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

Survey Methodology

Survey Methodology
50-2

Release date: December 20, 2024

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 2024

Volume 50

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 50, Number 2, December 2024

Contents

Waksberg invited paper series

Regular papers

Short note

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 147-148
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

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

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2026 Waksberg Award.

This issue of *Survey Methodology* opens with the 24th paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Maria Giovanna Ranalli (Chair), Denise Silva, Jae-Kwang Kim and Kristen Olson for having selected Richard Valliant as the author of 2024 Waksberg Award paper.

2024 Waksberg Invited Paper

Author: Richard Valliant

Richard Valliant is a Research Professor Emeritus at the University of Michigan and the Joint Program for Survey Methodology at the University of Maryland. He received his PhD from Johns Hopkins University in Biostatistics and an MS in statistics from Cornell University. He has over 45 years of experience in survey sampling, estimation theory, and statistical computing. He was formerly an Associate Director at Westat and a mathematical statistician with the Bureau of Labor Statistics (BLS). He has a range of applied experience in survey estimation and sample design on a variety of establishment, institutional, and household surveys, including the Consumer Price Index, the Current Population Survey, and other surveys done by BLS, the National Center for Education Statistics, the Consumer Product Safety Commission, the Department of Energy, and the National Agricultural Statistical Service among others. He is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute. He was an associate editor of the *Journal of the American Statistical Association–Theory and Methods Section* (1989-1993) and the *Applications and Case Studies Section* (1996-1999), *Journal of Official Statistics* (2003-2010), and *Survey Methodology* (1996-2007).

He is the co-author of three books: *Finite Population Sampling and Inference: A Prediction Approach* (2000) with A. Dorfman and R.M. Royall; *Survey Weights: A Step-by-step Guide to Calculation* (2018) with J.A. Dever; *Practical Tools for Designing and Weighting Survey Samples*, (2018, 2nd edition) with J.A. Dever and F. Kreuter. The first edition of the *Practical Tools* book was the winner of the 2020 Book Award from the American Association for Public Opinion Research. He is also the author of the R packages: PracTools and svydiags.

Waksberg Award honorees and their invited papers since 2001

2025 Michael A. **Hidiroglou**, Manuscript in preparation expected for the December 2025 issue.

2024 Richard **Valliant**, "Sample design using models". *Survey Methodology*, vol. 50, 2, 149-183.

2023 Raymond L. **Chambers**, "The missing information principle – A paradigm for analysis of messy sample survey data". *Survey Methodology*, vol. 49, 2, 219-256.

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.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 149-183
Statistics Canada, Catalogue No. 12-001-X

Sample design using models

Richard Valliant[1]

Abstract

Joseph Waksberg was an important figure in survey statistics mainly through his applied work in the design of samples. He took a design-based approach to sample design by emphasizing uses of randomization with the goal of creating estimators with good design-based properties. Since his time on the scene, advances have been made in the use of models to construct designs and in software to implement elaborate designs. This paper reviews uses of models in balanced sampling, cutoff samples, stratification using models, multistage sampling, and mathematical programming for determining sample sizes and allocations.

Key Words: Anticipated variances; Balanced samples; Cutoff samples; Design-based; Mathematical programming; Model-assisted; Model-based.

1. Introduction

Joseph Waksberg importantly influenced the practice of survey sampling and official statistics in several ways. At the US Census Bureau in the 1950-1970s, he led early studies of recall error, coverage error, small area estimation, sampling rare populations, advancements in household sampling including use of address lists, rotation of sample areas, use of administrative data, and improvements in telephone sampling methods. The interview with him by David Morganstein and David Marker in *Statistical Science* covers many of the areas he contributed to (Morganstein and Marker, 2000).

He mainly worked on sample design issues, but his thinking was not limited to mathematical considerations. Depending on the application he adapted methods to account for practicalities. In the early 1960's he and Neter studied telescoping in a consumer expenditure survey (Neter and Waksberg, 1964). Although response errors in expenditure surveys were a known problem (e.g., see Cole and Utting, 1956; Ferber, 1955), it had not often been studied directly. Neter and Waksberg (1964) conducted an experiment sponsored by the US Census Bureau to study the tendency of persons to misreport the time period when expenditures occurred. Large expenditures, in particular, were often reported to have occurred nearer to the present than when they actually occurred, i.e., they were *telescoped* forward. Based on their findings, they were the first to propose *bounded recall* as a potential solution. In the second or later interview in a continuing survey the respondent is told the expenditures that had been reported in the previous interview then asked for the additional expenditures since then.

Faulty data used in designing a sample was another topic he studied. When he became the head statistician on the Current Population Survey (CPS) in the early 1960's, the area probability methods were well established. But, the survey had to face new problems caused by the expanding American economy. The migration to the suburbs from cities was in full swing and data from the 1960 census was becoming progressively more stale. Maps being used for fieldwork were out-dated, and some area segments that had

1. Richard Valliant, Research Prof. Emeritus, Universities of Michigan & Maryland U.S.A. E-mail: valliant@umich.edu.

a few farm houses in the census were found with major housing developments built on them. Such fast growing neighborhoods led to bad measures of size based on the last census, which, in turn, led to intolerably expensive workloads if the original sampling plan was implemented. This led to his instituting the use of building permit samples to identify new construction in advance and avoid such "surprise" sample segments.

Coverage errors were recognized problems for censuses and surveys on which he also led research. In the 1960s decade, while he was head of the CPS, that survey and others at the Census Bureau introduced address-list sampling as a way of reducing the number of households inadvertently omitted by field listers. Their method for compiling an address list began with purchasing one from the Donnelley Corporation. As Waksberg explained in Morganstein and Marker (2000), "the post office had the mailing addresses in little slots. Dummy mailing pieces were prepared for all addresses on the Donnelley list and the postal carriers put the mail into these little slots and checked for missing addresses, filling out a card for each missing address." With this method plus some special procedures, like checking buildings that had been converted into apartments but with no apartment number designated, they compiled a more complete list to use for sampling within selected areas. This kind of inventiveness was characteristic of the way that he, Morris Hansen, and colleagues at Census solved practical problems.

Figure 1.1 Joseph Waksberg ca. 1998.

The Mitofsky-Waksberg (MW) method of random digit dialing (Waksberg, 1978) was another solution to a practical problem. In the early 1970s unrestricted random sampling of telephone numbers in the US was extremely inefficient for household sampling since about 80% of 10-digit phone numbers were assigned to businesses, institutions, government agencies, or were unassigned. The MW method treated the first eight digits in the sorted list of phone numbers as clusters (known as 100-blocks), screened clusters by phoning a randomly selected number in a sample 100-block, and retaining a cluster only if the contacted number was

residential. In a retained cluster additional 2-digit numbers were appended to the 8-digit cluster number and phoned to obtain the desired sample size. The MW method does not require knowledge of either the first- or second-stage selection probabilities, but it does produce an equal probability sample of telephone numbers. Because a high percentage of 100-blocks had no residential numbers, MW sampling is substantially more cost-efficient than unrestricted random sampling. This is another example of his very practical approach to sampling: given a specific problem, he devised a clever solution specific to the application.

One of his most important contributions to the field was training dozens of young statisticians. On-the-job training was his main way of doing this rather than formal teaching, as anyone who was fortunate enough to work with him can attest. He was adept at reducing complicated technical issues to intuitive, understandable explanations, which was especially valuable to clients and novices learning the field. One lesson that he emphasized was the importance for a sampling statistician to think not only about the specific questions that are asked, but also the broader aspects of these questions: whether the questions make sense and can be solved, or whether they should be modified or changed.

His approach to sample design was randomization-based with repeated sampling properties being paramount. The explicit use of models has gradually become a part of survey design and estimation over the years since Waksberg and his colleagues at the Census Bureau and Westat were at work. Their use of auxiliary data in sample design certainly has some model flavor, but they rarely, if ever, formally appealed to models for their work. Models have, of course, played a central role in the allied field of experimental design for many years (e.g., see Box, Hunter and Hunter, 2005; Wu and Hamada, 2021). This paper reviews some of the more explicit use of models to guide design for finite population samples in the last few decades. Section 2 reviews balanced sampling motivated by models. Cutoff sampling, discussed in Section 3, is sometimes used in establishment surveys when large units account for the bulk of a population total. Formation of strata using models is commonly used in business surveys and is covered in Section 4. Multistage sampling using models to estimate variance components is described in Section 5. Mathematical programming in Section 6 is very useful for finding efficient allocations in multipurpose surveys. Finally, Section 7 is a summary.

2. Balanced samples

Practitioners have long used systematic sampling from lists sorted by auxiliaries available on a sampling frame as a way of exercising control over the distribution of a selected sample. This technique is especially useful when there are several auxiliary variables (x's) on a frame, but the sample size is too small to allow all x's to be crossed to form separate strata. For example, a frame of schools might be stratified by geographic area and sorted within stratum by enrollment size as a way of controlling the sample distribution by area and size of school. A frame of city blocks can be numbered in a serpentine fashion so that blocks that are near each other in the serial numbering are also geographically close (Hansen, Hurwitz and Madow, 1953a; U.S. Census Bureau, 2006). A frame of hospitals might be stratified by number of emergency room

visits and sorted geographically within strata (Schroeder and Ault, 2001). Business establishments can be stratified by geography and industry code and then sorted by employment size within strata. Selecting systematically from each sorted list will, in expectation, produce a sample with a type of *balance* that depends on whether the sample is selected with equal probability or with probabilities proportional to a measure of size. The concept of *balanced samples* has been formalized by various authors as described below.

Balanced sampling was presented in the 1970s as a method of protecting against prediction bias (Royall and Herson, 1973a, b) using the model-based approach to sampling. For notation, let s denote the set of sample elements; U the population of N elements; n is the sample size; y_i is an analysis variable for element i; and x_i is an auxiliary variable known for every element in the population. As an example, take the ratio estimator, $\hat{\bar{y}}_R = \bar{y}_s\left(\bar{x}_U/\bar{x}_s\right)$ where $\bar{y}_s = \sum_{i\in s} y_i/n$, $\bar{x}_s = \sum_{i\in s} x_i/n$, and $\bar{x}_U = \sum_{i\in U} x_i/N$. The ratio estimator is the best linear unbiased (BLU) predictor of the mean, $\bar{y}_U = \sum_{i\in U} y_i/N$, under a model with mean $E_M\left(y_i\right) = \beta x_i$ and variance $V_M\left(y_i\right) = \sigma^2 x_i$. However, if the model mean is $E_M\left(y_i\right) = \alpha + \beta x_i$, $\hat{\bar{y}}_R$ has a model-bias (or prediction-bias), defined as $E_M\left(\hat{\bar{y}}_R - \bar{y}_U\right)$, equal to $\alpha\left(\bar{x}_U/\bar{x}_s - 1\right)$. Thus, if the model has an intercept rather than being through the origin, the ratio estimator has a model-bias. This bias is zero in any sample that is balanced in the sense that $\bar{x}_s = \bar{x}_U$. This result extends to more complicated cases where, e.g., the correct model is polynomial rather than straight-line with an intercept (Valliant, Dorfman and Royall, 2000, Section 3.1). Under a simple random sampling (srs) design, $E_\pi\left(\bar{x}_s - \bar{x}_U\right) = 0$, where E_π is the expectation with respect to repeated sampling, and $\bar{x}_s - \bar{x}_U \xrightarrow{p} 0$ in large srs's, under some standard conditions on how the population and sample grow as n and N get large. If there are other covariates, **z**, that should be in the model for y, srs does on average balance on their means also, even though the **z**'s may be unknown at the time of sampling. These results extend to other probability sampling designs that yield design-unbiased or asymptotically design-unbiased estimators of N and $\bar{x}_U$.

A key requirement for the model-based calculations above is that the sample not be *informative* in the sense defined by Pfeffermann and Sverchkov (2009). A sample is informative if the model that fits in a sample is different from the one that fits the population even after accounting for covariates. In symbols, being informative means that $f\left(y_i \mid x_i, i\in s\right) \neq f\left(y_i \mid x_i, i\in U\right)$ where $f(\cdot)$ is a density. Informativeness can be caused, for example, by the sampling itself or by response to the sample that depends on y.

The fact that the sample mean of x does aim at its population mean provides an argument for why srs or other probability sampling designs can be useful methods of sample selection to protect against unknown biases. These properties might suggest that using a probability design removes the need to worry about model-bias. Nonetheless, if the point estimator has a model-bias, the model-bias squared may not diminish fast enough to become a negligible part of the model mean squared error – reinforcing the notion that correct modeling is critical. In the example above, the model-bias squared of the ratio estimator in an srs design is $O_P\left(n^{-1}\right)$, where O_p is order with respect to the probability sampling design, and so is the model-variance. Royall and Cumberland (1985) illustrated that in srs a predictable percentage of samples will be poorly

balanced on any auxiliary regardless of how large the sample is. In those badly balanced samples, even confidence intervals constructed with a variance estimator, like the jackknife, that is robust to some types of model misspecification will have poor coverage. Thus, to eliminate worries about model-bias, a sampling plan is needed that reduces the order of the bias-squared faster than does srs. Kott (1986) did just that by showing that equal probability, systematic sampling from an ordered list was a way of achieving balance faster than the srs rate.

These model-bias results plus the cosmetic appeal of using "nicely distributed" samples provide an impetus for restricting samples at the design stage to ones that are in some sense balanced. Royall (1992) generalized the idea of balanced sampling to linear models of the form

$$E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}; \qquad V_M(\mathbf{Y}) = \mathbf{V}\sigma^2 \tag{2.1}$$

where $\mathbf{Y}$ is an N-vector of analysis variables, $\mathbf{X}$ is an $N \times p$ matrix of covariates, and $\mathbf{V} = \text{diag}\{v_i\}_{i \in U}$ is an $N \times N$ diagonal covariance matrix. Model (2.1) is reasonably flexible since covariates can be interactions or transformations of auxiliaries.

