios, where a bar shorter than 1 represents better (smaller) AeMSPE or APSD for the corresponding model relative to the independent FH model.
- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive, AeMSPE = Average empirical mean squared prediction error, APSD = Average posterior standard deviation.

For the sampled areas with direct estimates, the improvements in AeMSPE are less than 10% for the first group but more than 15% and 25% for the second and third groups (grouped by Moran's I values), respectively. Additionally, predictions from spatial models have a lower level of uncertainty, and for the third group, the SAR model predictions have more than 10% smaller ASPD.

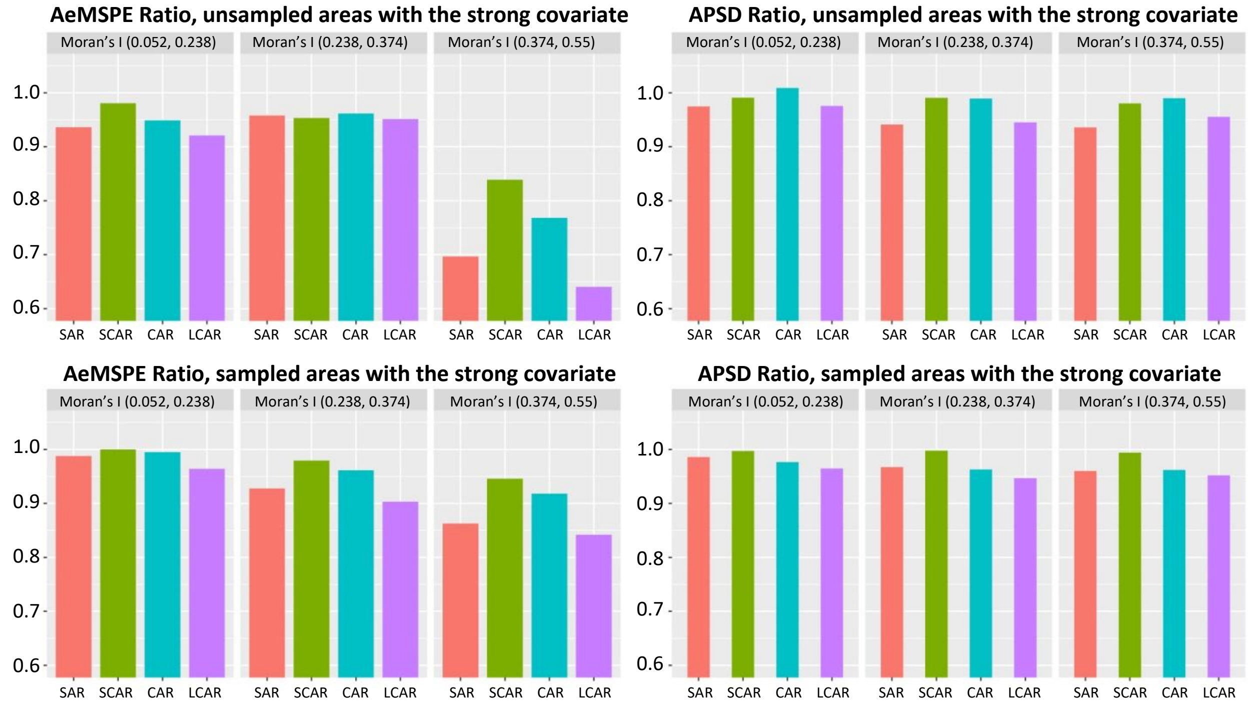
Similarly, various plots of Figure 4.2 summarize the ratios when the weak covariate $\mathbf{x}_2$ is excluded from the fitted models. In general, the spatial models continue to produce better predictions over the independent FH model. The LCAR model performs the best overall for the seven non-sampled areas, resulting in a reduction of over 10% in AeMSPE for the first and second groups, respectively, and around 25% in AeMSPE for the third group. In terms of uncertainty, the spatial models show more than 5% but less than 10% smaller APSD. For the sampled small areas, the SAR and LCAR models show an AeMSPE reduction of around 5%-13%, but APSD reductions are less than 5%. Unlike the previous results with the weak covariate, the improvements in AeMSPE and APSD are comparable across three groups that are categorized by Moran's I values. This is because the Moran's I values of small area means are mostly determined by the strong covariate, and once the strong covariate is present in the model to explain the spatial variability, the spatial variability in the residuals do not vary markedly across the three groups. We regroup the ratios in terms of the Moran's I values of the residuals obtained by regressing the strong covariate on the small area means, and summarize the ratios in Figure 4.3. Under this categorization, the LCAR model shows the best performance illustrating 5%, 10%, and 15% AeMSPE reductions for the first, second, and third groups, respectively. It can be also seen that the bigger the Moran's I value is, the more improvement the spatial models achieve.

Figure 4.2 AeMSPE-Ratios and ASPDE-Ratios predictions of non-sampled and sampled area means with the *strong* covariate.



- Vertical bars are drawn for the ratios, where a bar shorter than 1 represents better (smaller) AeMSPE or ASPDE for the corresponding model relative to the independent FH model.
- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive, AeMSPE = Average empirical mean squared prediction error, APSD = Average posterior standard deviation.

Figure 4.3 AeMSPE-Ratios and ASPDE-Ratios for predictions of non-sampled and sampled area means with the *strong* covariate. Results are grouped by the Moran's I values of the residuals, where the residuals are obtained by regressing the strong covariate on the small area means.



- Vertical bars are drawn for the ratios, where a bar shorter than 1 represents better (smaller) AeMSPE or ASPDE for the corresponding model relative to the independent FH model.
- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive, AeMSPE = Average empirical mean squared prediction error, APSD = Average posterior standard deviation.

5. Application to the Current Population Survey data

In this section, we evaluate the spatial models in terms of their prediction accuracy for some state-level population median incomes. The U.S. Department of Health and Human Service (HHS) annually needed accurate data for median incomes for states to implement a welfare program. While accurate national median income data are available from the Current Population Survey (CPS), the CPS data do not provide accurate state-level median income data. To supply accurate statistics to the HHS, the U.S. Census Bureau considered model-based small area estimation methods by utilizing auxiliary data from other federal programs. We apply proposed spatial models to estimate 1989 four-person family median incomes for the contiguous forty-nine U.S. states, including the District of Columbia. We use the direct estimates of 1990 CPS and compare our predictions with the more reliable statistics from the *long form* from the 1990 Census, i.e., we consider the statistics from the long form from the 1990 Census as the true values. Prediction performances are measured using all small areas and a subset of areas after leaving out multiple direct estimates.

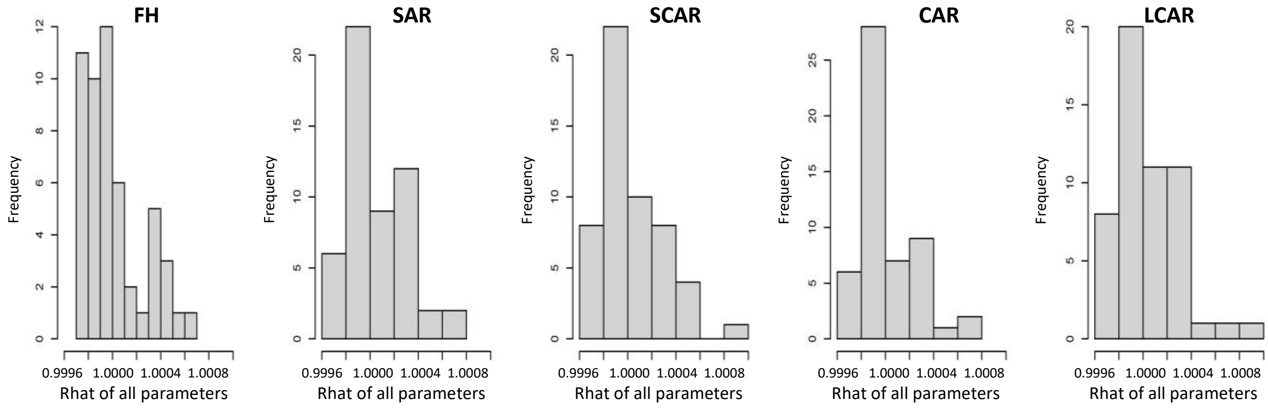
5.1 Four-person family median income estimation

Let θ_i be the true four-person family median income of the i^{th} state for the year 1989, where $i = 1, \ldots, 49.$ The states of Alaska and Hawaii are excluded as they are not geographically connected to the mainland. Let Y_i be the direct estimate of θ_i from the 1990 CPS. The covariates of interest are 1980 Census median income x_{i1} and an adjusted 1980 Census median income x_{i2}. The adjusted Census median income x_{i2} is defined as $\left(\text{PCI}_{i,1989} \,/\, \text{PCI}_{i,1979}\right) x_{i1},\ i = 1, \ldots, m,$ where $\text{PCI}_{i,1979}$ and $\text{PCI}_{i,1989}$ are the 1979 and 1989 per capita incomes of the i^{th} state provided by the Bureau of Economic Analysis of the U.S. Department of Commerce. It has been known that the adjusted Census median income is a good covariate which very effectively accounts for the variability of the small area median income.