The error variance or prediction variance of an estimator, $\hat{\theta}$, of a population quantity, θ_U, is $E_M\left(\hat{\theta} - \theta_U\right)^2$. For specificity, we consider estimators, $\hat{t}$ of the population total, $t_U = \sum_U y_i$. The BLU predictor of t_U is

$$\hat{t}_{\text{BLU}} = \sum_s y_i + \sum_{U-s} \hat{y}_i = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} \hat{e}_{Mi} \tag{2.2}$$

where $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, $\hat{e}_{Mi} = y_i - \hat{y}_i$, $\mathbf{x}_i^T$ is the ith row of $\mathbf{X}$, and $\hat{\boldsymbol{\beta}} = \left(\sum_s \mathbf{x}_i \mathbf{x}_i^T / v_i\right)^{-1} \sum_s \mathbf{x}_i y_i / v_i$ is the generalized least squares estimator of $\boldsymbol{\beta}$.

Define $\mathbf{1}_N$ to be an N-vector of all 1's and $\mathbf{1}_s$ to be an n-vector of 1's. Then, when both $\mathbf{V}\mathbf{1}_N$ and $\mathbf{V}^{1/2}\mathbf{1}_N$ are in the column space of $\mathbf{X}$, the sample that yields the minimum error variance for the BLU predictor is a weighted-balanced sample that satisfies

$$\frac{1}{n}\mathbf{1}_s^T \mathbf{V}_s^{-1/2} \mathbf{X}_s = \frac{\mathbf{1}_N^T \mathbf{X}}{\mathbf{1}_N^T \mathbf{V}^{1/2} \mathbf{1}_\mathbf{N}} \quad \text{or, equivalently} \quad \frac{1}{n}\sum_s \frac{\mathbf{x}_i}{v_i^{1/2}} = \frac{\bar{\mathbf{x}}_U}{\bar{v}_U^{(1/2)}} \tag{2.3}$$

where $\mathbf{V}_s$ is the $n \times n$ covariance matrix for the n sample units, $\mathbf{X}_s$ is the $n \times p$ matrix of covariates for the sample units, and $\bar{v}_U^{(1/2)} = N^{-1}\sum_{i \in U} v_i^{1/2}$. Identifying a set of elements that satisfies (2.3) in advance of sampling requires a frame with the individual x_i's and v_i's. If the latter depend on some function of the x's, it may be possible to derive them. If the v_i are all equal, then (2.3) reduces to simple balance, $\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_U$.

With a weighted-balanced sample, the BLU predictor of the total of the y's reduces to

$$\hat{t}_{\text{BLU}} = \frac{N\bar{v}_U^{(1/2)}}{n} \sum_{i \in s} \frac{y_i}{v_i^{1/2}} \tag{2.4}$$

See Valliant et al. (2000, Theorem 4.2.1) for details. Notice that (2.3) depends on y only through the covariance matrix $\mathbf{V}$ but the reduction to (2.4) requires that $v_i^{1/2}$ and v_i be linear combinations of the

columns of $\mathbf{X}$. Thus, if different y's have this same structure, weighted balance will be optimal for them also. Tam (1995) extended the idea of balance to clustered populations where elements within clusters have a model correlation. His results seem more difficult to implement in practice because the balance on auxiliary variables must account for the intracluster correlations which will depend on the y.

Model-assisted sample design uses both a model and random sample selection for analyses. A main tool in this approach is an anticipated variance (AV) of the form,

$$\mathrm{AV}\left(\hat{t}\right) = E_M E_\pi\left[\left(\hat{t} - t_U\right)^2\right] - \left[E_M E_\pi\left(\hat{t} - t_U\right)\right]^2.$$

When $\hat{t}$ is π-unbiased, i.e., $E_\pi(\hat{t} - t_U) = 0$, the anticipated variance reduces to $E_M V_\pi(\hat{t} - t_U)$. The optimality of weighted balanced sampling is closely related to earlier AV results on unequal probability sampling. The reduced form of the BLU predictor in (2.4) is equal to the π-estimator when the sample is selected with probabilities proportional to $v_i^{1/2}$. Godambe and Joshi (1965) and Isaki and Fuller (1982) presented circumstances where the anticipated variance of a regression estimator of the population mean is minimized when selection probabilities are proportional to the square root of a model variance. A key assumption is that the model errors are uncorrelated.

In the context of probability sampling, Deville and Tillé (2004) and Fuller (2009) give methods that restrict random samples to ones where weighted sample means of auxiliary variables are close to the corresponding population means, i.e.,

$$N^{-1}\sum_s \frac{\mathbf{x}_i}{\pi_i} \doteq \bar{\mathbf{x}}_U \tag{2.5}$$

where π_i is the selection probability of element i in a probability sample. (Also, see Ardilly, Haziza, Lavallée and Tillé (2024)). An estimator of the form on the left-hand side of (2.5) is generally called the π-estimator (Särndal, Swensson and Wretman, 1992). The Deville-Tillé method selects probability samples directly that approximately satisfy (2.5); Fuller's method rejects probability samples where (2.5) is not satisfied within a specified tolerance. In the terminology of Cumberland and Royall (1981) and Royall (1992), samples that satisfy (2.5) are π-balanced. Deville and Tillé (2004, 2005) cover calculation of weights and variance estimators for probability samples balanced using what they term the "cube" method that have a design-based interpretation. Nedyalkova and Tillé (2008) generalized the Godambe-Joshi and Fuller-Isaki results to show that an optimal model-assisted strategy (i.e., one that minimizes the AV) for the π estimator under model (2.1) is to select a fixed size, π-balanced sample on the x's in the model. A fixed size sample can be achieved by including $\mathbf{x}_i = \pi_i$ in the balance conditions.

The R package `sampling` (Tillé and Matei, 2023) will select weighted or unweighted balanced samples that satisfy either (2.3) or (2.5). If a probability sample is designed so that $\pi_i = nv_i^{1/2}/(N\bar{v}_U^{(1/2)})$, the result will be both model and design optimal, at least for the key y variable used to assign the selection probabilities. A variance estimator like the jackknife that has both good design and model properties can then be used. If $V_M\left(y_i\right) \neq v_i$ for some y's, then a probability sample selected with $\pi_i = nv_i^{1/2}\Big/\left(N\bar{v}_U^{(1/2)}\right)$

may be somewhat model inefficient for those y's but will still allow for design-unbiased or consistent estimation.

The French Institut National de la Statistique et des Études Économiques (INSEE) has used the cube method to select a master sample of PSUs (Costa, Guillo, Paliod, Merly-Alpa, Vincent, Chevalier and Deroyon, 2018) and in its census to select samples of municipalities with equal probabilities that are balanced on a set of demographic variables (see Deville and Tillé, 2004). The INSEE application has a key feature that many applications lack: municipalities cannot be nonrespondents in the census. In cases where units can nonrespond, the initial balance of a sample may be lost, which, at best, is a nuisance, and, at worst, leads to biases. In principle, substituting for a nonrespondent with an element with the same $\mathbf{x}_i$ value will preserve model unbiasedness and optimality under (2.1). However, this would perturb design properties because a substitution is an imputation that adds variance and, possibly, bias.

Restricting the geographic configuration of first-stage sample units has long been a desire when designing area samples (Kish, 1965, Section 12.8) and is related to balancing. Having the sample PSUs spread over a map of the universe is especially pleasing in area probability samples. In addition, there may be a number of potential stratification variables like population density, education level, and concentration of ethnic groups whose use could improve efficiency of estimators but cannot be fully used because of a limited sample size – much like the cases of systematic sampling noted above. Goodman and Kish (1950) proposed a one-PSU per stratum method called *controlled selection* geared toward these types of restrictions. This method and others, like Latin squares (Frankel and Stock, 1942), restrict the configurations of sample units and assign a probability of selection to each allowable configuration (Hansen et al., 1953a, Section 11.4). Although these methods can achieve sample restriction, the variance estimators that were in use at that time did not reflect gains in precision (if any) due to the restricted sampling.

Grafström, Lundström and Schelin (2012) and Grafström (2012) introduced other methods that control the spread of a sample over a population. Instead of a mapping, these methods use distance between units to create small joint inclusion probabilities for nearby units, forcing the samples to be well spread. Grafström and Tillé (2013) proposed a method that is doubly balanced in the sense of selecting samples that are balanced on a number of auxiliary variables and also are well spread for other variables like topographical coordinates. Grafström and Tillé (2013) used a linear model of the form $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$. The model errors are assumed to have the covariance structure, $\text{cov}_\xi(\varepsilon_i, \varepsilon_j) = \sigma_i\sigma_j\rho^{d_{ij}}$ where d_{ij} is a measure of distance between elements i and j and $0 < \rho < 1$. As a result, the correlation diminishes as elements get farther apart. In area sampling, a latitude/longitude centroid is often associated with each first-stage unit and can be used for computing the distance between any pair of units.

All of the balancing methods are available in the R package `BalancedSampling` (Grafström, Lisic and Prentius, 2023). Since the balancing methods are probability samples with known single and joint selection probabilities, standard design-based variance estimators can be used. Practical limitations often mean that exactly balanced samples cannot be selected. In such cases or ones where additional x's are found to be predictive of the analysis variables, either general regression (GREG) or purely model-based

estimators of means and totals can be used along with variance estimators in Valliant et al. (2000); Valliant (2002) or Särndal et al. (1992).

The discussion above uses linear models, so the question naturally arises whether any balancing results extend to nonlinear models like $E_M(y_i) = \mu(\mathbf{x}_i, \boldsymbol{\theta})$ with $V_M(y_i) = v_i(\boldsymbol{\theta})$ where $\mu(\mathbf{x}_i, \boldsymbol{\theta})$ is a function of a vector of covariates and an unknown parameter $\boldsymbol{\theta}$. The mean, $\mu(\mathbf{x}_i, \boldsymbol{\theta})$, can be linear or nonlinear in $\mathbf{x}_i$ so that quantitative or categorical y's are covered. A model-assisted estimator of a total, designed for this nonlinear model, is (Breidt and Opsomer, 2009)

$$\hat{t}_{\text{MA}} = \sum_U \hat{\mu}_i + \sum_s \frac{e_{\text{MA}i}}{\pi_i}$$

where $e_{\text{MA}i} = y_i - \hat{\mu}_i$. Although similar in form to $\hat{t}_{\text{BLU}}$ in (2.2), no balancing results are available for $\hat{t}_{\text{MA}}$ or for the model-calibrated estimator of Wu and Sitter (2001), which is also based on a nonlinear model.

3. Cutoff samples

In applications where a few units account for an inordinate share of population totals, the standard approach is to include the large units in the sample with certainty and select a random sample from the remainder of the population. A more extreme method is to select a cutoff sample. Cutoff samples are ones in which only elements with a specified characteristic are sampled. The cutoffs are often quantitative like amounts of revenue in business establishments or production levels of oil refineries. If estimates are desired for a full population, they can be justified if (a) the sample and nonsample units follow the same superpopulation model or (b) there is some randomness to the cutoff so that the propensity of being included in the sample can be modeled based on element-level covariates. Cutoff samples can also be considered as special cases of stratified designs that have take-none, take-some, and take-all strata as discussed in Section 4.

Such samples are even mentioned in Hansen et al. (1953a, pages 486-490), who note that this design can be effective in establishment populations where a small number of large units account for a large percentage of the totals being estimated and where collecting data from small units would be uneconomical. Restricting a sample in some way may be unavoidable if some members of a target population are inaccessible. For example, if data collection must be done by telephone, non-telephone households are excluded; institutionalized persons (e.g., incarcerated or in nursing homes) may be excluded from household surveys because of difficulties in collecting their data. If estimates are needed for the whole population, a critical requirement for justifying (a) above is that predictions for the nonsample units can be made from the sample units. This can be done when the same model holds for the sample as for the nonsample. Inclusion in the cutoff sample must also be ignorable, i.e., it cannot depend directly on the y's to be analyzed.

In some applications, a nonrandom cutoff sample will be model-optimal for estimating a total. For example, consider the ratio model, $E_M(y_i) = \beta x_i$ and $V_M(y_i) = \sigma^2 x_i$ with the y's being independent. Assume that the

goal is to estimate the total of the y's for the full population. Values for nonsample units are predicted as $\hat{\beta}x_i$, and the error variance of the ratio estimator is $V_M\left(\hat{\bar{y}}_R - \bar{y}_U\right) = \left(1-n/N\right)\sigma^2\bar{x}_r\bar{x}_U/\left(n\bar{x}_s\right)$ where $\bar{x}_r = \left(N-n\right)^{-1}\sum_{i\in U-s}x_i$. In this case, the optimal sample design that minimizes the error variance is nonrandom and consists of taking the n units with the largest values of x. If y is a current period value and x is a census value for the same variable at a previous time period and economic conditions have not changed radically from the census, the ratio model may fit well.

In extremely skewed populations where the largest units account for a high percentage of the population total, making poor predictions for smaller units is less of a worry. However, this type of cutoff sampling is risky because it eliminates the possibility of testing how well the model fits for smaller units. If domain estimates for small and medium sized units are required, a cutoff sample should not be used because of the risk that the domains follow a different model from the one that fits the cutoff sample. Another worry is misclassification. If a large unit that should be in the cutoff sample is erroneously classified as part of the take-none part of the population, the cutoff sample can exclude an important contributor to the population total. Model breakdown over time is also a concern. In volatile economic times, a model may fit well for a while but fail when a recession or other downturn occurs. As partial protection against this, Benedetti, Bee and Espa (2010) extended the cutoff idea by stratifying the population into three strata – take-all, take-some, and take-none – and developed an algorithm for dividing a population into the strata and allocating the sample to them.

Another concern is nonresponse. If one or more extremely large units do not cooperate, adjusting for that nonresponse can be difficult in either cutoff samples or more conventional samples. If a large unit is unique, weighting up the respondents or imputing for a nonrespondent may not be a workable solution. For example, in an agricultural survey of crop production if farms owned by a large agribusiness refuse to provide data, a big part of production of corn, wheat, etc. will be missing. Values for respondents may have limited use as sources for imputation in such cases. Having large, nonresponding units that were sampled with certainty may scuttle the goal of estimating for the entire population unless a good method of imputing for them can be devised. In the US, a Census of Agriculture is conducted every five years of all farms and ranches. The census data may be useful for imputing missing crop production data in noncensus-year sample surveys provided all large units respond to the census. Other surveys of business establishments or other institutions may not have this luxury, though.