With the noninformative prior (2.14) with $\alpha = 0,$ we fit all five models as described by (2.8)-(2.10) with $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1, \mathbf{x}_2]$ and $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1],$ where for the second covariate setting, we exclude the adjusted Census median income from the fitted model. For each model considered, we run 4 parallel HMC chains for 2,500 iterations after 5,000 burn-in iterations using `rstan` (Stan Development Team, 2018). We retain every 10^{th} iteration and concatenate the 4 chains to obtain a posterior sample of size 10,000. For all models and parameters, the potential scale reduction factors ($\hat{R};$ Gelman and Rubin, 1992) are all one indicating no lack of convergence. The potential scale reduction factors are provided in Figure 5.1 and Table 5.1.

Using the posterior means, $\hat{\theta}_i$'s, we calculate the squared prediction errors from respective θ_i's and obtain the mean squared prediction error (MSPE), defined in Section 4, by averaging the $m = 49$ squared deviations. The average posterior standard deviations (APSD) associated with $\hat{\theta}_i$'s are used to quantify the uncertainty of predictions, and the widely applicable information criterion (WAIC; Watanabe and Opper, 2010) is used to evaluate and compare the models, where a smaller WAIC value indicates a better model fit.

Figure 5.1 The potential scale reduction factor $\hat{R}$ of all parameters when no non-sampled area exists.



- All values are practically one indicating no evidence of lack of convergence.
- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

Table 5.1
The potential scale reduction factor $\hat{R}$ of the hyperparameter σ_v^2 and ρ and corresponding 95% upper confidence limit for the dataset with no non-sampled area

Hyperparameter	Covariate included	Potential scale reduction factor (upper 95% confidence limit)				
		FH	**SAR**	**SCAR**	**CAR**	**LCAR**
σ_v^2	x_1, x_2	0.995 (1.017)	0.987 (1.010)	0.995 (1.018)	0.985 (1.007)	1.008 (1.031)
	x_1	0.993 (1.015)	0.991 (1.012)	0.999 (1.022)	0.987 (1.009)	0.987 (1.01)
ρ	x_1, x_2	–	1.005 (1.028)	0.997 (1.023)	0.998 (1.036)	0.994 (1.017)
	x_1	–	1.005 (1.025)	0.997 (1.016)	0.998 (1.017)	1.019 (1.039)

Note: All upper limits are within the acceptable range, below 1.1 (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin, 2013, Chapter 11.5), indicating no evidence of lack of convergence.

- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

Table 5.2 summarizes various evaluation measures we considered and the respective percentage improvements (PI) of MSPE, APSD. When both covariates are available, the LCAR model has approximately 14% smaller MSPE and 4% smaller APSD than the independent FH model. In terms of MSPE, the second best performing model is the SAR having approximately 9.5% smaller MSPE. When only x_1 (week covariate) is included in the fitted model, the SAR model has approximately 40% smaller MSPE and 14% smaller APSD than the independent FH model. The CAR and LCAR models show competitive performances having approximately 36% smaller MSPE and 13% smaller APSD over the independent FH model. By removing the strong covariate x_2 from the full model, the MSPE of the SAR and LCAR models increase approximately 66% and 84%, respectively, whereas the MSPE of the independent FH model increases more than 150%.

Table 5.2
Mean squared prediction error, average posterior standard deviation, and respective percentage improvements (PI) of spatial models over the independent FH model

Covariate Included	x_1 , x_2					x_1				
	MSPE	**MSPE-PI**	**APSD**	**APSD-PI**	**WAIC**	**MSPE**	**MSPE-PI**	**APSD**	**APSD-PI**	**WAIC**
FH	2.88	–	1.93	–	*259.06 (7.13)*	7.27	–	2.31	–	267.75 (8.44)
SAR	2.61	*9.55%*	1.94	0.34%	261.46 (7.47)	4.34	*40.22%*	1.98	*14.25%*	265.76 (8.16)
SCAR	3.03	-5.14%	1.95	-0.91%	259.37 (7.01)	5.62	22.62%	2.22	3.52%	263.41 (7.29)
CAR	2.64	8.47%	1.91	1.24%	261.61 (7.86)	4.62	*36.35%*	2.01	*12.97%*	*263.32 (7.96)*
LCAR	2.47	*14.50%*	1.85	*4.19%*	261.79 (8.01)	4.54	*37.51%*	1.97	*14.36%*	263.35 (8.08)

- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive, MSPE = Mean squared prediction error, APSD = Average posterior standard deviations, WAIC = Widely applicable information criterion.

In terms of the goodness of fit, the independent FH model shows the best fit (the smallest WAIC) when both covariates are included. Conversely, when only x_1 (week covariate) is included in the fitted model, the independent FH model shows the worst fit having the largest WAIC value. However, considering the standard errors given in parentheses, there is no significant difference in the model fit.

Table 5.3 summarizes the posterior distributions of ρ in terms of the posterior mean, mode, and standard deviation. When all covariates are included in the fitted model, the posterior distributions of ρ indicate no strong spatial dependency having posterior means centered around zero with large standard deviations. In contrast, when only x_1 (week covariate) is included in the fitted model, ρ becomes very significant illustrating the posterior distributions concentrated near the upper limit of its support.

In summary, when the weak covariate does not adequately explain existing spatial variation, the spatial models produce significantly better predictions accounting for the spatial variation. When no substantial spatial variation remains in the residual, they make marginally better predictions than the independent FH model without sacrificing model fit.

Table 5.3
Posterior mean/mode (standard deviation) of ρ

Covariate included	SAR	SCAR	CAR	LCAR
x_1, x_2	0.10/0.40 (0.48)	-0.06/0.04 (0.14)	0.21/0.83 (0.55)	0.57/0.80 (0.27)
x_1	0.76/0.80 (0.14)	0.14/0.17 (0.04)	0.93/0.99 (0.09)	0.85/0.97 (0.13)

Note: The first and second rows of the table respectively summarizes posterior distributions of ρ when both covariates and only x_1 are included in the fitted model.

- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

5.2 Estimation of some non-sampled state means excluding their CPS values

In this section, we evaluate the spatial models in terms of their prediction accuracy for non-sampled small areas using the 1980 Census median income x_1. Specifically, we randomly exclude CPS estimates (direct estimates) of multiple states at each instance and make predictions for θ_i's of the excluded states. As there are 49 small areas (states), we created 12 datasets that lack direct estimates for $m_1 = 4$ or 5 areas, where m_1 is the number of non-sampled small areas as in Section 2.2. Excluded states for each dataset are listed in Table 5.4.

For each dataset, we fit the independent FH model and four spatial models as specified in (2.11)-(2.13) with the noninformative prior (2.13) with $\alpha = 0$ running HMC chains under the same setting as in Section 5.1. The $\hat{R}$ values show no evidence of lack of convergence, where detailed values are provided in Figure 5.2 and Table 5.5. For each non-sampled area, the squared prediction error $\text{SPE}_i = (\hat{\theta}_i - \theta_i)^2$ and posterior standard deviation $\text{sd}(\theta_i)$ are obtained for each model. Based on these, the prediction performance is compared with the following ratio: for $i = 1, \ldots, m$ and $k = 2, \ldots, 5,$

$$\text{SPE}-\text{Ratio}_{ki} = \frac{\text{SPE}_{ki}}{\text{SPE}_{1i}}, \quad \text{PSD}-\text{Ratio}_{ki} = \frac{\text{sd}_k(\theta_i)}{\text{sd}_1(\theta_i)} \tag{5.1}$$

where $\text{sd}_k(\theta_i)$ is the posterior standard deviation of θ_i under the k^{th} model. A value of $\text{SPE}-\text{Ratio}_{ki}$ $(\text{PSD}-\text{Ratio}_{ki})$ less than one indicates that the k^{th} spatial model has a smaller squared prediction error (posterior standard deviation) than the independent FH model. In Figures 5.3 and 5.4, we display the ratios using a red and blue color scheme to denote ratios greater than and less than one, respectively, where a darker color represents a more extreme value.