Yorgason, Bridgman, Cheng, Dorfman, Lent, Liu, Miranda and Rumburg (2011) review some applications of cutoff sampling by US federal government agencies. In particular, the Energy Information Administration (EIA) of the US Department of Energy conducts monthly surveys of crude oil and natural gas producers using cutoff samples that cover at least 85 percent of the total oil and gas production of each state (U.S. Energy Information Administration, 2018). Production for nonsample companies is implicitly imputed using a ratio estimator. EIA also surveys electric utilities each month using similar cutoff samples of large producers (Kirkendall, 1992; Knaub, 2008). Rapid changes in the energy economy in 2008 illustrate the riskiness of cutoff samples. Per Yorgason et al. (2011, page 3): "When petroleum and natural gas prices

began to rise rapidly in 2008, the large (in-sample) natural gas well operators increased their production rates faster than did the smaller (non-sample) operators. In addition to the production incentive created by rising prices, technological advances allowed some large companies to increase their shale gas extraction rates. The actual EIA-914 sample coverage rates increased, and the estimated coverage rates, based on prior data, failed to reflect the changes quickly enough. As a result, EIA overestimated natural gas production for some states." Industry analysts claimed that the overstated production estimates had artificially deflated market prices for natural gas.

Haziza, Chauvet and Deville (2010) cite a Statistics Canada tax data example from Fecteau and Jocelyn (2005). Unincorporated Canadian businesses may declare their financial statement either on paper or electronically. Owing to the high costs of converting data collected on paper to an electronic format, the paper filers are deliberately excluded from possible selection in the sample. Population estimates are based on a sample selected from the electronic-filers only. In this tax example, if it is reasonable to conceive of the situation as one where there is a probability of filing a return either via paper or electronically, the propensity of being an electronic filer can be estimated based on covariates. (Note the covariates must not include the variable that determines whether an element is in the cutoff sample or not.) Then, inverse propensity weighting of electronic filers can be used for estimation as is sometimes done in nonprobability samples (e.g., see Elliott and Valliant, 2017). If there is zero probability that an electronic filer would file via paper, inverse propensity weighting would fail, but knowing that in a given situation seems dubious. This would be similar to having hardcore nonrespondents who have a zero probability of responding.

4. Stratification using models

Suppose the population is divided into $h = 1, \ldots, H$ strata with N_h elements in stratum h. The population of elements in stratum h is denoted by U_h. The proportion of units in stratum h is $W_h = N_h / N$, and the population mean of y is $\bar{y}_U = \sum_{h=1}^{H} W_h \bar{y}_{Uh}$ where $\bar{y}_{Uh}$ is the population mean of y within stratum h. For design-based analyses assume that a simple random sample is selected without replacement in each stratum (stsrswor). The sample size in stratum h is n_h, the set of sample units in h is s_h, and the total sample size is $n = \sum_{h=1}^{H} n_h$. An estimator of the population mean is then $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{sh}$ where $\bar{y}_{sh}$ is the sample mean in stratum h. The design variance of $\bar{y}_{st}$ is $V_\pi(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(n_h^{-1} - N_h^{-1}\right) S_{yUh}^2$ where S_{yUh}^2 is the population variance of y in stratum h. The design relvariance of $\bar{y}_{st}$ is defined as $V_\pi(\bar{y}_{st}) / \bar{y}_U^2$; the coefficient of variation (CV) of $\bar{y}_{st}$ is the square root of the relvariance. Basic design questions are how best to form the strata and how to allocate the sample to the strata.

Models are most useful in forming strata in populations where single-stage sampling, like stsrswor, is used. These include populations of business establishments, schools, hospitals, or other institutions. In some samples, strata may be dictated by publication goals in which models may have limited use. For example, in a business survey, separate statistics may be needed for retail, wholesale, manufacturing, and other sectors. However, within a sector a model might be used to form substrata using methods described below.

When a y variable is related to a single, quantitative auxiliary x, known for all elements in a population, a model can be used to guide the formation of strata. This line of reasoning is often used in business or institution surveys and is referred as stratification by size. There is a large literature on how to form strata; e.g., see Rivest (2002) and his list of references. The usual formulation is to sort the sampling frame by x, divide the population into strata, and determine an optimal allocation of an stsrswor to the strata. With that scenario the goal is to find stratum boundaries, $(b_{h-1}, b_h]$ for $h = 1, \ldots, H-1$, that lead to the design-variance of the π -estimator of a mean or total being minimized or to n being minimized subject to a target CV.

4.1 Model-assisted analyses

Lavallée and Hidiroglou (1988) present iterative algorithms for finding strata boundaries that will minimize the total sample size subject to a constraint on the coefficient of variation (CV) of $\bar{y}_{st}$, assuming that the allocation to strata is a power allocation (Bankier, 1988). In a power allocation, the proportion of the total sample size allocated to strata h is proportional to $\left(W_h \bar{y}_{Uh}\right)^p$ for $p \in (0,1]$.

Rivest (2002) extended the algorithm of Lavallée and Hidiroglou (1988) when either (i) $\log(y) = \alpha + \beta_{\log} x + \varepsilon$ where ε is normally distributed with mean 0 and a constant variance that does not depend on x, or (ii) $y = \beta_{\text{lin}} x + \varepsilon$ with ε having mean 0 and variance $\sigma^2_{\text{lin}} x^\gamma$ where γ is non-negative. Rivest gave results for both power and Neyman allocations. Rivest assigns stratum H to be a take-all (or certainty) stratum – a procedure often used in business surveys for large units – so that $n_H = N_H$. For $h < H$ the sample size in stratum h can be written as $\left(n - n_H\right) a_h$ where n is the total sample size and $a_h = \left(W_h \bar{y}_{Uh}\right)^p \Big/ \sum_{h=1}^{H-1} \left(W_h \bar{y}_{Uh}\right)^p$ for a power allocation and $a_h = W_h S_{yUh} \Big/ \sum_{h=1}^{H-1} W_h S_{yUh}$ for Neyman allocation. Solving $V_\pi\left(\bar{y}_{st}\right)$ for n and using variances conditional on stratum membership leads to

$$n = NW_H + \frac{\sum_{h=1}^{H-1} W_h^2 V_M\left(y \mid b_h \geq x \geq b_{h-1}\right) \Big/ a_{hX}}{\bar{y}_U^2 c^2 + \sum_{h=1}^{H-1} W_h^2 V_M\left(y \mid b_h \geq x \geq b_{h-1}\right) \Big/ N} \tag{4.1}$$

where V_M denotes model-variance, $\bar{y}_U$ is the population mean of y, c^2 is the target relvariance for $\bar{y}_{st}$, and a_{hX} is written using the model relationship between y and x. Expression (4.1) leads to differential equations for $\partial n / \partial b_h$ and $\partial n / \partial b_{h-1}$, which are solved iteratively using an algorithm of Sethi (1963). A practical limitation of the Lavallée-Hidiroglou (LH) and Rivest iterative algorithms is that they may converge to boundaries that do not give the true minimum of n or may not converge at all for some configurations of y (Slanta and Krenzke, 1994, 1996; Rivest, 2002).

Gunning and Horgan (2004) and Horgan (2006) presented another algorithm for finding the stratum boundaries, $(b_{h-1}, b_h]$, based on a single, skewed measure of size x. Their solution was to compute the stratum boundaries as $b_h = b_0 r^h$ $(h = 1, \ldots, H)$ where $r = \left(b_H / b_0\right)^{1/H}$ with b_H and b_0 being the maximum and minimum values of x. That is, the boundaries follow a geometric progression. If the distribution of x is uniform within each stratum, this set of boundaries approximately equalizes the stratum coefficients of variation of x defined as $S_{xh} / \bar{x}_h$, where S_{xh} is the standard deviation among units in the frame for stratum

h and $\bar{x}_h$ is the frame mean for the stratum h units. The algorithm is not motivated by a model but was competitive with the LH algorithm and the $\text{cum}(\sqrt{f})$ method of Dalenius and Hodges (1959) and is computationally easier to implement.

Baillargeon and Rivest (2009) extended Rivest (2002) to account for stratum-specific non-response rates and allow take-none, tale-some, and take-all strata. Excluding some units from the sample via a take-none stratum may be reasonable when some units have y values near zero or are so small relative to the large units that they contribute little to a population total. (The extreme case of this is cutoff sampling in Section 3.) A take-none stratum can be a way of reducing the relative root mean square error of $\bar{y}_{st}$ but does lead to it being biased. They allocate the sample to the strata using a general rule that features proportional, Neyman, and power allocations as special cases. As noted in Baillargeon and Rivest (2009), their solution for n requires an iterative solution for which they use an algorithm due to Kozak (2004). When non-response is accounted for in the take-all stratum or when there is a take-none stratum, an unconstrained solution can be negative. Thus, a constrained minimization for n is found over the boundaries $\{b_h\}$ that give a positive sample size.

The R package `stratification` (Baillargeon and Rivest, 2011) implements several methods of stratification, including $\text{cum}\sqrt{f}$, geometric, and LH. The LH algorithm that finds stratum boundaries that minimize the total sample size n while achieving a target CV can be implemented with either the Sethi or Kozak algorithms. With Kozak's algorithm the package will also find stratum boundaries that minimize the CV of $\bar{y}_{st}$ for a fixed sample size n rather than minimizing n for a predetermined CV.

4.2 Purely model-based analyses

Dorfman and Valliant (2000) investigated the model-based properties of stratification by size from a purely model-based point-of-view. Some of their results are summarized here. When a common model holds for the entire population as in (2.1) and $\mathbf{V}\mathbf{1}_N$ and $\mathbf{V}^{1/2}\mathbf{1}_N$ are in the column space of $\mathbf{X}$, the BLU predictor with a weighted balanced sample is the best strategy, as noted in Section 2. Stratification by size is then, at best, a mechanism for selecting a weighted balanced sample. Nonetheless, further model-based analyses will illuminate the justification for different variations on stratification by size that are sometimes used in practice.

First, consider a stratified version of model (2.1) in which parameters may depend on strata:

$$E_M\left(\mathbf{Y}_h\right)=\mathbf{X}_h\boldsymbol{\beta}_h;\quad V_M\left(\mathbf{Y}_h\right)=\sigma_h^2\mathbf{V}_h;\quad h=1,\ldots,H \tag{4.2}$$

where $\mathbf{Y}_h$ is $N_h\times 1$, $\mathbf{X}_h$ is $N_h\times p_h$, σ_h^2 is a positive scalar, $\mathbf{V}_h=\text{diag}\left(v_{hi}\right)$ is $N_h\times N_h$, and $\boldsymbol{\beta}_h$ is a $p_h\times 1$ parameter vector. The BLU predictor is then the sum of the BLU predictors in each stratum.

In stratum h define a weighted balanced sample to be one that satisfies

$$\frac{1}{n_h}\mathbf{1}_{sh}^T\mathbf{V}_{sh}^{-1/2}\mathbf{X}_{sh}\;=\;\frac{\mathbf{1}_{Nh}^T\mathbf{X}_h}{\mathbf{1}_{Nh}^T\mathbf{V}_h^{1/2}\mathbf{1}_{Nh}} \tag{4.3}$$

where $\mathbf{1}_{sh}$ is a vector of n_h 1's, $\mathbf{1}_{Nh}$ is a vector of N_h 1's, $\mathbf{V}_{sh}$ is the $n_h \times n_h$ diagonal covariance matrix for the sample units, and $\mathbf{X}_{sh}$ is the $n_h \times p_h$ matrix of auxiliaries for the sample units. Any stratum sample satisfying (4.3) will be denoted by $B(\mathbf{X}_h\colon \mathbf{V}_h)$, and, when (4.3) is satisfied in each stratum, the entire sample is a stratified weighted balanced sample.

If both $\mathbf{V}_h\mathbf{1}_{Nh}$ and $\mathbf{V}_h^{1/2}\mathbf{1}_{Nh}$ are in the column space of $\mathbf{X}_h$, then the BLU predictor achieves its minimum variance when each stratum sample is $B(\mathbf{X}_h\colon \mathbf{V}_h)$. In that case, the BLU predictor reduces to

$$\hat{t}_{\text{BLU}} = \sum_{h=1}^{H} N_h \bar{v}_h^{(1/2)} \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{v_{hi}^{(1/2)}} \tag{4.4}$$

and the error variance is

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \sum_h \left[\frac{1}{n_h}\left(N_h \bar{v}_h^{(1/2)}\right)^2 - N_h \bar{v}_h\right] \sigma_h^2 \tag{4.5}$$

where $\bar{v}_h^{(1/2)} = N_h^{-1}\sum_{U_h} v_{hi}^{1/2}$ and $\bar{v}_h = N_h^{-1}\sum_{U_h} v_{hi}$.

Thus, in a stratified weighted balanced sample, the optimal estimator reduces to a sum of mean-of-ratios estimators. As in the unstratified case, a weighted balanced sample is the best that can be selected in the sense of making the error variance of the BLU predictor small. The estimator of the total, $\hat{t}_{\text{BLU}}$, is also the π-estimator when each within-stratum sample is selected with probabilities proportional to $v_{hi}^{1/2}$. Although a probability sample selected with probabilities proportional to $v_{hi}^{1/2}$ is balanced in expectation, the model-based result does not require that the balanced sample be obtained via probability sampling. However, if a probability sample is desired, the methods of Deville and Tillé (2004) can be used.

The optimal, cost-constrained allocation to strata can be computed using standard methods. Assume that the cost function is $C = C_0 + \sum_h c_h n_h$ where C is the total cost, C_0 is the cost that does not vary with sample size, and c_h is the cost per unit in stratum h. Suppose that $\mathbf{V}_h\mathbf{1}_{Nh}$ and $\mathbf{V}_h^{1/2}\mathbf{1}_{Nh}$ are in the column space of $\mathbf{X}_h$ and that a weighted balanced sample, $B(\mathbf{X}_h : \mathbf{V}_h)$, is selected in each stratum. In that case, the allocation that minimizes the variance of $\hat{t}_{\text{BLU}}$ subject to a fixed total cost is

$$\frac{n_h}{n} = \frac{N_h \bar{v}_h^{(1/2)} \sigma_h \big/ \sqrt{c_h}}{\sum_{h'} N_{h'} \bar{v}_{h'}^{(1/2)} \sigma_{h'} \big/ \sqrt{c_{h'}}}. \tag{4.6}$$

When all costs are equal, the BLU with the optimal, balanced sample allocation is then equal to

$$\hat{t}_{\text{BLU}} = \frac{1}{n}\left(\sum_h N_h \bar{v}_h^{(1/2)} \sigma_h\right) \sum_h \sum_{s_h} \frac{y_{hi}}{v_h^{1/2}\sigma_h}. \tag{4.7}$$

and its error variance is

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{1}{n}\left(\sum_h N_h \bar{v}_h^{(1/2)} \sigma_h\right)^2 - \sum_h N_h \bar{v}_h \sigma_h^2. \tag{4.8}$$

To investigate how to form strata, take the case of a single model fitting the entire population, i.e., the special case of (2.1) and (4.2) defined by

$$E_M\left(\mathbf{Y}_h\right) = \mathbf{X}_h\boldsymbol{\beta}, V_M\left(\mathbf{Y}_h\right) = \mathbf{V}_h\sigma^2. \tag{4.9}$$

Suppose we select a stratified weighted balanced sample and use the optimal allocation given in (4.6) for the equal cost case. Using (4.7) with $\sigma_h = \sigma$ and $\overline{v}^{(1/2)} = N^{-1}\sum_h\sum_{U_h} v_{hi}^{1/2}$, the BLU predictor with the optimal allocation is

$$\hat{t}_{\text{BLU}} = \frac{1}{n}\left(\sum_h N_h\overline{v}_h^{(1/2)}\right)\sum_h\sum_{s_h}\frac{y_{hi}}{v_{hi}^{1/2}} = \frac{1}{n}\left(N\overline{v}^{1/2}\right)\sum_h\sum_{s_h}\frac{y_{hi}}{v_{hi}^{1/2}}$$

which is the form of the BLU in (2.4) in a weighted balanced sample for an unstratified sample. In other words, stratification with optimal allocation of a stratified weighted balanced sample here gains nothing at all compared to the strategy of selecting an unstratified sample with overall weighted balance.

An important special case of a single population model occurs when there is a single x and the model is polynomial:

$$E_M\left(y_{hi}\right) = \delta_0\beta_0 + \delta_1\beta_1 x_{hi} + \cdots + \delta_J\beta_J x_{hi}^J, V_M\left(y_{hi}\right) = \sigma^2 x_{hi}^\gamma$$

where $\delta_j = 1$ if the jth order term is in the model and 0 if not; and the β_j's are regression parameters. Among the models in this class is the one for the ratio estimator: $E_M\left(y_{hi}\right) = \beta_1 x_{hi}, V_M\left(y_{hi}\right) = \sigma^2 x_{hi}$. With the variance specification, $V_M\left(y_{hi}\right) = \sigma^2 x_{hi}^\gamma$, the optimal, cost-constrained allocation is given by specializing (4.6):

$$\frac{n_h}{n} = \frac{N_h\overline{x}_h^{(\gamma/2)}\Big/\sqrt{c_h}}{\sum_{h'}N_{h'}\,\overline{x}_{h'}^{(\gamma/2)}\Big/\sqrt{c_{h'}}}.$$

where $\overline{x}_h^{(\gamma/2)} = N_h^{-1}\sum_{U_h} x_{hi}^{\gamma/2}$. The error variance with that allocation is

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \sigma^2\sum_h\left[\frac{1}{n_h}\left(N_h\,\overline{x}_h^{(\gamma/2)}\right)^2 - N_h\,\overline{x}_h^{(\gamma)}\right]. \tag{4.10}$$

The problem of how to create strata is most conveniently studied when an equal number of sample units, $n_h \equiv n_0$, is allocated to each stratum. In that case,

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n_0}\sum_h\left(N_h\,\overline{x}_h^{(\gamma/2)}\right)^2 - N\overline{x}^{(\gamma)}\,\sigma^2 \tag{4.11}$$

with $\overline{x}^{(\gamma)} = N^{-1}\sum_h\sum_{U_h} x_{hi}^\gamma$.

Let $Z_h = N_h\overline{x}_h^{(\gamma/2)}$. Optimal stratification occurs when the leading term in (4.11), $\sum_h Z_h^2 = \sum_h\left(N_h\,\overline{x}_h^{(\gamma/2)}\right)^2$, is minimized. Adding and subtracting $\sigma^2 H\overline{Z}^2/n_0$, where $\overline{Z} = \sum_{h=1}^H Z_h/H$, gives

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n_0} S_Z^2 + \frac{\sigma^2}{n_0} \frac{\left(N\bar{x}^{(\gamma/2)}\right)^2}{H} - \sigma^2 N\bar{x}^{(\gamma)} \tag{4.12}$$

where $S_Z^2 = \sum_h \left(Z_h - \bar{Z}\right)^2$. The one term in (4.12) that depends on the formation of the strata is the first, which is eliminated by making the Z_h all equal. Expression (4.12) then becomes $V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n}\left(N\bar{x}^{(\gamma/2)}\right)^2 - \sigma^2 N\bar{x}^{(\gamma)}$.

Equalizing $Z_h = N_h \bar{x}_h^{(\gamma/2)}$ leads to several "equal aggregate size" rules for forming strata found in the literature, for example, Cochran (1977, page 172), Godfrey, Roshwalb and Wright (1984), and Hansen et al. (1953a, page 382). When $\gamma = 0$, equal values of $N_h \bar{x}_h^{\gamma/2}$ correspond to equal numbers of units N_h in each stratum. When $\gamma = 1$, we have equal aggregate square root of size, and $\gamma = 2$ gives equal aggregate x. Thus, model-based analyses clarify when the different methods of stratification by size will be efficient.

The results in subsections 4.1 and 4.2 refer to a single y and a pre-determined number of strata H, but extensions without those restrictions have been done. For practical reasons, a single set of strata is usually used with the understanding that it will not be equally efficient for all estimates. Consequently, a compromise set of strata and an allocation are needed that does reasonably well for different estimates while adhering to budget, workload, and other constraints on the design. Mathematical programming, discussed in Section 6, is especially useful in that regard. Benedetti, Espa and Lafratta (2008) and Ballin and Barcaroli (2013) address the problem of design for multipurpose surveys, i.e., ones with multiple y's, using tree and genetic algorithms. Their solutions identify an optimal set of strata based on crossing a set of categorical covariates and an allocation to those strata. The total number of strata H is a byproduct of their solutions. The Ballin and Barcaroli (2013) algorithms are implemented in the `SamplingStrata` R package (Barcaroli, Ballin, Odendaal, Pagliuca, Willighagen and Zardetto, 2022).

5. Multistage sampling and anticipated variances

For decades multistage sampling has been a standard tool in household surveys that require in-person data collection. A nested sequence of geographic areas is selected until, at the last stage, households or persons within households are sampled. Multistage sampling is also common in education surveys where schools and students within schools are the stages and in business surveys where establishments then employees are sampled. To design these surveys, estimates of variance components are needed. Anticipated variances can be useful to avoid the problem of negative variance estimates as described below.

5.1 Two-stage designs

Take the case of a two-stage sample in which primary sampling units (PSUs) are selected with varying probabilities and with replacement (ppswr) while the second-stage elements are selected by simple random sampling without replacement. With-replacement designs may not often be used in practice but have simple variance formulae which facilitate sample size calculation. Let y_k be the value of an analysis variable for

element k, m be the number of sample PSUs, M the number of PSUs in the population, s the set of sample PSUs, N_i the number of elements in the population for PSU i, n_i the number of sample elements in sample PSU i, and s_i the set of sample elements in PSU i. The "probability with replacement" pwr-estimator of the total of the y's is

$$\hat{t}_{\text{pwr}} = \frac{1}{m}\sum_{i\in s}\frac{\hat{t}_i}{p_i}$$

where $\hat{t}_i = \frac{N_i}{n_i}\sum_{k\in s_i} y_k$ is the estimated total for PSU i from a simple random sample, and p_i is the one-draw selection probability of PSU i. The design variance of $\hat{t}_{\text{pwr}}$ from Cochran (1977, pages 308-310) is

$$V_\pi\left(\hat{t}_{\text{pwr}}\right) = \frac{1}{m}\sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i\in U}\frac{N_i^2}{mp_in_i}\left(1-\frac{n_i}{N_i}\right)S_{U2i}^2 \tag{5.1}$$

where t_U is the population total of y and $S_{U2i}^2 = \left(N_i-1\right)^{-1}\sum_{k\in s_i}\left(y_k-\bar{y}_{Ui}\right)^2$ with $\bar{y}_{Ui}$ being the population mean of y in PSU i.

Computing sample sizes when the n_i are allowed to vary is difficult, but, to control workloads, samples are often designed to select the same number of elements in each PSU. Making the assumption that $\bar{n}$ elements are selected in each PSU and that the within-PSU sampling fraction, $\bar{n}/N_i$, is negligible, the design relvariance of $\hat{t}_{\text{pwr}}$, defined as $V_\pi\left(\hat{t}_{\text{pwr}}\right)/t_U^2$, is approximately (Valliant, Dever and Kreuter, 2018, Section 9.2.4):

$$\frac{V_\pi\left(\hat{t}_{\text{pwr}}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}}\,k\left[1+\delta\left(\bar{n}-1\right)\right] \tag{5.2}$$

where

$$B^2 = S_{U1(\text{pwr})}^2/t_U^2,\ S_{U1(\text{pwr})}^2 = \sum_{i\in U} p_i\left(\tfrac{t_i}{p_i}-t_U\right)^2,$$

t_i is the population total of y for PSU i,

$$W^2 = t_U^{-2}\sum_{i\in U} N_i^2\,\tfrac{S_{U2i}^2}{p_i},\ k = (B^2+W^2)/\tilde{V},$$

and $\delta = B^2/\left(B^2+W^2\right)$ is a measure of homogeneity. The unit (i.e., population) relvariance is $\tilde{V} = S_U^2/\bar{y}_U^2$ with S_U^2 being the population variance of y and $\bar{y}_U$ the population mean of y.

The design-based variance component estimator of B^2 found in, e.g., Särndal et al. (1992), can be negative, depending on the configuration of the data. Using anticipated variances permits the relvariance of the pwr-estimator to be written in terms of model variance components. The model components can be estimated using algorithms that avoid the numerical problems that the basic design-based, analysis of variance formulas have. Examples in the literature of using model-based variance component estimates in survey design seem limited, even though practitioners often employ the technique. A few examples are

Chromy and Myers (2001), Hunter, Bowman and Chromy (2005), Judkins and Van de Kerckhove (2003), Valliant and Gentle (1997), and Waksberg, Sperry, Judkins and Smith (1993). Searle, Casella and McCulloch (1992) review the methods that are available, including minimum variance quadratic unbiased estimation (MIVQUE0), maximum likelihood, and restricted maximum likelihood (REML). The use of anticipated variances also clarifies the key role, noted below, that PSU and SSU sizes have in determining measures of homogeneity.

As noted in Section 2, when the estimator is design-unbiased or approximately so, i.e., $E_\pi(\hat{t}_{\text{pwr}}) \doteq t_U$, the anticipated variance is $\text{AV}\left(\hat{t}_{\text{pwr}}\right) = E_M\left[V_\pi\left(\hat{t}_{\text{pwr}} - t_U\right)\right]$. Thus, the model expectation E_M of a formula like (5.2) can be computed, giving model variance components that can be estimated using standard software. In a clustered population, the simplest model to consider is one with common mean, μ, and random effects for clusters, α_i, and elements, ε_{ij}:

$$y_k = \mu + \alpha_i + \varepsilon_{ik},\ k \in U_i, \tag{5.3}$$

with $\alpha_i \sim (0, \sigma_\alpha^2)$, $\varepsilon_{ik} \sim (0, \sigma_\varepsilon^2)$, and the errors being independent. This model is overly simple, but results can be extended to a case with $E_M(y_k) = \mathbf{x}_k^T\boldsymbol{\beta}$.

The case of sampling clusters with probability proportional to the size N_i is of particular practical importance, i.e., $p_i = N_i/(M\bar{N})$ where $\bar{N} = \sum_{i\in U} N_i/M$ is the average number of elements per cluster. In that special case, after some algebra, the model expectations of B^2 and W^2 are

$$\begin{aligned} E_M\left(B^2\right) &= \frac{1}{\mu^2}\left\{\sigma_\alpha^2\left[1 - \frac{1}{M^2}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \frac{\sigma_\varepsilon^2}{\bar{N}}\right\} \\ E_M\left(W^2\right) &= \sigma_\varepsilon^2/\mu^2, \end{aligned}$$

$\nu_N^2 = \sum_{i\in U}\left(N_i - \bar{N}\right)^2 \Big/ \left[\left(M-1\right)\bar{N}^2\right]$ is the relvariance of the cluster sizes. When M is large, the anticipated measure of homogeneity is approximately

$$E_M(\delta) \doteq \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}}{\sigma_\alpha^2 + \sigma_\varepsilon^2\left(1 + 1/\bar{N}\right)} \doteq \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}. \tag{5.4}$$

Expression (5.4) is the correlation under model (5.3) of any two elements in the same cluster. Note that there is no assumption that all PSUs have the same size $\left(N_i \equiv \bar{N}\right)$ to obtain (5.4). As long as M is large, ν_N^2 has a limited effect on B^2 and δ. This contrasts to the result when both stages are selected via srs where the variation in PSU sizes plays an important role in determining δ even when M is large (see Valliant et al. (2018, equation (9.43)), Valliant, Dever and Kreuter (2015)).