Overall, SAR, SCAR, CAR, and LCAR models exhibit smaller SPEs in 35, 41, 36, and 36 states, respectively. The SCAR model has the greatest number of states in which its predictions perform better than those of the independent FH model, but the overall improvements are the least. In more than 35 states, the SAR, CAR, and LCAR models produce more accurate predictions than the independent FH model, and in three states (New Mexico, Oregon, and Wisconsin), their SPEs are more than 100 times smaller. For California, Minnesota, and South Carolina, all spatial models make worse predictions than the independent FH model. California and Minnesota have much higher median incomes than surrounding states, whilst South Carolina has a significantly lower median income. Among the 49 states, these three states have the second, seventh, and nineteenth smallest local Moran's I values. This illustrates that if the small area mean of an non-sampled area is significantly different from the means of surrounding areas, then the spatial models may produce inferior predictions. The model that demonstrates the best fit in terms of WAIC is either the SCAR, CAR, or LCAR model, where the exact numbers are provided in Table 5.4.

Table 5.4
States whose CPS estimates are excluded for each dataset and corresponding widely applicable information criterion (WAIC)

Excluded states	FH	SAR	SCAR	CAR	LCAR
AZ MS OK SD	246.15 (7.79)	244.37 (7.50)	*241.71 (6.67)*	242.46 (7.50)	242.21 (7.57)
AR CO DE TN	245.79 (7.83)	241.23 (7.69)	241.05 (6.70)	*239.70 (7.61)*	239.85 (7.69)
MD MI NV WV	246.06 (7.83)	242.23 (7.58)	241.34 (6.78)	240.13 (7.32)	*240.02 (7.43)*
MT NC NE NY	248.91 (8.10)	245.97 (9.51)	243.85 (6.87)	242.88 (8.34)	*242.40 (8.36)*
DC GA ID ND	245.07 (7.91)	241.02 (6.61)	240.09 (6.44)	*239.16 (6.75)*	239.48 (6.75)
AL MO VT WY	245.57 (8.31)	245.13 (7.67)	*242.74 (7.38)*	242.89 (7.65)	243.36 (7.66)
FL LA UT WA	247.38 (7.39)	245.64 (7.29)	*243.08 (6.44)*	243.25 (7.20)	243.48 (7.33)
MA MN SC TX	248.32 (9.73)	242.43 (8.95)	244.75 (8.83)	240.99 (8.69)	*240.34 (8.56)*
KY RI VA WI	243.86 (8.09)	241.74 (7.06)	240.05 (6.76)	239.04 (6.91)	*238.87 (6.90)*
IL IN NH PA	244.69 (7.10)	245.12 (7.41)	*243.31 (6.59)*	244.45 (7.60)	244.39 (7.68)
CA ME NJ OH	248.62 (8.54)	244.06 (8.26)	245.28 (7.68)	*242.00 (7.78)*	242.01 (7.90)
CT IA KS NM OR	239.86 (7.95)	239.81 (7.76)	*237.28 (7.29)*	237.78 (7.75)	237.75 (7.60)

Note: The numbers in parentheses are the standard errors of WAIC estimates.

- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.
- AL = Alabama, AZ = Arizona, AR = Arkansas, CA = California, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District of Columbia, FL = Florida, GA = Georgia, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiana, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = New Mexico, NY = New York, NC = North Carolina, ND = North Dakota, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvania, RI = Rhode Island, SC = South Carolina, SD = South Dakota, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginia, WA = Washington, WV = West Virginia, WI = Wisconsin, WY = Wyoming.

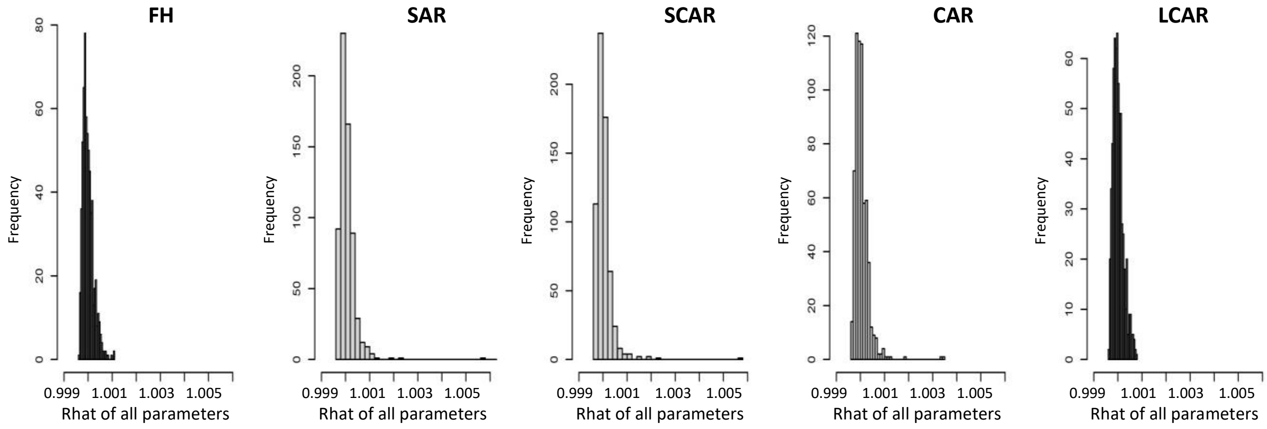
Table 5.5
The potential scale reduction factor $\hat{R}$ of the hyperparameter σ_v^2 and ρ and corresponding 95% upper confidence limit for the 12 datasets with non-sampled areas

	Excluded states	FH	SAR	SCAR	CAR	LCAR
Potential scale reduction factor (upper 95% confidence limit) of σ_v^2	AZ MS OK SD	0.991 (1.013)	1.012 (1.036)	0.995 (1.017)	1.005 (1.028)	1.005 (1.030)
	AR CO DE TN	0.997 (1.019)	1.010 (1.034)	0.999 (1.023)	0.985 (1.008)	0.984 (1.007)
	MD MI NV WV	0.992 (1.015)	1.007 (1.030)	1.023 (1.047)	1.001 (1.025)	1.011 (1.035)
	MT NC NE NY	0.990 (1.012)	0.984 (1.009)	0.997 (1.020)	0.999 (1.024)	0.983 (1.006)
	DC GA ID ND	1.002 (1.024)	1.007 (1.032)	1.001 (1.024)	1.011 (1.036)	0.997 (1.020)
	AL MO VT WY	0.997 (1.019)	1.006 (1.030)	0.998 (1.021)	1.002 (1.026)	1.019 (1.045)
	FL LA UT WA	1.014 (1.037)	1.005 (1.027)	1.002 (1.025)	0.999 (1.023)	1.019 (1.043)
	MA MN SC TX	1.008 (1.031)	1.001 (1.025)	1.003 (1.025)	0.994 (1.019)	1.007 (1.032)
	KY RI VA WI	1.011 (1.034)	1.005 (1.029)	1.009 (1.033)	0.989 (1.011)	1.011 (1.035)
	IL IN NH PA	0.992 (1.013)	1.018 (1.041)	0.992 (1.013)	0.989 (1.014)	0.989 (1.012)
	CA ME NJ OH	1.002 (1.025)	1.009 (1.033)	1.001 (1.024)	1.011 (1.035)	1.003 (1.026)
	CT IA KS NM OR	0.997 (1.018)	1.008 (1.032)	1.007 (1.030)	1.009 (1.034)	1.012 (1.041)
Potential scale reduction factor (upper 95% confidence limit) of ρ	AZ MS OK SD	–	1.005 (1.026)	1.001 (1.029)	1.004 (1.040)	0.995 (1.018)
	AR CO DE TN	–	1.019 (1.042)	0.993 (1.018)	1.000 (1.034)	1.005 (1.028)
	MD MI NV WV	–	1.005 (1.027)	0.992 (1.017)	0.988 (1.021)	1.002 (1.026)
	MT NC NE NY	–	1.001 (1.024)	0.999 (1.030)	0.986 (1.023)	0.998 (1.023)
	DC GA ID ND	–	0.999 (1.020)	0.995 (1.024)	0.991 (1.026)	1.000 (1.024)
	AL MO VT WY	–	1.003 (1.026)	1.011 (1.037)	0.989 (1.025)	1.008 (1.031)
	FL LA UT WA	–	1.004 (1.025)	0.996 (1.026)	1.000 (1.033)	1.005 (1.029)
	MA MN SC TX	–	1.002 (1.041)	1.014 (1.045)	0.984 (1.018)	0.996 (1.02)
	KY RI VA WI	–	1.010 (1.032)	0.990 (1.016)	0.993 (1.031)	0.986 (1.009)
	IL IN NH PA	–	1.030 (1.055)	0.993 (1.016)	0.994 (1.029)	1.003 (1.025)
	CA ME NJ OH	–	1.003 (1.026)	0.997 (1.026)	0.996 (1.032)	0.989 (1.012)
	CT IA KS NM OR	–	1.017 (1.040)	0.997 (1.022)	1.009 (1.044)	0.981 (1.003)