5.2 Three-stage designs

Maze (2021) has extended the analysis above to three-stage sampling where both secondary sampling units (SSUs) and third stage elements are stratified. Consider a three-stage design in which the stages are PSUs, SSUs, and housing units (HUs). Using HUs as the third stage units is illustrative. The formulation below also applies to other applications. Suppose that m PSUs are selected with ppswr, the SSUs are

stratified within each PSU and n_{ia} are selected with ppswr within PSU i, SSU stratum a. Housing units are stratified within each SSU and q_{iajb} are selected via simple random sampling without replacement (srswor) from the population total of Q_{iajb} HUs within PSU i, SSU j in SSU stratum $a\,(a=1,\ldots,A)$, and HU substratum $b\,(b=1,\ldots,B)$. Assume that SSU strata are defined the same in each PSU and that HU strata definitions are the same in every PSU/SSU.

The SSU strata might be defined based on the percentage of the SSU population in some domain (e.g., Hispanics) that is important to the survey. HU strata could be defined by the race-ethnic group of the head-of-household. For example, the University of Michigan's Health and Retirement Study (HRS, https://hrs.isr.umich.edu/about) is a longitudinal panel study of persons aged 50 and older, supported by the US National Institute on Aging and the Social Security Administration. Its PSUs are counties or groups of counties and its SSUs are census blocks or groups of blocks. SSUs are stratified by the concentration of African Americans and Hispanics. Periodically, HRS recruits a new age cohort of persons who become age-eligible. In 2016 the "Late Baby Boom" cohort (birth years 1960-65) was recruited with HUs stratified by the race-ethnicity of the head of household as coded on a commercial list of housing units (Valliant, Hubbard, Lee and Chang, 2014). The availability of commercial lists in the US and household panels like NORC's AmeriSpeak panel at the University of Chicago

(see https://amerispeak.norc.org/us/en/amerispeak/about-amerispeak/panel-design.html)

make such targeted samples feasible for non-governmental organizations. In other countries, population registries allow government agencies to implement similar designs.

The pwr estimator of a population total, t_U, of y's is

$$\hat{t}_{\text{pwr}} = \frac{1}{m}\sum_{i\in s_1}\frac{1}{p_i}\sum_{a=1}^{A}\frac{1}{n_{ia}}\sum_{j\in s_{ia}}\frac{1}{p_{j|ia}}\sum_{b=1}^{B}\frac{Q_{iajb}}{q_{iajb}}\sum_{k\in s_{iajb}} y_k \tag{5.5}$$

where p_i is the 1-draw probability of selecting PSU i; s_1 is the set of sample PSUs; $p_{j|ia}$ is the conditional 1-draw probability of selecting SSU j within PSU i, SSU stratum a; s_{ia} is the set of sample SSUs in PSU i, SSU stratum a; s_{iajb} is the set of sample HUs in PSU i, SSU j in SSU stratum a and HU stratum b.

To simplify design-based variance and sample size calculations, a standard workaround is to suppose that the same number of SSUs, $\bar{n}_a$, is selected in each PSU and SSU stratum and that the same number of HUs, $\bar{\bar{q}}_{ab}$ is sampled in each $iajb$ combination. Also, define U to be the universe of PSUs, U_{ia} the universe of SSUs in ia, U_{iajb} the universe of HUs in $iajb$, $K_a = t_{Ua}/t_U$ to be the proportion of the population total of y that is in SSU stratum a, and $K_{ab} = t_{Uab}/t_U$ to be the proportion in the strata combination ab. After some calculation, the design relvariance of the estimator can be written as the sum of three terms, which are similar to those in Hansen, Hurwitz and Madow (1953b, Chapter 9):

$$\frac{V_\pi\left(\hat{t}_{\text{pwr}}\right)}{t_U^2} = \frac{B^2}{m} + \sum_{a=1}^{A} K_a^2\frac{W_{2a}^2}{m\bar{n}_a} + \sum_{a=1}^{A}\sum_{b=1}^{B} K_{ab}^2\frac{W_{3ab}^2}{m\bar{n}_a\,\bar{\bar{q}}_{ab}} \tag{5.6}$$

where $B^2 = \frac{S_{U1(\text{pwr})}^2}{t_U^2}$,

$$W_{2a}^2 = \frac{1}{t_{U_a}^2}\sum_{i\in U}\frac{S_{U2(\text{pwr})ia}^2}{p_i},\ W_{3ab}^2 = \frac{1}{t_{U_{ab}}^2}\sum_{i\in U}\frac{1}{p_i}\sum_{j\in U_{ia}}\frac{Q_{iajb}^2 S_{U3iajb}^2}{p_{j|ia}}$$

with

$$S_{U1(\text{pwr})}^2 = \sum_{i\in U} p_i\left(\frac{t_{U_i}}{p_i} - t_U\right)^2,\ S_{U2(\text{pwr})ia}^2 = \sum_{j\in U_{ia}} p_{j|ia}\left(\frac{t_{U_{iaj}}}{p_{j|ia}} - t_{U_{ia}}\right)^2,$$

and

$$S_{U3iajb}^2 = \frac{1}{Q_{iajb}-1}\sum_{k\in U_{iajb}}\left(y_k - \overline{y}_{U_{iajb}}\right)^2.$$

The population totals, t_{Ui} and t_{Uiaj}, are for the y's in PSU i and SSU iaj; $\overline{y}_{U_{iajb}}$ is the mean per element in $iajb$. This assumes that the sampling fraction at the third stage is negligible. Details of the derivation are in Maze (2021, Section 2.3). The coefficient of variation of $\hat{t}_{\text{pwr}}$ is

$$\text{CV}_\pi\left(\hat{t}_{\text{pwr}}\right) = \sqrt{V_\pi\left(\hat{t}_{\text{pwr}}\right)/\, t_U^2}.$$

For a differentiable, nonlinear function, $\hat{\theta}_{\text{pwr}}$, like a ratio mean, a linear approximation to $\hat{\theta}_{\text{pwr}}$ can be made and a formula analogous to (5.6) derived. A complication not dealt with here is that some PSUs may be certainties (i.e., selected with probability 1). The relvariance in (5.6) is then split between certainties and non-certainties with no PSU variance component used for the certainties. The non-certainty PSUs are also typically stratified by geography or other characteristics. The extension to these other complications is straightforward.

A well-known limitation of (5.6) is that estimates of B^2 and W_{2a}^2 involve subtractions so that the estimates can be negative in some samples. As in two-stage sampling, anticipated variances can be used to circumvent this problem. Existing routines for estimating model variance components can then be used that constrain parameter estimates to be positive.

In the three-stage case, a simple model for y_k has a common mean, μ, and random effects α_i for PSUs, γ_{iaj} for SSUs, and ε_{iajbk} for HUs in SSU/HU substratum ab:

$$y_k = \mu + \alpha_i + \gamma_{iaj} + \varepsilon_{iajbk}$$

with

$$\alpha_i \sim \left(0, \sigma_\alpha^2\right),\ \gamma_{iaj} \sim \left(0, \sigma_\gamma^2\right),\ \varepsilon_{iajbk} \sim \left(0, \sigma_{\varepsilon_{ab}}^2\right)$$

and the errors being independent, such that $V_M\left(y_k\right) = \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_{\varepsilon_{ab}}^2$ and $E_M\left(y_k\right) = \mu$ for $k \in U_{iajb}$. Extensions to models where $E_M(y_k) = \mathbf{x}_k^T\boldsymbol{\beta}$ are possible.

6. Mathematical programming solutions for sample allocations

Most national surveys of households, establishments, and institutions have multiple goals. Separate estimates for demographic groups or types of business are often desired. These may be implemented as target sample sizes for subgroups or target coefficients of variation for estimates. In addition, there are

usually constraints on the budget, workload assignments for data collectors, maximum number of attempts to contact a sample unit for cooperation, among other things. One way to determine a sample allocation to strata, PSUs, SSUs, and other stages of sampling is trial and error. By trying enough combinations, a designer may iteratively find an allocation that meets most goals. A more formal and accurate way of finding an allocation is mathematical programming (MP), which can be applied with either model-based or design-based calculations. MP is an extremely useful technique for finding allocations in complex problems where it is impossible to obtain a direct, closed-form solution.

Methods for finding approximate solutions have been developed in the field of operations research. Although MP does appear in the literature for sample allocation, it seems underused in practice. Some of the strata formation methods in Section 4 can be considered MP algorithms. Bethel (1989) gave a specialized nonlinear algorithm for some stratified allocation problems. Ballin and Barcaroli (2008) extended Bethel's algorithm to the tasks of creating strata for stsrs and finding an efficient allocation. Hughes and Rao (1979) covered optimum allocation to strata with multiple constraints. Chromy (1987) presented a specialized algorithm for finding sample sizes that minimize cost subject to various constraints. Valliant and Gentle (1997) described allocation in a two-stage sample with smoothed anticipated variances used for components and with constraints on sample sizes and CVs of a set of estimators. Choudhry, Rao and Hidiroglou (2012) use nonlinear programming in stsrs to solve allocation problems with constraints on CV's of stratum and domain estimators for domains that cut across strata. More recently, de Moura Brito, Silva, Semaan and Maculan (2015) examined stsrs allocations using integer programming. Valliant et al. (2018, Chapter 5) give an introduction to MP along with a number of examples of the use of MP for sample allocation. An especially intricate application is allocating outlets and items for pricing in the US Consumer Price Index (Gomes and Johnson, 2016; Leaver and Solk, 2005). This section gives two examples – one simple and the other more complex – of using MP to determine multicriteria allocations.

Solving MP problems requires specialized software that implements the sophisticated algorithms developed in operations research. A shortcoming of the early papers is that they were not accompanied by publicly available software – that is no longer the case. Schwendinger and Borchers (2023) give a long list of R packages that have optimization functions. The `alabama` (Varadhan, 2023) and `nloptr` (Ypma, Johnson, Borchers, Eddelbuettel, Ripley, Hornik, Chiquet, Adler, Dai, Stamm and Ooms, 2022) packages in R, in particular, solve the types of nonlinear problems needed for sample allocation. Also useful are the procedures `NLP` and `OPTMODEL` in SAS® (http://www.sas.com) and the Solver add-on (http://www.solver.com) that comes bundled with Microsoft Excel® – the latter especially so because of its user-friendly interface.

Valliant et al. (2014) give a fairly simple application of MP to finding sample sizes using data from the 2011-12 US National Survey of Family Growth (NSFG). We sketch the application here; more details are in the paper. Using a national probability sample of households, the NSFG gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health for persons age 15-44 (Groves, Mosher, Lepkowski and Kirgis, 2009). The goal of the example here is to obtain

target sample sizes for one age group using an imperfect commercial list of addresses for sampling households. In the US, survey organizations can purchase these address lists from private vendors. In European and other countries with population registries, governmental agencies may have access to better lists with extensive auxiliary data.

One approach to obtaining a target sample size for a particular age (or other demographic) group is to select an equal probability sample of HUs, roster all persons in the HU, and retain all or a subset of any that are in the desired group. With no advance information about the HUs, this may be the only option. The drawback to this approach that is many HUs may have to be screened especially when the target domain is small. An alternative, less expensive tactic is to use an address list that eliminates some screening by pinpointing HUs that are likely to be in the desired group. Even when the list is not altogether accurate, this may be more efficient than equal probability sampling.

In the NSFG screener, a roster of all persons is collected from each responding HU along with limited demographic data. In particular, the age of each person is obtained and is the variable used to define a domain in this example. Although the target age range in NSFG is 15-44, the age of each person in the HU is obtained during screening. The 2011-12 NSFG respondents' addresses were sent to a commercial list vendor for matching. The NSFG screener data were considered to be correct and could then be compared to demographic data on the commercial list to assess the list's accuracy in classifying people by age and in designing future editions of the survey.

In this illustration, we want to obtain a target sample size in the 65+ age group by stratifying HUs using age data from the commercial list. To formulate the problem, define the following notation:

d = target age domain (65+);

h = sampling stratum based on commercial list information on individual addresses; $h = 1, \ldots, 4$ as given in Table 6.1;

$p_h(d)$ = proportion of HUs in sampling stratum h that have at least 1 person in the age 65+ domain based on NSFG;

$a_h(d)$ = average number of persons per HU in sampling stratum h that are actually in domain d based on NSFG data; this average is based on all HUs, including those with no persons in the domain;

n_h = number of sample HUs allocated to stratum h;

$n(d) = \sum_h n_h a_h(d)$ = expected number of sample persons eligible by being in domain d.

As shown in the strata descriptions in Table 6.1 here, the commercial list may not have a record for an address. When the list has a record, it may or may not show that the HU has persons that are 65+. In fact, from Valliant et al. (2014, Table 1) 36.8% of the records had no age information; and in the 65+ group, overall the list included 74.6% of the persons found in NSFG. In stratum 2, "List has record; 1 or more persons in age group", 67.1% $\left(p_h(d)\right)$ of HUs actually do have someone 65+ with an average of 0.947

$(a_h(d))$ persons aged 65+ per HU. Although stratum 2 has, by far, the highest incidence of persons 65+, about 1/3 of the HUs in that stratum do not have anyone in the target age group. The other three strata do have small percentages of HUs with someone 65+, even though the list does not say that. Thus, an efficient allocation will assign the most units to stratum 2, but the other strata should be sampled to ensure complete coverage of the age group.

Table 6.1
Strata based on whether commercial list has persons within the 65+ age group. Proportions and averages are estimated from NSFG data.