Note: All upper limits are within the acceptable range, below 1.1 (Gelman et al., 2013, Chapter 11.5), indicating no evidence of lack of convergence.

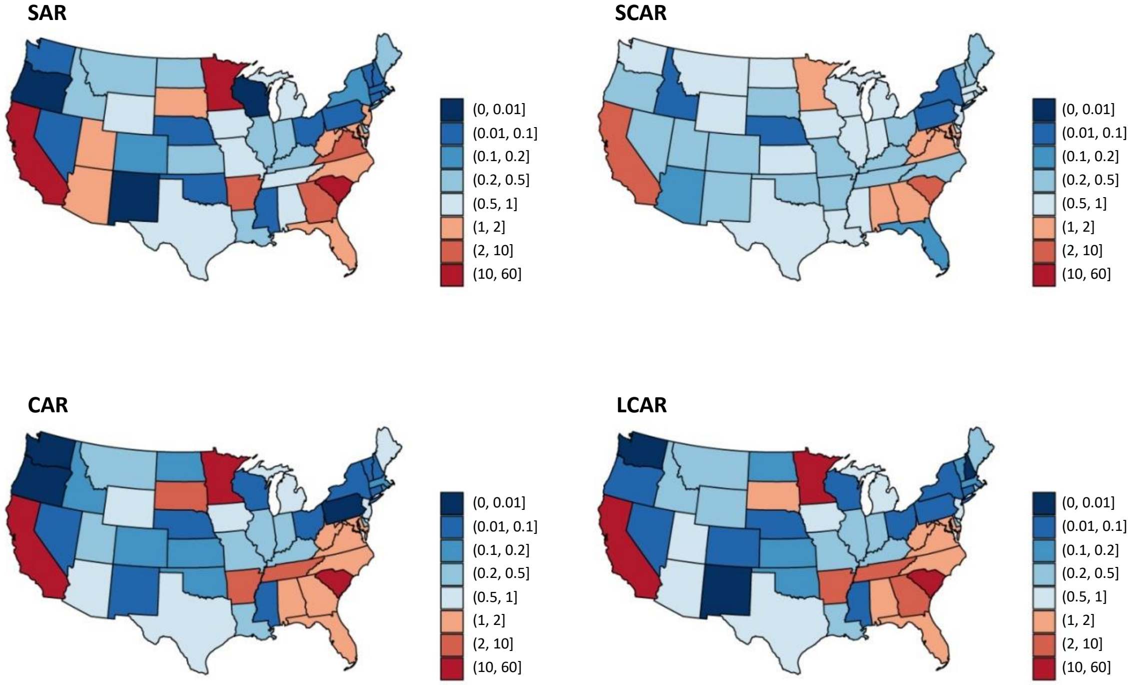
- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.
- AL = Alabama, AZ = Arizona, AR = Arkansas, CA = California, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District of Columbia, FL = Florida, GA = Georgia, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiana, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = New Mexico, NY = New York, NC = North Carolina, ND = North Dakota, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvania, RI = Rhode Island, SC = South Carolina, SD = South Dakota, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginia, WA = Washington, WV = West Virginia, WI = Wisconsin, WY = Wyoming.

Figure 5.2 The potential scale reduction factor $\hat{R}$ of all parameters, where $\hat{R}$ values of the 12 datasets are all combined.



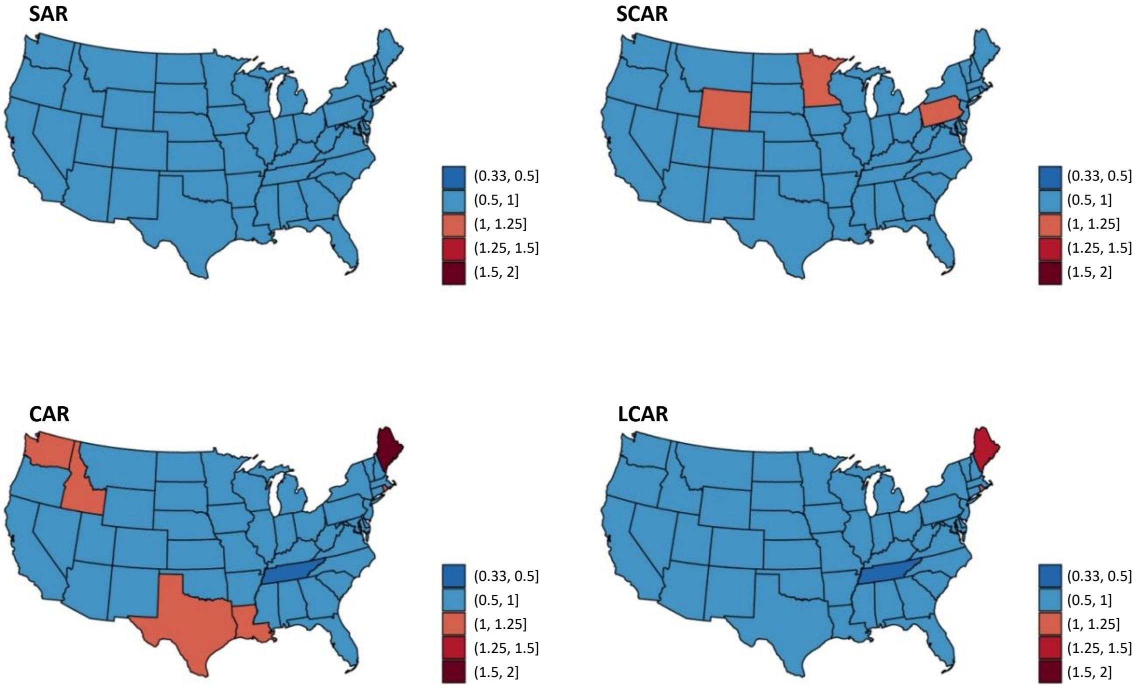
- All values are practically one indicating no evidence of lack of convergence.
- FH = Fay-Herriot, SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

Figure 5.3 SPE-Ratios.



- The values smaller (larger) than one are represented in blue (red) color and indicate a spatial model has a smaller (larger) squared deviation.
- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

Figure 5.4 PSD-Ratios.



- The values smaller (larger) than one are represented in blue (red) color and indicate a spatial model has a smaller (larger) posterior standard deviation.
- SAR = Simultaneous autoregressive, SCAR = Simple conditional autoregressive, CAR = Conditional autoregressive, LCAR = Leroux's conditional autoregressive.

6. Conclusions

In this paper, we followed a Bayesian approach to investigate four spatial random-effects models as alternatives to the independent Fay-Herriot model to estimate small area means. In particular, we considered four spatial models with different autocorrelation structures. We further extended the spatial models to allow multiple small areas without any direct estimates in predicting small area means for all the areas. For a class of noninformative priors, we established the propriety of posterior densities of the proposed models for both setups.

A simulation study in Section 4 illustrates that prediction accuracy can be greatly improved by considering spatial models when effective covariates are unavailable. Datta, Hall and Mandal (2011) noted that the prediction accuracy of small area estimation models largely depends on the availability of good covariates. In other words, when suitable covariates are unavailable, the independent Fay-Herriot model may not provide a significant advantage over direct estimates. The simulation results indicated that, in such cases, the spatial models considerably increase the prediction accuracy by exploiting information from adjacent areas.

We applied the proposed spatial random-effects models to estimate four-person family median incomes. Even when a good covariate exists, the spatial models exhibited noticeable improvements in terms of mean squared deviation and average posterior standard deviations. When a good covariate is unavailable, spatial models provided significantly more accurate median income predictions with much smaller variability, which agrees with the simulation results. Furthermore, the SAR and LCAR models provide more precise small area estimates when direct estimates of some states are excluded in model fitting.

In summary, the spatial models considered in this paper outperform the independent Fay-Herriot model. A significant improvement can be expected when effective covariates are unavailable. Since useful covariates are not always available, the utility of the proposed models in small area estimation can be substantial. Our simulation study and real data analysis demonstrate no clear winner among the proposed models. Nonetheless, the SAR and LCAR models show better performance compared with other spatial models. Also, the LCAR model performs robustly well with simulated data from the SAR model and real data with unknown spatial dependence. Thus, in the context of real applications where true dependency is unknown, we recommend the LCAR model.

This work assumes that all areas have at least one neighborhood. In real applications, however, there are many situations that data contain small areas with no neighborhood (stand-alone areas). Although the proposed models can accommodate stand-alone areas by adjusting the diagonal entries of the precision matrices as in Brown, Datta and Lazar (2017), we find that this approach results in a counterintuitive prior, where stand-alone areas have smaller prior random effect variances than areas with neighborhoods. Also, we find that this prior can considerably deteriorate predictions of stand-alone areas. This is a practically important problem as many countries have islands, and this will possibly be our future research to pursue.

Disclaimer and acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not those of the U.S. Census Bureau, or the University of Georgia. The authors are grateful to Dr. William R. Bell for his insightful comments on an earlier version of this work that led to an improved manuscript.

Appendix

A. Proof of the propriety of the posterior pdf

Proof of Theorem 1. For convenience of notation, we denote $\boldsymbol{\Omega}_k(\rho)$ by $\boldsymbol{\Omega}_k$, and, for a given square matrix $\mathbf{A}$, the determinant of $\mathbf{A}$ is denoted by $|\mathbf{A}|$. We use K to denote a generic positive constant, not depending on the variables we are integrating out.

Let $m_1 \geq 0$ be the number of small areas with no direct estimates and let $m_2 = m - m_1$. Also, let $\mathbf{Y}_{(2)}$ be the $m_2 \times 1$ vector with direct estimates of the sampled small areas. Without loss of generality, we assume that $\theta_1, \ldots, \theta_m$ are arranged so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^{\top}, \boldsymbol{\theta}_{(2)}^{\top})^{\top}$. Let $\mathbf{D}_{(2)} = \{D_i\}_{i=m_1+1}^{m}$ be the diagonal matrix with sampling variances corresponding to the components of $\mathbf{Y}_{(2)}$ and $\delta = \max_{m_1 < i \leq m} D_i < \infty$.

The joint pdf of $\mathbf{Y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2$ and ρ is given by

$$f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho) = N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)})\, N_m(\boldsymbol{\theta} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho), \tag{A.1}$$

where $N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)})$ is the normal pdf with mean $\boldsymbol{\theta}_{(2)}$ and covariance matrix $\mathbf{D}_{(2)}$. The posterior pdf $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho \,|\, \mathbf{y}_{(2)})$ will be proper if and only if the function $f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)$ is integrable with respect to $\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2$ and ρ. Since

$$N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}) \leq K \exp\left\{-\frac{1}{2\delta}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})^{\top}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})\right\},$$

we have from (A.1)

$$\begin{aligned} f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho) &\leq K \exp\left\{-\frac{1}{2\delta}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})^{\top}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})\right\} N_m(\boldsymbol{\theta} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho) \\ &= K \int \exp\left\{-\frac{1}{2\delta}(\mathbf{y} - \boldsymbol{\theta})^{\top}(\mathbf{y} - \boldsymbol{\theta})\right\} d\mathbf{y}_{(1)}\, N_m(\boldsymbol{\theta} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho), \end{aligned} \tag{A.2}$$

where $\mathbf{y} = (\mathbf{y}_{(1)}^{\top}, \mathbf{y}_{(2)}^{\top})^{\top}$. By integrating both sides of (A.2) with respect to $\boldsymbol{\theta}$, we get

$$\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} \leq Kg(\sigma_v^2)\, h(\rho) \int N_m(\mathbf{y} \,|\, \mathbf{X}\boldsymbol{\beta}, \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}^{-1})\, d\mathbf{y}_{(1)}. \tag{A.3}$$

Partition $\mathbf{X}$ as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]^{\top}$, where $\mathbf{X}_1^{\top}$ is $m_1 \times p$ and $\mathbf{X}_2^{\top}$ is $m_2 \times p$. We assume that $\operatorname{rank}(\mathbf{X}_2) = p$. Let $\mathbf{d} = (\mathbf{0}_{m_1}^{\top}, \mathbf{y}_{(2)}^{\top})^{\top}$, $\boldsymbol{\phi} = (\mathbf{y}_{(1)}^{\top}, \boldsymbol{\beta}^{\top})^{\top}$ and

$$\mathbf{G} = \begin{bmatrix} -\mathbf{I}_{m_1} & \mathbf{X}_1^{\top} \\ \mathbf{0}_{m_2,\, m_1} & \mathbf{X}_2^{\top} \end{bmatrix}.$$

Then, we can write

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{d} - \mathbf{G}\boldsymbol{\phi},$$

where $\mathbf{G}$ is $m \times (m_1 + p)$, $\boldsymbol{\phi}$ is $(m_1 + p) \times 1$. Hence, (A.3) can be written as

$$\int f\,(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\,\boldsymbol{\theta} \ \le \ Kg\,(\sigma_v^2)\, h(\rho) \int N_m\,(d \mid \mathbf{G}\boldsymbol{\phi}, \delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}^{-1})\, d\mathbf{y}_{(1)}. \tag{A.4}$$

By integrating both sides of (A.4) with respect to $\boldsymbol{\beta}$, we get

$$\int f\,(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\,\boldsymbol{\theta}\, d\,\boldsymbol{\beta} \ \le \ Kg\,(\sigma_v^2)\, h(\rho) \int N_m\,(\mathbf{d} \mid \mathbf{G}\boldsymbol{\phi}, \delta\mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}^{-1})\, d\boldsymbol{\phi}. \tag{A.5}$$

Since $\text{rank}\,(\mathbf{X}_2) = p$, we immediately get that $\text{rank}\,(\mathbf{G}) = m_1 + p$. Thus $\mathbf{G}$ has full column rank. We denote $m_1 + p$ by q. For $k = 2, \ldots, 5$, we now derive upper bounds for

$$\left| \delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}_k^{-1} \right|^{-1/2} \int \exp\left\{ -\frac{1}{2}\, (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^{\top}\, (\delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}_k^{-1})^{-1}\, (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \right\} d\boldsymbol{\phi}$$

which will be integrable with respect to σ_v^2 and ρ as follows.