Stratum h	Description	Proportion of HUs with 1+ persons 65+, $p_h(d)$	Average no. of 65+ persons per HU, $a_h(d)$
1	List has record; 0 persons in age group	0.062	0.071
2	List has record; 1 or more persons in age group	0.671	0.947
3	List has record; no age information	0.122	0.159
4	No list record	0.102	0.128
	Total	0.176	0.236

To approximate costs, suppose that the cost of screening and dropping an ineligible HU is c_S and the average cost of screening an HU and interviewing all persons in an eligible HU is c_{S+I}. The expected cost of a randomly selected sample HU in sampling stratum h, when screening is done to locate a member of domain d, is

$$c_h(d) = p_h(d)c_{S+I} + \left[1 - p_h(d)\right]c_S.$$

Define the design effect due to using unequal weights (Kish, 1992) as $\text{deff}_w = 1 + \sum_{i=1}^{n}\left(w_i - \bar{w}\right)^2 \Big/ \left(n\bar{w}^2\right)$ where n is the sample size, w_i is the sampling weight for element i, and $\bar{w}$ is their mean. The effective sample size is $n_{\text{eff}} = n/\text{deff}_w$ and the expected domain sample size in an equal probability sample of HUs is $n_{\text{eq}}(d)$. The statement of the optimization problem is:

Objective: Find $\{n_h\}_{h=1}^4$ to minimize total screening and interviewing cost, $C_d = \sum_{h=1}^4 n_h c_h(d)$.

Subject to the constraints:

(1) Minimum stratum sample size of HUs: $n_h \geq n_{\min}$;

(2) Effective sample size of persons: $n_{\text{eff}}(d) = n_{\text{eq}}(d)$;

(3) Maximum stratum sample size of HUs: $n_h \leq N_h$ with N_h being the number of HUs in the population in stratum h;

(4) Design effect for person weights: $\text{deff}_w(d) \leq d_0$, a fixed constant.

The second constraint is used to facilitate comparing the MP allocation to one using an equal probability sample of HUs. If the two did not have the same effective sample size and the MP sample size is much smaller, the MP allocation would look unrealistically good compared to equal probability. The third constraint has no effect in the large US household population, but in some applications could be necessary.

The constraint parameters were set to be $n_{\text{eq}}(d) = 2{,}000$, $n_{\min} = 250$, and $d_0 = 1.5$. The unit costs in terms of person-hours were $c_S = 3$ and $c_{S+I} = 10$. Although the objective function is linear in the substratum sample sizes, the design effect, $\text{deff}_w(d)$, has the sample sizes in the denominators of the sampling weights, making this a nonlinear programming (NLP) problem.

The national NSFG estimate of the average number of persons 65+ per HU was 0.236. The approximate number of HUs to screen in an equal probability sample to locate 2,000 persons 65+ would be 8,475 ≐ 2,000 / 0.236. On the other hand, the MP solution to obtain an effective sample size of 2,000 persons was 4,746, i.e., 56% of the equal probability sample. The expected cost of the MP allocation was 19% less than that of the equal probability sample. Valliant et al. (2014) also presented results for the 18-44 age group. Because that age is much more prevalent, MP using the imperfect HU list was less efficient than equal probability sampling for that age group.

Mathematical programming can be applied to situations much more complex than the NSFG example, which does not consider precision of estimates. A household survey like the HRS, introduced in Section 5.2, will serve as a motivating example. This survey has many goals, including estimating statistics for income sources, assets, and health status for financial units, which are similar to households, and persons. The HRS has sample size targets for a set of domains denoted $d = 1, \ldots, D$. The HRS also relies on commercial HU lists for each PSU that classify an HU by race-ethnicity and age of the head of household. These are used to assign HUs to the b substrata. However, the lists are not always accurate – a problem that needs to accounted for in the sample allocation. Define $p_{ab}(d)$ to be the proportion of HUs in SSU stratum/HU substratum ab that are correctly identified by the commercial list data as being in domain d.

Assume there are costs per sample PSU, sample SSU in stratum a, and sample HU in substratum ab, denoted as C_1, C_{2a}, and C_{3ab}, respectively. Using the formulation that $\bar{n}_a$ SSUs are selected in stratum a in every PSU and $\bar{\bar{q}}_{ab}$ HUs are sampled in every ab substratum in each PSU/SSU, a simple cost function is

$$C = C_0 + C_1 m + \sum_{a=1}^{A} C_{2a} m \bar{n}_a + \sum_{a=1}^{A} \sum_{b=1}^{B} C_{3ab} m \bar{n}_a \bar{\bar{q}}_{ab}. \tag{6.1}$$

Let $\hat{\theta}_\ell$, $\ell = 1, \ldots, L$ be a set of estimators that are important in the sample design. The optimization problem is to find $\left\{m, \bar{n}_a, \bar{\bar{q}}_{ab};\ a = 1, \ldots, A, b = 1, \ldots, B\right\}$ that minimize the weighted sum of the relvariances (i.e., the objective function),

$$\phi = \sum_{\ell=1}^{L} \omega_\ell \, \text{CV}^2\left(\hat{\theta}_\ell\right)$$

where the $\hat{\theta}_\ell$ are estimates to be computed from the sample and ω_ℓ is an importance weight for estimate ℓ.

The size of the importance weight, ω_ℓ, assigned to each analysis variable ℓ included in the optimization, depends on the goals of the survey. In some surveys it may be possible to identify variables that are the main outcomes of interest, giving them more weight in the optimization. For example, the HRS variables, income, assets, and health status might be given more weight in the objective function above. The CVs are computed with (5.6) and are used rather than variances because CVs are unitless. This permits estimates that are measured on different scales, like mean income, mean value of owned homes, and proportions of people with poor health or who donate to charities to be included in ϕ without some of them dominating its value as they would if variances of estimators were used.

A variety of constraints can be used on individual CV's and sample sizes at different stages. The ones below are based on a household survey but could be adapted to other types of samples.

(a) Maximum PSU sample size: $m \le m_{\max}$, a maximum set by the sample designer;

(b) Minimum PSU sample size: $m \ge m_{\min}$, e.g., $m_{\min} = 2$ to accommodate variance estimation;

(c) Maximum SSU strata sample size: $\bar{n}_a \le \min\{N_{ia} \mid i = 1, \ldots, m\}$ for all a, i.e., the number of sample SSUs cannot exceed the population count of SSUs in ia;

(d) Minimum SSU strata sample size: $\bar{n}_a \ge \bar{n}_{a,\min}$ for all a;

(e) Maximum HU substrata sample size: $\bar{\bar{q}}_{ab} \le \min\{Q_{iajb} \mid i = 1, \ldots, m;\ j = 1, \ldots, \bar{n}_a\}$ for all ab, i.e., $\bar{\bar{q}}_{ab}$ is bounded above by the smallest value of Q_{iajb} across the PSU/SSU combinations;

(f) Minimum HU substrata sample size: $\bar{\bar{q}}_{ab} \ge \bar{\bar{q}}_{ab,\min}$;

(g) Minimum and Maximum sample size of HUs per PSU: $\mathrm{HU}_{\min} \le q_{i\cdot\cdot} \equiv \sum_a \sum_b \bar{n}_a \bar{\bar{q}}_{ab} \le \mathrm{HU}_{\max}$, i.e., a minimum and maximum number of HUs sampled per PSU; this might be set considering workload requirements for data collectors;

(h) Fixed costs: Total variable cost is less than a budgeted amount $C - C_0 \le C_{\text{budget}}$;

(i) Target sample sizes for analytical domains $d = 1, \ldots, D$ accounting for inaccuracy of listings in commercial list data: The expected number of sample HUs found to be eligible by being in analytical domain d is $q(d) = \sum_{a=1}^{A} \sum_{b=1}^{B} m \bar{n}_a \bar{\bar{q}}_{ab} p_{ab}(d)$. Constraints can be set on this number, e.g., $q(d) = q_0(d)$. Alternatively, constraints can be set on the proportion of HUs allocated to each domain without constraining their totals. For example, if about the same size HU samples are desired in each domain, the constraint might be

$$\frac{q(d)}{\sum_{d=1}^{D} q(d)} = \frac{1}{D} \pm \epsilon$$

for some tolerance ϵ;

(j) Maximum design effects for weights in each domain: $\mathrm{deff}_w(d) \le \mathrm{deff}_{\max}$ where $\mathrm{deff}_w(d)$ is the (Kish, 1992) design effect due to differential weighting in domain d.

The last constraint may or may not be useful. Its intent is to keep base weights from varying too much. However, constraining $\text{deff}_w(d)$ in every domain may conflict with other constraints like target sample sizes for domains. A variation on the above problem would be to use a pre-selected sample of PSUs and optimize the sample allocation within that set. This might be done in a continuing survey that uses the same PSU sample for extended periods of time. Setting up a math program with conflicting constraints is a fairly common issue and will lead to a problem with no feasible solution. Good software will let you know this.

The application in Maze (2021) uses HRS data and incorporates 11 different estimators into the objective function, ϕ. The results are lengthy and are not shown here, but an allocation of PSUs SSUs, and HUs, using anticipated variance component estimates, could be found that meets all of the sample size constraints within a specified, realistic budget.

Putting the above MP into practice requires a number of steps:

(1) Estimate variance components, B^2, W_{2a}^2, and W_{3ab}^2, to use in the relvariance formula (5.6). These will be different for each y;

(2) Estimate the proportions, K_a and K_{ab}, of population totals of y's in the a and ab strata;

(3) Estimate the accuracy rates, $p_{ab}(d)$, of the list being used;

(4) Obtain the population counts, Q_{iajb}, of HUs within PSU i, SSU j in SSU stratum a, and HU substratum b;

(5) Obtain the unit costs, C_0, C_1, C_{2a}, and C_{3ab} needed for the cost function.

All of the above will be facilitated if previous additions of a survey have been conducted, and their data are available for analysis. Even when an MP problem has been carefully formulated, a solution may yield unusual or perplexing results. Meticulous review of the outputs is always wise and may lead to reformulating the problem.

The designer of the sample generally has some leeway in how to formulate an MP problem for sample size allocation. The budget is almost always the most important constraint. How to achieve estimation goals within a fixed budget comes second. The goals may be stated in terms of target sample sizes for analytic subgroups or CV's for important estimates. Constraints, other than budget, may be determined by workloads that data collectors can handle and, also, the need to estimate variances of estimators from the collected data. The last can, for example, dictate that at least two first-stage units be selected within each stratum of PSUs. There is invariably flexibility in how to formulate the allocation problem, the solution to which is part of the craft of sample design. Sample designers have to account for the concerns above one way or another. MP is a formal way of doing so and can often find more efficient solutions than less systematic approaches.

7. Summary

Using models for finite population estimation has gotten much more attention than their use in sample design in the literature. Valliant (2024) reviews many of the alternatives that have been studied, including

model-based estimation (e.g., best linear unbiased prediction and multilevel regression with poststratification) and model-assisted estimation (e.g., general regression, model calibrated, and empirical likelihood estimators). However, models can also play an important role in designing efficient samples. In the design of finite population samples, models provide a way of formally accounting for the effects of auxiliary data available prior to selection of the sample. Often the predictive power of auxiliaries is considered informally in sample design when creating strata or determining selection probabilities, but explicit appeal to models can help create more efficient designs and clarify how efficient a sample will be for different analytic variables to be collected in a sample.

Even in the age of big data, when huge amounts of data can be scraped from the web, sample design still has a place. Editing costs of cleaning web-scraped data may be exorbitant since a survey may have its own specialized definitions for variables like employment status and quality adjusted prices for a price index. Web-scraped data may not be in the form required by a survey, leading to extensive and expensive editing. A well-designed subsample from the big dataset can reduce editing requirements and provide as much information at a lower cost.

The techniques considered here are balanced sampling, cutoff sampling, strata creation, and multistage sampling. In surveys with multiple goals, mathematical programming is a useful technical tool that can formally account for a variety of constraints that sample designers must consider. One area that relies heavily on models but is not covered here is adaptive design as presented in Groves and Heeringa (2006), Schouten, Shlomo and Skinner (2011), Tourangeau, Brick, Lohr and Li (2017), Wagner and Raghunathan (2010), and many other papers. An excellent review and critique of adaptive methods is the Waksberg paper by Tourangeau (2021).

Joe Waksberg was adroit at meeting the challenges of survey design, relying on experience and a sharp intuition. Since his time, there have been advances that, relying on well chosen models, make many sophisticated tools available for designing surveys for an expanded array of challenging goals. This is especially true of mathematical programming, now available in several R packages and spreadsheets. Such software permits complicated, single and multistage sample allocations to be found, subject to many practical constraints.

Acknowledgements

The author thanks Jill Dever, Alan Dorfman, the associate editor, and the referees for their useful comments that improved the coverage of the paper.

References

Ardilly, P., Haziza, D., Lavallée, P. and Tillé, Y. (2024). Jean-Claude Deville's contributions to survey theory and official statistics. *Survey Methodology*, 49, 2, 257-298. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2023002/article/00017-eng.pdf.

Baillargeon, S., and Rivest, L. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77(3), 331-344. DOI: https://doi.org/10.1111/j.1751-5823.2009.00093.x.

Baillargeon, S., and Rivest, L.-P. (2011). The construction of stratified designs in R with the package *stratification*. *Survey Methodology*, 37, 1, 53-65. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11447-eng.pdf.

Ballin, M., and Barcaroli, G. (2008). Optimal stratification of sampling frames in a multivariate and multidomain sample design. Technical report, Insituto Nationale di Statistica, Rome Italy. https://www.istat.it/it/files//2018/07/10_2008.pdf.

Ballin, M., and Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 39, 2, 369-393. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11884-eng.pdf.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42(3), 174-177. DOI: https://doi.org/10.1080/00031305.1988.10475556.

Barcaroli, G., Ballin, M., Odendaal, H., Pagliuca, D., Willighagen, E. and Zardetto, D. (2022). *SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys*, v.1.5-4. https://cran.r-project.org/package=SamplingStrata.

Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4), 651-671.

Benedetti, R., Espa, G. and Lafratta, G. (2008). A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology*, 34, 2, 195-203. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10760-eng.pdf.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1989001/article/14578-eng.pdf.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. New York: John Wiley & Sons, Inc., 2nd edition. ISBN:978-0-471-71813-0.

Breidt, F.J., and Opsomer, J.D. (2009). Nonparametric and semiparametric estimation in complex surveys. In *Handbook of Statistics, Sample Surveys: Inference and Analysis*, (Ed., C.R. Rao), Volume 29B, Chapter 27, 103-119. Amsterdam: Elsevier.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11682-eng.pdf.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 194-199.

Chromy, J.R., and Myers, L.E. (2001). Variance models applicable to the NHSDA. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Cole, D., and Utting, J.E.G. (1956). Estimating expenditure, saving and income from household budgets. *Journal of the Royal Statistical Society, Series A*, 119, 371-392.

Costa, L., Guillo, C., Paliod, N., Merly-Alpa, T., Vincent, L., Chevalier, M. and Deroyon, T. (2018). Le tirage coordonné du nouvel échantillon-maître nautile avec l'échantillon de l'enquête emploi en continu. *Journées de Méthodologie Statistique, INSEE*.

Cumberland, W.G., and Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society B*, 43, 353-367.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54(285), 88-101.

de Moura Brito, J.A., Silva, P.L.N., Semaan, G.S., and Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41, 2, 427-442. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-eng.pdf.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2), 411-425.

Dorfman, A.H., and Valliant, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264. DOI: https://doi.org/10.1214/16-STS598.

Fecteau, S., and Jocelyn, W. (2005). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incorporées. *Méthodes d'Enquêtes et Sondages*, (Eds., P. Lavallée and L.-P. Rivest), 405-411. Paris : Dunod.

Frankel, L.R., and Stock, S. (1942). On the sample survey of unemployment. *Journal of the American Statistical Association*, 37(443), 77-80. DOI: https://doi.org/10.1080/01621459.1942.10500615.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944. DOI: https://doi.org/10.1093/biomet/asp042.

Godambe, V.P., and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations – I. *The Annals of Mathematical Statistics*, 36, 6, 1707-1723.

Godfrey, J., Roshwalb, A. and Wright, R. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*, 2(1), 1-9.

Gomes, H., and Johnson, W.H. (2016). Sample size optimization of the Consumer Price Index: An implementation using R. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 2137-2151.

Goodman, R., and Kish, L. (1950). Controlled selection – A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142, 139-147.

Grafström, A., Lisic, J. and Prentius, W. (2023). *BalancedSampling: Balanced and Spatially Balanced Sampling,* R package version 1.6.3. https://CRAN.R-project.org/package=BalancedSampling.

Grafström, A., Lundström, N.L.P. and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520. DOI: https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Grafström, A., and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24, 120-131.

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, 3, 439-457. DOI: http://dx.doi.org/10.1111/j.1467-985x.2006.00423.x.

Groves, R.M., Mosher, W.D., Lepkowski, J. and Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. Vital Health Statistics, Series 1, No. 48, National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_01/sr01_048.pdf.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2004002/article/7749-eng.pdf.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953a). *Sample Survey Methods and Theory, Volume I. Methods and Applications*. New York: John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953b). *Sample Survey Methods and Theory, Volume II. Theory*. New York: John Wiley & Sons, Inc.

Haziza, D., Chauvet, G. and Deville, J.-C. (2010). Sampling and estimation in the presence of cut-off sampling. *Australia & New Zealand Journal of Statistics*, 52(3), 303-319. DOI: https://doi.org/10.1111/j.1467-842X.2010.00584.x.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1), 67-76. https://www.jstor.org/stable/25472691.

Hughes, E., and Rao, J.N.K. (1979). Problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics – Theory and Methods*, 8(15), 1551-1574.

Hunter, S.R., Bowman, K.R. and Chromy, J.R. (2005). Results of the variance component analysis of sample allocation by age in the National Survey on Drug Use and Health. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3132-3136.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377), 89-96. DOI: https://doi.org/10.2307/2287773.

Judkins, D., and Van de Kerckhove, W. (2003). RECS 2005 optimization. Prepared for U.S. Department of Energy, no. 16.3, Task 98-010, contract no.: DE-AC01-96E123968. Technical report, Westat, Rockville MD.

Kirkendall, N.J. (1992). When is model-based sampling appropriate for EIA surveys? *Proceedings of the Section on Survey Methods Research*, 637-642. http://www.asasrms.org/Proceedings/index.html.

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

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, 8(2), 183-200.

Knaub, J.R. (2008). Cutoff vs. design-based sampling and inference for establishment surveys. *InterStat*. http://interstat.statjournals.net/YEAR/2008/abstracts/0806005.php.

Kott, P.S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, 73(2), 485-491.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-eng.pdf.

Leaver, S., and Solk, D.T. (2005). Handling program constraints in the sample design for the commodities and services component of the U.S. Consumer Price Index. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3290-3298.

Maze, A. (2021). *Using Commercial List Information in Screening Eligible Housing Units*. PhD thesis, University of Maryland. DOI: https://doi.org/10.13016/xdzx-dto7.

Morganstein, D., and Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15(3), 299-312.

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

Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59(305), 18-55. DOI: https://doi.org/10.1080/01621459.1964.10480699.

Pfeffermann, D., and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of Statistics, Sample Surveys: Design, Methods, and Applications*, (Ed., C.R. Rao), Volume 29A, Chapter 39. Amsterdam: Elsevier.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 2, 191-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-eng.pdf.

Royall, R.M. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Methodology*, 18, 2, 179-185. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14488-eng.pdf.

Royall, R.M., and Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80(390), 355-359.

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

Royall, R.M., and Herson, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association*, 68(344), 890-893.

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

Schouten, B., Shlomo, N. and Skinner, C.J. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 1-24.

Schroeder, T., and Ault, K. (2001). The NEISS Sample (Design and Implementation) from 1979 to 1996. Technical report, U.S. Consumer Product Safety Commission, Washington DC. https://www.cpsc.gov/s3fs-public/2001d010-6b6.pdf.

Schwendinger, F., and Borchers, H.W. (2023). CRAN Task View: Optimization and Mathematical Programming. Technical report, The R Foundation, Vienna Austria. https://CRAN.R-project.org/view=Optimization.

Searle, S., Casella, G. and McCulloch, C. (1992). *Variance Components*. New York: John Wiley & Sons, Inc.

Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Slanta, J.G., and Krenzke, T.R. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 693-698.

Slanta, J.G., and Krenzke, T.R. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Survey Methodology*, 22, 1, 65-75. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14384-eng.pdf.

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Tillé, Y., and Matei, A. (2023). *sampling: Survey Sampling*, R package version 2.10. http://CRAN.R-project.org/package=sampling.

Tourangeau, R. (2021). Science and survey management. *Survey Methodology*, 47, 1, 3-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00006-eng.pdf.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society, Series A*, 180(1), 203-223. DOI: http://onlinelibrary.wiley.com/doi/10.1111/rssa.12186.

U.S. Census Bureau (2006). Current Population Survey: Design and Methodology. https://www2.census.gov/programs-surveys/cps/methodology/tp-66.pdf.

U.S. Energy Information Administration (2018). EIA-914 monthly crude oil and lease condensate, and natural gas production report methodology. Technical report, US Department of Energy, Washington DC. https://www.eia.gov/petroleum/production/pdf/eia914methodology.pdf.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology*, 28, 1, 103-114. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002001/article/6424-eng.pdf.

Valliant, R. (2024). Hansen lecture 2022: The evolution of the use of models in survey sampling. *Journal of Survey Statistics and Methodology*, 12(2), 275-304. DOI: https://doi.org/10.1093/jssam/smad021.

Valliant, R., Dever, J.A. and Kreuter, F. (2015). Effects of cluster sizes on variance components in two-stage sampling. *Journal of Official Statistics*, 31(4), 763-782. DOI: http://dx.doi.org/10.1515/JOS-2015-0044.

Valliant, R., Dever, J.A. and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, 2nd edition. New York: Springer.

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

Valliant, R., and Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25(3), 337-360. DOI: https://doi.org/10.1016/S0167-9473(97)00007-8.

Valliant, R., Hubbard, F., Lee, S. and Chang, C. (2014). Efficient use of commercial lists in U.S. household sampling. *Journal of Survey Statistics and Methodology*, 2(2), 182-209. DOI: https://doi.org/10.1093/jssam/smu006.

Varadhan, R. (2023). *alabama: Constrained Nonlinear Optimization*, R package version 2023.1.0. https://CRAN.R-project.org/package=alabama.

Wagner, J., and Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29(9), 1014-1024. DOI: https://doi.org/10.1002/sim.3834.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73(361), 40-46. DOI: https://www.tandfonline.com/doi/abs/10.1080/01621459.1978.10479995.10479995.

Waksberg, J., Sperry, S., Judkins, D. and Smith, V. (1993). The National Survey of Family Growth, Cycle IV, evaluation of linked design. *Vital and Health Statistics*, 2, (117), (PHS) 93-1391.

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(453), 185-193. http://www.jstor.org/stable/2670358.

Wu, C.F.J., and Hamada, M. (2021). *Experiments: Planning, Analysis, and Optimization*, 3rd edition. New York: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781119470007.

Yorgason, D., Bridgman, B., Cheng, Y., Dorfman, A., Lent, J., Liu, Y.K., Miranda, J. and Rumburg, S. (2011). Cutoff sampling in federal surveys: An inter-agency review. Technical report, Bureau of Labor Statistics, Washington DC. https://www.bls.gov/osmr/research-papers/2011/pdf/st110050.pdf.

Ypma, J., Johnson, S.G., Borchers, H.W., Eddelbuettel, D., Ripley, B., Hornik, K., Chiquet, J., Adler, A., Dai, X., Stamm, A. and Ooms, J. (2022). *nloptr: R Interface to NLopt*, version 2.0.3. https://CRAN.R-project.org/package=nloptr.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 185-207
Statistics Canada, Catalogue No. 12-001-X

Design consistent random forest models for data collected from a complex sample

Daniell Toth and Kelly S. McConville[1]

Abstract

Random forest models, which are the result of averaging the estimated values from a large number of tree models, represent a useful and flexible tool for modeling the data nonparametrically to provide accurately predicted values. There are many potential applications for these types of models when dealing with survey data. However, survey data is usually collected using an informative sample design, so it is necessary to have an algorithm for creating random forest models that account for this design during model estimation.

The tree models used in the forest are typically obtained by estimating tree models on bootstrapped samples of the original data. Since the models depend on the observed data and the values observed in the sample depend on the informative sample design, the usual method for estimation is likely to lead to a biased random forest model when applied to survey data.

In this article, we provide an algorithm and a set of conditions that produce consistent random forest models under an informative sample design and compare this method to the usual random forest modeling method. We show that ignoring the design can lead to biased model estimates.

Key Words: Machine learning; Nonparametric; Sample design; Survey data; Tree models.

1. Introduction

Recursive partitioning algorithms were first suggested by Morgan and Sonquist (1963) as a method for analyzing survey data because of the complicated relationships, including interaction effects, among variables that are typical of these datasets. Variables collected from a survey are often highly correlated (even collinear) with each other, are frequently categorical, and can contain many missing values. These complications can make it difficult to make inference about the target population with this data using traditional parametric models. Tree models, which are estimated by applying a recursive partitioning algorithm to the dataset, handle this type of data easily. The variables used in the model along with any interaction effects are selected automatically and the resulting binary split structure makes these models easy to interpret and identifies complicated interaction effects between the variables in the dataset (Phipps and Toth, 2012; Earp, Toth, Phipps and Oslund, 2018).

Given a set of n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$, of a random response variable Y and d random predictor variables $\mathbf{X} = (X_1, \ldots, X_d)$, from an informative sample, we want to estimate k new values $\{y_i,\}_{i=n+1}^{n+k}$, given predictor values $\{\mathbf{x}_i\}_{i=n+1}^{n+k}$ for non-sampled units in the population. By estimating the mean function $E[Y \mid \mathbf{X} = \mathbf{x}] = h(\mathbf{x})$ from the observed data we can get predictions of y_i from the model $\tilde{y}_i = \tilde{h}(\mathbf{x})$.

Using survey data to estimate a model in order to obtain good predictions rather than estimates of a finite population parameter is a topic with increasing interest (Wieczorek, 2023). For example, Hong and He (2010) use longitudinal study data to fit a model that can be used to predict the functional mobility status

1. Daniell Toth, Office of Survey Methodology, U.S. Bureau of Labor Statistics. E-mail: toth.daniell@bls.gov; Kelly S. McConville, Department of Statistics, Harvard University.

among the elderly. Meanwhile, Kshirsagar, Wieczorek, Ramanathan and Wells (2017) and Krebs, Reeves and Baggett (2019) both use machine learning models to predict poverty levels and under-story vegetation structure respectively. Similarly, we are interested in estimating the regression function using a machine learning approach using data from an informative sample design. Like Nalenz, Rodemann and Augustin (2024) we propose an approach to modeling random forests using survey data.

A tree model, $\tilde{h}(\mathbf{x})$, is a nonparametric model obtained from an algorithm that recursively partitions the observed data and then estimates the desired statistic for each final partitioning box (end node) separately. The recursive partitioning algorithm consists of choosing a variable X_j among all the available d variables given by vector $\mathbf{X}$ and a value a in which to split the set of observations into two nodes: the observations where $\mathbf{x}_j \le a$ and $\mathbf{x}_j > a$. This procedure is then repeated for each node until there are not enough observations to split or some stopping criteria have been reached (Hothorn, Hornik and Zeileis, 2006). This algorithm results in a set of q boxes, $Q^n = \{B_1, \ldots, B_q\}$ which completely partition the support of $\mathbf{X}$ and depend on the values of the observed data.

Though tree models are easy to interpret, making them ideal for many inference applications, they are not very efficient models for producing point estimates. They are particularly inefficient for models that have linear effects (Loh, 2008). A random forest model, in contrast to the easily interpretable tree model, estimates the expected value of the response variable conditionally on the predictor variables, by averaging the estimates from a set of regression tree models.

Given a set of M regression tree models, $\{\tilde{h}_j\}_{j=1}^{M}$, the random forest estimator of $h(\mathbf{x})$ is

$$\mathcal{F}_0(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\tilde{h}_j(\mathbf{x}), \tag{1.1}$$

where each tree model, $\tilde{h}_j$, is fit using a random subset of the predictor variables on a bootstrap sample of the observed data (Breiman, 2001). Though these models lose the feature of easy interpretation that tree models possess, they are known to provide very accurate predictions and still retain the applicability to a wide range of data types (Breiman, 2001). This provides a useful and flexible tool for accurately modeling the response variable of a given dataset which could have many applications in analyzing data from an informative sample.