A.1 Details for the SCAR Model

We first consider the CAR model where $k = 2$. Let $\mathbf{P_W}$ be an orthogonal matrix such that $\mathbf{P_W^{\top}}\, \mathbf{W}\mathbf{P_W} = \text{diag}\{\lambda_i\}_{i=1}^{m} = \boldsymbol{\Lambda}$. Then $\boldsymbol{\Omega}_2\,(\rho)^{-1} = \mathbf{P_W}\,\{\mathbf{I} - \rho\boldsymbol{\Lambda}\}^{-1}\, \mathbf{P_W^{\top}}$, and hence,

$$\begin{aligned}(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^{\top}\, (\delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}_2^{-1})^{-1}\, (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) &= (\mathbf{P_W^{\top}}\, \mathbf{d} - \mathbf{P_W^{\top}}\, \mathbf{G}\boldsymbol{\phi})^{\top} \left(\delta\mathbf{I}_m + \sigma_v^2\, \{\mathbf{I} - \rho\boldsymbol{\Lambda}\}^{-1} \right)^{-1} (\mathbf{P_W^{\top}}\, \mathbf{d} - \mathbf{P_W^{\top}}\, \mathbf{G}\boldsymbol{\phi}) \\ &= (\mathbf{d}_* - \mathbf{G}_*\, \boldsymbol{\phi})^{\top} \left(\delta\mathbf{I}_m + \sigma_v^2\, \{\mathbf{I} - \rho\,\boldsymbol{\Lambda}\}^{-1} \right)^{-1} (\mathbf{d}_* - \mathbf{G}_*\, \boldsymbol{\phi}),\end{aligned}$$

where $\mathbf{d}_* = \mathbf{P_W^{\top}}\, \mathbf{d}$ and $\mathbf{G}_* = \mathbf{P_W^{\top}}\, \mathbf{G}$. Suppose the rows of $\mathbf{G}_*$ corresponding to distinct q indices $\{i_1, \ldots, i_q\} \subseteq \{1, \ldots, m\}$ are linearly independent. We denote these rows by $\mathbf{g}_{i_k^*}^{\top}$, $k = 1, \ldots, q$. Let $\mathbf{A}$ be the $q \times q$ non-singular matrix $[\mathbf{g}_{i_1^*}, \ldots, \mathbf{g}_{i_q^*}]^{\top}$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^{\top} = \mathbf{A}\boldsymbol{\phi}$. Note that

$$(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^{\top}\ (\delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}_2^{-1})^{-1}\, (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \ \ge \ \sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2\, (1 - \rho\, \lambda_{i_k})^{-1}}.$$

From this, we get that

$$\begin{aligned}\int \exp\left\{ -\frac{1}{2}\, (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^{\top} (\delta\mathbf{I}_m + \sigma_v^2\, \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \right\} d\boldsymbol{\phi} &\le \int \exp\left\{ -\frac{1}{2} \sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2\, (1 - \rho\, \lambda_{i_k})^{-1}} \right\} d\boldsymbol{\phi} \\ &= \int \exp\left\{ -\frac{1}{2} \sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2 (1 - \rho\lambda_{i_k})^{-1}} \right\} d\boldsymbol{\eta} \left| \mathbf{A}^{\top}\mathbf{A} \right|^{-1/2} \\ &= K \prod_{k=1}^{q} \left\{ \delta + \sigma_v^2\, (1 - \rho\, \lambda_{i_k})^{-1} \right\}^{1/2},\end{aligned} \tag{A.6}$$

where $K > 0$ is a finite generic constant. Also, we know that

$$\left|\,\delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_2^{-1}\right|^{-1/2} = \prod_{i=1}^{m}\left\{\delta + \sigma_v^2\,(1-\rho\lambda_i)^{-1}\right\}^{-1/2}. \tag{A.7}$$

By (A.6) and (A.7), we get

$$\begin{aligned}\left|\,\delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_2^{-1}\right|^{-1/2} \int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^{\top}(\delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_2^{-1})^{-1}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} \\ \le K \prod_{i\notin\{i_1,\ldots,i_q\}}\left\{\delta + \sigma_v^2\,(1-\rho\lambda_i)^{-1}\right\}^{-1/2} \\ \le K\left\{I\,(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2}\prod_{i\notin\{i_1,\ldots,i_q\}}(1-\rho\,\lambda_i)^{1/2}\,I\,(\sigma_v^2 > N)\right\}\end{aligned} \tag{A.8}$$

for any positive number N. Recall that $\lambda_m^{-1} < \rho < \lambda_1^{-1}$. We know $1-\rho\,\lambda_i$ is an eigenvalue of $\boldsymbol{\Omega}_2$. Thus, for $\lambda_m^{-1} < \rho < \lambda_1^{-1}$, for $i = 1, \ldots, m$, $1-\rho\lambda_i > 0$. Also, $\sum_{i=1}^{m}(1-\rho\lambda_i) = m$. These imply that $0 < 1-\rho\lambda_i < m$. Then from (A.8), we get

$$\begin{aligned}\left|\,\delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_2^{-1}\right|^{-1/2} \int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^{\top}(\delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_2^{-1})^{-1}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} \\ \le K\left\{I\,(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2}\,I\,(\sigma_v^2 > N)\right\}.\end{aligned} \tag{A.9}$$

From (A.5) and (A.9), it follows under the conditions of the theorem that the desired integral $\int f\,(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} d\boldsymbol{\beta} d\sigma_v^2 d\rho$ is finite.

A.2 Details for the SAR Model

We now consider $k = 3$ for the SAR model. With $\mathbf{W}_* = \mathbf{L}^{-1/2}\mathbf{W}\mathbf{L}^{-1/2}$, we have

$$\begin{aligned}\boldsymbol{\Omega}_3 &= (\mathbf{I}_m - \rho\tilde{\mathbf{W}})^{\top}(\mathbf{I}_m - \rho\tilde{\mathbf{W}}) \\ &= (\mathbf{L} - \rho\mathbf{W})^{\top}\mathbf{L}^{-2}(\mathbf{L} - \rho\mathbf{W}) \\ &= \mathbf{L}^{1/2}(\mathbf{I}_m - \rho\mathbf{W}_*)\,\mathbf{L}^{-1}(\mathbf{I}_m - \rho\mathbf{W}_*)\mathbf{L}^{1/2}.\end{aligned}$$

First, $\operatorname{tr}\boldsymbol{\Omega}_3 = m + \rho^2\sum_i\sum_j \tilde{w}_{ij}^2 \le m + \rho^2\sum_i\sum_j \tilde{w}_{ij} = m + \rho^2 m < 2m$ since $0 \le \tilde{w}_{ij} \le 1$, $\sum_j \tilde{w}_{ij} = 1$, and $-1 < \rho < 1$.

Note that the eigenvalues $\nu_1, \ldots, \nu_m$ of $\mathbf{W}_*$ are all real (since $\mathbf{W}_*$ is symmetric). Also, $\mathbf{W}_*$ and $\tilde{\mathbf{W}}$ have identical eigenvalues. Being a stochastic matrix, $\tilde{\mathbf{W}}$ has at least one eigenvalue which is one and the remaining eigenvalues are bounded above by 1, that is $|\nu_i| \le 1$ and $\max_i \nu_i = 1$. As $-1 < \rho < 1$ and $1-\rho\nu_i > 0$, $|\boldsymbol{\Omega}_3| = \prod_{i=1}^{m}(1-\rho\nu_i)^2 > 0$. Thus, the eigenvalues of $\boldsymbol{\Omega}_3$ are positive and bounded above by $2m$. Let $l_{(1)} = \min l_i$ and $l_{(m)} = \max l_i$, where $\mathbf{L} = \operatorname{diag}\{l_i\}_{i=1}^{m}$. Then $l_{(1)} > 0$ and $l_{(m)}$ is bounded above. By writing

$$\boldsymbol{\Sigma}_3 = \delta\,\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_3^{-1} = \mathbf{L}^{-1/2}\left\{\delta\mathbf{L} + \sigma_v^2\,(\mathbf{I}_m - \rho\,\mathbf{W}_*)^{-1}\,\mathbf{L}\,(\mathbf{I} - \rho\,\mathbf{W}_*)^{-1}\right\}\mathbf{L}^{-1/2},$$

we have

$$\begin{aligned} |\mathbf{\Sigma}_3| &= |\mathbf{L}|^{-1} \left| \delta \mathbf{L} + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \mathbf{L} (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \right| \geq |\mathbf{L}|^{-1} l_{(1)}^m \left| \delta \mathbf{I}_m + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-2} \right| \\ &= |\mathbf{L}|^{-1} l_{(1)}^m \prod_{i=1}^m \left\{ \delta + \sigma_v^2 (1 - \rho \nu_i)^{-2} \right\}, \end{aligned} \quad \text{(A.10)}$$

Letting $\mathbf{P}_{\mathbf{W}_*}$ be the matrix of eigenvectors of $\mathbf{W}_*$ such that $\mathbf{P}_{\mathbf{W}_*}^\top \mathbf{W}_* \mathbf{P}_{\mathbf{W}_*} = \text{diag}\{\nu_i\}_{i=1}^m = \mathbf{N}_*,$ we also have

$$\begin{aligned} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top \mathbf{\Sigma}_3^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) &= (\mathbf{L}^{1/2} \mathbf{d} - \mathbf{L}^{1/2} \mathbf{G}\boldsymbol{\phi})^\top \left\{ \delta \mathbf{L} + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \mathbf{L} (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \right\}^{-1} (\mathbf{L}^{1/2} \mathbf{d} - \mathbf{L}^{1/2} \mathbf{G}\boldsymbol{\phi}) \\ &= (\mathbf{r} - \mathbf{S}\boldsymbol{\phi})^\top \left\{ \delta \mathbf{L} + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \mathbf{L} (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \right\}^{-1} (\mathbf{r} - \mathbf{S}\boldsymbol{\phi}) \\ &\geq (l_{(m)}^{-1/2} \mathbf{r} - l_{(m)}^{-1/2} \mathbf{S}\boldsymbol{\phi})^\top \left\{ \delta \mathbf{I}_m + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-2} \right\}^{-1} (l_{(m)}^{-1/2} \mathbf{r} - l_{(m)}^{-1/2} \mathbf{S}\boldsymbol{\phi}) \\ &= (\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi})^\top \left\{ \delta \mathbf{I}_m + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{N}_*)^{-2} \right\}^{-1} (\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi}) \\ &\geq \sum_{k=1}^q \frac{(\tilde{r}_{i_k} - \tilde{s}_{i_k}^\top \boldsymbol{\phi})^2}{\delta + \sigma_v^2 (1 - \rho \nu_{i_k})^{-2}}, \end{aligned} \quad \text{(A.11)}$$

where $\mathbf{r} = \mathbf{L}^{1/2}\mathbf{d},$ $\mathbf{S} = \mathbf{L}^{1/2}\mathbf{G},$ $\tilde{\mathbf{r}} = l_{(m)}^{-1/2} \mathbf{P}_{\mathbf{W}_*} \mathbf{r},$ $\tilde{\mathbf{S}} = l_{(m)}^{-1/2} \mathbf{P}_{\mathbf{W}_*} \mathbf{S},$ and $\{i_1, \ldots, i_q\}$ is a subset of $\{1, \ldots, m\}$ so that the $q \times q$ matrix $[\tilde{\mathbf{s}}_{i_1}, \ldots, \tilde{\mathbf{s}}_{i_q}]^\top = \tilde{\mathbf{S}}_1,$ a submatrix of $\tilde{\mathbf{S}},$ is non-singular. Note that $\tilde{\mathbf{S}}_1$ is determined by $\mathbf{W}.$ Using (A.11) we get

$$\int \exp\left\{ -\frac{1}{2} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top \mathbf{\Sigma}_3^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \right\} d\boldsymbol{\phi} \leq (2\pi)^{q/2} \left| \tilde{\mathbf{S}}_1^\top \tilde{\mathbf{S}}_1 \right|^{-1/2} \prod_{k=1}^q \left\{ \delta + \sigma_v^2 (1 - \rho \nu_{i_k})^{-2} \right\}^{1/2}. \quad \text{(A.12)}$$

Based on (A.10) and (A.12), we get that

$$\begin{aligned} \int |\mathbf{\Sigma}_3|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top \mathbf{\Sigma}_3^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \right\} d\boldsymbol{\phi} &\leq K \prod_{i \notin \{i_1, \ldots, i_q\}} \left\{ \delta + \sigma_v^2 (1 - \rho \nu_i)^{-2} \right\}^{-1/2} \\ &\leq K \left\{ I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} I(\sigma_v^2 > N) \prod_{i \notin \{i_1, \ldots, i_q\}} (1 - \rho \nu_i) \right\} \\ &\leq K \left\{ I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} I(\sigma_v^2 > N) \right\}, \end{aligned} \quad \text{(A.13)}$$

where we use the fact that $-1 < \rho < 1$ and $-1 \leq \nu_i \leq 1$ to claim $0 < 1 - \rho \nu_i < 2.$ From (A.5) and (A.13), it follows by proceeding along the lines we did for the CAR model that the desired integral $\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} d\boldsymbol{\beta} d\sigma_v^2 d\rho$ is finite under the conditions of the theorem.

A.3 Details for the CAR Model

We now consider $k = 4$ for the IAR model where

$$\mathbf{\Omega}_4 = \mathbf{L} - \rho \mathbf{W} = \mathbf{L}^{1/2} (\mathbf{I}_m - \rho \mathbf{W}_*) \mathbf{L}^{1/2}.$$

Let $\mathbf{\Sigma}_4 = \delta \mathbf{I}_m + \sigma_v^2 \mathbf{\Omega}_4^{-1} = \mathbf{L}^{-1/2} \left\{ \delta \mathbf{L} + \sigma_v^2 (\mathbf{I}_m - \rho \mathbf{W}_*)^{-1} \right\} \mathbf{L}^{-1/2}.$ Then

$$|\boldsymbol{\Sigma}_4| \geq |\mathbf{L}|^{-1} k_*^m \prod_{i=1}^m \left\{\delta + \sigma_v^2 (1-\rho\nu_i)^{-1}\right\}, \tag{A.14}$$

where $k_*^m = \min\{l_{(1)}, 1\}$. Proceeding along the same line as in (A.11), we get that

$$(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^{\top} \boldsymbol{\Sigma}_4^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \geq \sum_{k=1}^{q} \frac{(\tilde{r}_{i_k} - \tilde{s}_{i_k}^{\top} \boldsymbol{\phi})^2}{\delta + \sigma_v^2 (1-\rho\nu_{i_k})^{-1}}. \tag{A.15}$$

Again, as we had for the two previous cases, we can use (A.14) and (A.15) to establish that the desired integral is finite under the conditions stated in the theorem.

A.4 Details for the LCAR Model

Finally, we consider $k = 5$, where for the LCAR case we have

$$\boldsymbol{\Omega}_5 = \rho\mathbf{R} + (1-\rho)\,\mathbf{I}_m.$$

Suppose $r_1, \ldots, r_m$ are the eigenvalues of $\mathbf{R}$ and $\mathbf{P_R}$ is an orthogonal matrix such that $\mathbf{P_R^{\top} R P_R} = \operatorname{diag}\{r_i\}_{i=1}^m$. Since $\mathbf{R}$ is a non-negative definite matrix, $r_i \geq 0$, $i = 1, \ldots, m$, and $\sum_{i=1}^m r_i = \operatorname{tr}\mathbf{R} = \sum_{i=1}^m l_i$, implying that $r_1, \ldots, r_m$ are all bounded between 0 and $l = \sum_{i=1}^m l_i$. Then we can write

$$\boldsymbol{\Omega}_5 = \mathbf{P_R}\left\{\operatorname{diag}\{\rho r_i + 1 - \rho\}_{i=1}^m\right\}\mathbf{P_R^{\top}},$$

and claim that for $0 < \rho < 1$, the eigenvalues of $\boldsymbol{\Omega}_5$ are all positive and bounded above by $\sum_{i=1}^m r_i + 1 = l + 1$. Then, with $\tilde{\mathbf{r}} = \mathbf{P_R^{\top} d}$, and $\tilde{\mathbf{S}} = \mathbf{P_R^{\top} G}$, we can establish an inequality similar to (A.11). Note that the nonsingular matrix $\tilde{\mathbf{S}}_1$ is a submatrix of $\tilde{\mathbf{S}}$ and is free from ρ. Boundedness of the eigenvalues of $\boldsymbol{\Omega}_5$ will lead to an inequality similar to (A.14). Finally, we get the desired integral is finite under the conditions of the theorem.

References

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer Science & Business Media.

Besag, J., and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

Brown, D.A., Datta, G.S. and Lazar, N.A. (2017). A Bayesian generalized CAR model for correlated signal detection. *Statistica Sinica*, 27, 1125-1153.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., and Smith, D.D. (2003). On propriety of posterior distributions of variance components in small area estimation. *Journal of Statistical Planning and Inference*, 112, 175-183.

Datta, G.S., Hall, P. and Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106, 362-374.

Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92, 183-196.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

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

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, CRC Press, 3rd ed.

Ghosh, M. (1992). Hierarchical and empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (Eds., M. Ghosh and P.K. Pathak), Institute of Mathematical Statistics, 151-177.

Hodges, J.S. (2019). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Chapman and Hall/CRC.

Leroux, B.G., Lei, X. and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer, 179-191.

MacNab, Y.C. (2003). Hierarchical Bayesian spatial modelling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.

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

Porter, A.T., Wikle, C.K. and Holan, S.H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, 57, 15-29.

Porter, A.T., Holan, S.H., Wikle, C.K. and Cressie, N. (2014). Spatial Fay-Herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10, 27-42.

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

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

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

Speckman, P.L., and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90, 289-302.

Stan Development Team (2018). RStan: The R interface to Stan. R package version 2.17.3.

Sun, D., Tsutakawa, R.K. and Speckman, P.L. (1999). Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika*, 86, 341-350.

Torabi, M. (2012). Hierarchical Bayes estimation of spatial statistics for rates. *Journal of Statistical Planning and Inference*, 142, 358-365.

Trevisani, M., and Gelfand, A. (2013). Spatial misalignment models for small area estimation: A simulation study. In *Advances in Theoretical and Applied Statistics*, Springer, 269-279.

Watanabe, S., and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41, 434-449.

You, Y., and Zhou, Q.M. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37, 1, 25-37. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11445-eng.pdf.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2022.

- C. Bayart, *University Claude Bernard Lyon 1*
- A. Béliveau, *University of Waterloo*
- D. Bonnery, *University of Cambridge*
- H.J. Boonstra, *Statistics Netherlands*
- J. Breidt, *NORC at the University of Chicago*
- J.M. Brick, *Westat Inc.*
- J. Brown, *University of Technology Sydney*
- E. Bueno, *Stockholm University*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- Q. Chen, *Columbia University*
- S. Chen, *University of Oklahoma Health Sciences Center*
- Y. Cheng, *USDA National Agricultural Statistics Service*
- J. Chipperfield, *Australian Bureau of Statistics*
- S. Cho, *Iowa State University*
- R. Clark, *Australian National University*
- M. Cohen, *American Institutes for Research*
- M. Couper, *University of Michigan*
- M. Dagdoug, *Laboratoire de Mathématiques de Besançon*
- G. Datta, *University of Georgia*
- T. De Waal, *Centraal Bureau voor de Statistiek*
- J. Dever, *RTI International*
- M. Elliott, *University of Michigan*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- V. Estevao, *Statistics Canada*
- K. Foster, *East Tennesse State University*
- W.A. Fuller, *Iowa State University*
- K.C.M. Gonçalves, *Universidade Federal do Rio de Janeiro*
- W. Gross, *University of Wollongong*
- P. Han, *University of Michigan*
- S. Hasslet, *Australian National University*
- D. Haziza, *University of Ottawa*
- D. Hedlin, *Stockholm University*
- M.A. Hidiroglou, *Statistics Canada*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
- M. Hyllienmark, *Stockholms universitet Samhallsvetenskapliga Fakulteten*
- D. Judkins, *ABT Associates Inc Bethesda*
- M.G.M. Khan, *University of the South Pacific*
- J.K. Kim, *Iowa State University*
- P.S. Kott, *RTI International*
- T. Kubokawa, *The University of Tokyo*
- P. Lahiri, *University of Maryland*
- R. Lehtonen, *University of Helsinki*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- A. Matei, *Université de Neuchâtel*
- K. McConville, *Reed College*
- F. Mecatti, *University of Milan-Bicocca*
- I. Molina, *Universidad Complutense de Madrid*
- B. Nandram, *Worcester Polytechnic Institute*
- T. Nayak, *The George Washington University*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- L. Rizzo, *WESTAT*
- F.J. Scheuren, *National Opinion Research Center*
- Z. Seeskin, *NORC*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- L. Stokes, *Southern Methodist University Dedman College of Humanities and Sciences*
- S. Sugasawa, *The University of Tokyo*
- N. Thuan, *Oregon Health and Science University*
- Y. Tillé, *University of Neuchatel Faculty of Sciences*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- R. Valliant, *University of Michigan*
- J. van den Brakel, *Statistics Netherlands*
- F. Verret, *Statistics Canada*
- J. Wagner, *University of Michigan*
- L. Wang, *National Cancer Institute*
- H. Wang, *University of Connecticut*
- M. Witt, *Abt Associates*
- C. Wu, *University of Waterloo*
- C. Xu, *The University of Oklahoma Health Sciences Center*
- D. Yang, *US Bureau of Labor Statistics*
- C. Yu, *Iowa State University*
- W. Yung, *Statistics Canada*
- L.-C. Zhang, *University of Southampton*
- P. Zhao, *Yunnan University*

Acknowledgements are also due to those who assisted during the production of the 2022 issues: Céline Ethier of International Cooperation and Methodology Innovation Centre; Patrick O'Leary of Social Statistics Methods Division; the team from Dissemination Division, in particular: Sophie Pilon, Eric Chenier, Camil Séguin, Kathy Charbonneau, Travis Robinson and Karo-Lynn Audy as well as our partners in the Communications Division

ANNOUNCEMENTS

Nominations Sought for the 2024 Waksberg Award

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

The recipient of the Waksberg Award will receive an honorarium and give the 2024 Waksberg Invited Address at the Statistics Canada Symposium, expected to be held in the autumn of 2024. The paper will be published in an upcoming issue of *Survey Methodology* (Targeted for December 2024).

The author of the 2024 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*. **Nomination of individuals to be considered should be sent by email before February 15, 2023 to the chair of the committee, Maria Giovanna Ranalli (maria.ranalli@unipg.it). Nominations should include a CV and a nomination letter.** Nominations will remain active for 5 years.

Members of the Waksberg Paper Selection Committee (2022-2023)

Maria Giovanna Ranalli, *University of Perugia* (Chair)
Denise Silva, *Brazilian Institute of Geography and Statistics*
Jae-Kwang Kim, *Iowa State University*
Kristen Olson, *University of Nebraska-Lincoln*

Past Chairs:

Graham Kalton (1999-2001)
Chris Skinner (2001-2002)
David A. Binder (2002-2003)
J. Michael Brick (2003-2004)
David R. Bellhouse (2004-2005)
Gordon Brackstone (2005-2006)
Sharon Lohr (2006-2007)
Robert Groves (2007-2008)
Leyla Mojadjer (2008-2009)
Daniel Kasprzyk (2009-2010)
Elizabeth A. Martin (2010-2011)
Mary E. Thompson (2011-2012)
Steve Heeringa (2012-2013)
Cynthia Clark (2013-2014)
Louis-Paul Rivest (2014-2015)
Tommy Wright (2015-2016)
Kirk Wolter (2016-2017)
Danny Pfeffermann (2017-2018)
Michael A. Hidiroglou (2018-2019)
Robert E. Fay (2019-2020)
Jean Opsomer (2020-2021)
Jack Gambino (2021-2022)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 38, No. 2, June 2022

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 38, No. 3, September 2022

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 50, No. 1, March/mars 2022

Special Issue
Functional and object data analysis/L’analyse de données fonctionnelles et d’objets

Introduction

Research Articles

Erratum

Acknowledgement

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 50, No. 2, June/juin 2022

Research Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* as a guide and note particularly the points below. Articles must be submitted in Word or Latex, preferably in Word with MathType for the mathematical expressions. A pdf version is also required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract and Introduction

2.1 The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

2.2 The last paragraph of the introduction should contain a brief description of each section.

3. Style

3.1 Avoid footnotes and abbreviations.

3.2 Limit the use of acronyms. If an acronym is used, it must be defined the first time it occurs in the paper.

3.3 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.

3.4 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in Section 4.

3.5 Bold fonts should normally be used to distinguish vectors and matrices from scalars.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the top of tables or figures. Use a two-level numbering system based on the section of the paper. For example, Table 3.1 is the first table in Section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors’ names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The first time a reference is cited in the text, the name of all authors must be written. For subsequent occurrences, the names of all authors can again be written. However, if the reference contains three or more authors, the names of the second and subsequent authors can be replaced with “et al.”.

5.3 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words, including tables, figures and references.