For example, Buskirk (2018) and Bilton, Jones, Ganesh and Haslett (2017) present applications of regression trees and random forests on data collected using a complex sample design. Unfortunately, the standard random forest algorithms are meant for independent and identically distributed (i.i.d.) data and many surveys use a complex sample design to collect observations, violating the i.i.d. assumption and in many applications of tree-based models, the available sample design information is often ignored, likely leading to biased estimates as demonstrated by the results in Toth and Eltinge (2011).

Dagdoug, Goga and Haziza (2021) extended the work of McConville and Toth (2019) by using a forest model instead of a single tree as the assisting model in a model assisted estimator to estimate a finite population total. They point out that, if the variables used to determine the sample design are available, it is

possible much of the bias could be reduced by including them in the model. These variables are extremely useful for estimation of population parameters in the context of model assisted estimation, but these variables are not available for making predictions of values for units outside of the sample.

It is desirable then to have an algorithm that allows consistent estimation of a regression function for the population, estimated using data from an informative design, that can be used for prediction. For example work being done at the BLS requires a model to predict respondent-burden for households selected in the Consumer Expenditure Survey from household characteristics which are believed to be associated with burden Yang and Toth (2022).

In this article, we propose a design consistent random forest model for the regression function that uses a weighted average of end-nodes obtained from a set of purely random trees that incorporate sampling weights in their estimation. This process avoids having to produce sensible bootstrap samples from a general sample design. Forests constructed from completely random trees have been studied in the literature with the method commonly referred to as the *uniform random forest algorithm* (Biau, Devroye and Lugosi, 2008; Scornet, 2016; Arlot and Genuer, 2014) but these models are generally not effective in practice. They are used primarily to study the theoretical properties and to understand the behavior and limits of ensemble methods. In the standard uniform random forest algorithm, the end-node estimates are simple averages whereas the end-node weights of our method enables good predictive properties.

To our knowledge our algorithm is the first to propose using weights at the end-node, rather than at the tree level. Because the trees are produced using completely random splits of the predictor space all the real work comes from these weights, providing an adaptive and more efficient estimator. We show that this model provides design consistent estimates and is, therefore, more appropriate for use with data collected using an informative sample design.

In Section 2 we introduce the tree model and provide the necessary assumptions for its design consistency. Section 3 contains the method for using tree models in a random forest model that requires weighting each tree model estimate and the statement of the main theoretical result of the article, which is that the proposed random forest estimator is design consistent estimator of the regression function. Appendix contains all necessary auxiliary lemmas and proofs of the results. Section 4 summarizes simulation studies where we compare the performance of our proposed method to the standard random forest estimator on data from simple random samples (SRS) and from probability proportional to size (PPS) samples. In particular, we apply our proposed random forest model to the Academic Performance Index (API) score data from standardized test results of students computed for all California schools with at least 100 students and the U.S. Bureau of Labor Statistics' (BLS) Consumer Expenditure (CE) data. These results demonstrate that ignoring the sample weights using data from an informative sample design could lead to biased forest estimators. Lastly, also in Section 4, we conduct a simulation study on generated data to explore consistency of our proposed method and the standard random forest model.

2. Design consistent tree models

Consider a finite population of size N, $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^{N}$, generated from a super-population model ξ, where Y_i is the variable of interest associated with unit i and $\mathbf{X}_i$ is a d -vector of potential predictors variables associated with unit i that are part of the released data available to the analyst. We use $\mathbf{Z}_i$ to denote a d^*-vector of variables associated with unit i, known by the survey designer but not released with the survey data for analysis.

A random sample $S \subset U = \{1,\ldots,N\}$ of size n is selected using a sample design with inclusion probabilities $\pi_i = P(i \in S)$. The sample design, defined by the inclusion probabilities, can depend on variables associated with the unit where some are known and some are unknown to the analyst, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i)$. If these inclusion probabilities are associated with Y, the design can affect the estimates and inference for the population that result from using the sample data. Such sample designs are called *informative* sample designs. If the sample design only depends on variables available to the data analyst, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i) = P(i \in S \mid \mathbf{x}_i)$, design-consistent models may be able to be obtained by incorporating all of these variables used to define the inclusion probabilities into the modeling process (Gelman, King and Liu, 1998; Little, 2004). However, in most publicly released survey datasets, many of the variables used in the survey design are not released to the data analyst. Instead, the data is released with a set of survey weights $\{w_i\}_{i \in S}$ intended to be used by the data analyst to account for the survey design in the analysis (Lavallée and Beaumont, 2015; Pfeffermann, 1993). Besides accounting for the probability of selection into the sample, these weights often include adjustments for nonresponse and/or known totals of key auxiliary information. Though our arguments in this article could be used for general survey weights, for simplicity of exposition, we assume that the weight associated with unit i is the inverse of the probability of selection for that unit, π_i^{-1}.

In order to study large sample properties of the estimator in this context, it is necessary to consider a sequence of populations that are increasing in size and distributed i.i.d. from super-population and a sequence of associated sample designs. From each population-design pair a corresponding sequence of samples is drawn, also increasing in size and each selected according to the sample design. More concretely, suppose we have a sequence of finite populations $\{(Y_1, \mathbf{X}_1), \ldots, (Y_{N_\nu}, \mathbf{X}_{N_\nu})\}$, indexed by ν, so that $U_1 \leq \ldots, \leq U_\nu$, with sizes $N_1 \leq \ldots \leq N_\nu$. Each finite population is generated by taking i.i.d. draws from the distribution of the super-population ξ. The random samples, $S_1 \subset U_1, \ldots S_\nu \subset U_\nu$, are drawn from each finite population using the corresponding sample design, with increasing sizes $n_1 \leq \ldots \leq n_\nu$. It is the behavior of the sequence of estimates obtained from these samples that is considered.

If a tree model, $\tilde{h}_\nu(\mathbf{x})$, is obtained by recursively partitioning the observed sample data and then estimating the mean in each box, the resulting tree model is an estimator of the conditional mean function $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{x}]$. Toth and Eltinge (2011) provide an algorithm for estimating $h(\mathbf{x})$ and a set of conditions for which this estimator is an L^2 consistent estimator of $h(\mathbf{x})$. In this article, we intend to propose a random forest model that is constructed from a weighted average of these design consistent tree models. For the rest of this section, we review the notation and results necessary to establish a design consistent algorithm for

random forest models including a discussion of the conditions and the main result for the tree model given in Toth and Eltinge (2011).

Let $Q^{n_\nu} = \{B_1^{n_\nu}, \ldots, B_q^{n_\nu}\}$ be the set of partitioning boxes that result from applying a recursive partitioning algorithm to the observed sample S_ν. To facilitate discussion of the predicted value of an observation with predictor variables $\mathbf{x}$ for a given tree, we now define some functions that help simplify the notation. Let $B^{n_\nu}(\mathbf{x})$ denote the box in Q^{n_ν} containing the value $\mathbf{x}$. The functions $\#B^{n_\nu}(\mathbf{x}) = \sum_{i \in S} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}$ and $\tilde{\#}B^{n_\nu}(\mathbf{x}) = \sum_{i \in S} \pi_i^{-1} \mathbb{I}_{\{x_i \in B^{n_\nu}(\mathbf{x})\}}$ provide the number of observed sample units and estimated number of population units in box $B^{n_\nu}(\mathbf{x})$ respectively. The estimated mean of the box containing the value $\mathbf{x}$ is define as

$$\tilde{\mu}(\mathbf{x}) = \left[\tilde{\#}B^{n_\nu}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \tag{2.1}$$

Note that this is the standard Hájek estimator of the mean for population units contained in the partitioning box $B^{n_\nu}(\mathbf{x})$ (Hájek, 1960).

Each box B^{n_ν} in a given partition Q^{n_ν} has two corresponding index vectors which define the borders of the box in all d dimensions. Given a box of a partition, B^{n_ν}, let $\mathbf{a}(B^{n_\nu}) = (a_1(B^{n_\nu}), \ldots, a_d(B^{n_\nu}))$ and $\mathbf{b}(B^{n_\nu}) = (b_1(B^{n_\nu}), \ldots, b_d(B^{n_\nu}))$, where for every $\mathbf{x} \in B^{n_\nu}$, $a_l(B^{n_\nu}) \le x_l < b_l(B^{n_\nu})$, for $l = 1 \ldots d$.

For a given value $\mathbf{x}$ in the support of $\mathbf{X}$, we use the notation $\tilde{F}_i^{\nu}(\cdot)$ to denote the empirical marginal distribution function of x_l conditioned on the partition. That is, for a constant c, and given value $\mathbf{x}$ the empirical marginal distribution function of x_l conditioned on the partition is

$$\begin{aligned} \tilde{F}_l^{\nu}(c \mid Q^{n_\nu}) &= \tilde{F}_l^{\nu}\left(c \mid B^{n_\nu}(\mathbf{x})\right) \\ &= \left(\tilde{\#}_{N_\nu}\left(B^{n_\nu}(\mathbf{x})\right)\right)^{-1} \sum_{i \in S_\nu} \pi_{\nu i}^{-1} \mathbb{I}_{\{x_{li} \le c\}} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \end{aligned} \tag{2.2}$$

The left continuous conditional empirical marginal distribution function $\tilde{F}_l^{-}$ is defined by replacing the indicator function $\mathbb{I}_{\{x_{li} \le x\}}$ in the above definition with $\mathbb{I}_{\{x_{li} < x\}}$. We will also use the empirical probability function $\tilde{P}_{n_\nu}(\mathcal{A})$ of a given event $\mathcal{A}$. The empirical probability function is defined as

$$\tilde{P}_{n_\nu}(\mathcal{A}) = \tilde{N}_\nu^{-1} \sum_{i \in S_\nu} \pi_{\nu i}^{-1} \mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i), \tag{2.3}$$

where $\mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i) = 1$, if the event $\mathcal{A}$ is satisfied for observation $\mathbf{x}_i$ and where $\tilde{N}_\nu = \sum_{i \in S_\nu} \pi_i^{-1}$.

Next we define the l -norm of partition Q^{n_ν} relative to $\tilde{F}_l$ by

$$\| Q^{n_\nu} \|_l^{\tilde{F}} = \sum_{B^{n_\nu} \in Q^{n_\nu}} \left\{ \left[\tilde{F}_l\left(b_l(B^{n_\nu})\right) - \tilde{F}_l\left(a_l(B^{n_\nu})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_\nu}) \right\} \tag{2.4}$$

and the l -norm of partition Q^{n_ν} relative to $\tilde{F}^{-}$

$$\| Q^{n_\nu} \|_l^{\tilde{F}^{-}} = \sum_{B^{n_\nu} \in Q^{n_\nu}} \left\{ \left[\tilde{F}_l^{-}\left(b_l(B^{n_\nu})\right) - \tilde{F}_l^{-}\left(a_l(B^{n_\nu})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_\nu}) \right\}. \tag{2.5}$$

The following conditions on the super-population model, sample design and partition created from the algorithm are sufficient to show that a regression tree model based on the sample data is an L^2 - consistent estimator of the true conditional mean of the variable of interest, Y. In Section 3, we show that these conditions (with the strengthening of one condition) are sufficient to obtain consistent forest estimators as well. The proofs for consistent regression trees require only a finite second moment of the variable of interest, but we require a finite fourth moment to prove consistency of the proposed forest estimator.

Many of the conditions to obtain consistency require understanding the rate at which things converge. Before specifying the conditions on the population, sample design and algorithm, we first define two scalar functions that will be used as the rates of convergence. Let $\gamma(x)$ and $k(x)$ be functions bounded above 0 for all $x > 0$ satisfying:

1: $\gamma(x) \to \infty$

2: $x^{-1}k(x) \to 0$

3: $k(x)^{-1}\gamma(x)x^{1/2} \to 0,$

as $x \to \infty$. These constraints require both functions to be unbounded where $\gamma(x)$ grows to ∞ slower than $\sqrt{x}$, while $k(x)$ grows faster than $\sqrt{x}$, but slower than x. Note that there are an infinite number of function pairs that satisfy these three constraints. Below we use these functions to specify the relative speeds at which different terms converge relative to the sizes of the population N_v and sample n_v. We will also use the sampling fraction, defined as $f_v = n_v / N_v$.

Condition 1: $\lim_{v\to\infty} N_v^{-1}\sum_{i=1}^{N_v} Y_i^4 < \infty$

Condition 2: $\limsup_{v\to\infty}(N_v \min_{i\in U_v}\pi_{vi})^{-1} = O(n_v^{-1})$

Condition 3: $\limsup_{v\to\infty}\max_{i,j\in U_v, i\neq j}\left|\frac{\pi_{vij}}{\pi_{vi}\pi_{vj}} - 1\right| = O(N_v^{-1})$

Condition 4: $f_v^{-1} = O(n_v^{1/2}\gamma(n_v)^{-1})$

Condition 5: $E_p[\delta_{vi}\delta_{vj} \mid Q^{n_v}] = \pi_{vij} \quad \forall i, j \in U_v$

Condition 6: $\tilde{P}\left(k(n_v)^{-1}\#_{n_v}(B^{n_v}(\mathbf{x})) \geq 1\right) \to_p 1$

Condition 7: $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}} \to_p 0$ and $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}^-} \to_p 0,$ for $l = 1, \ldots, d$

where the above conditions are all assumed with ξ -probability 1.

Condition 1 is the only condition directly on the distribution of the super-population model. The data do not need to follow any predefined distribution, requiring only that the outcome variable, Y, has a finite fourth moment. This generality makes these models applicable to a wide class of problems. We use this condition on the fourth moment for establishing design consistency of the proposed forest estimator, but as mentioned above, this condition could be weakened to require only a finite second moment in the case of design consistent tree estimators.

Conditions 2 through 4 are standard conditions on the sample design requiring that every unit in the population can be selected with some minimum probability, the effect of clustering shrinks relative to the population size and a mild requirement on the sampling rate (Isaki and Fuller, 1982; Breidt and Opsomer, 2000). Condition 4 is a weak limit on how big the finite populations can grow relative to the sample size in the sense that it allows for an arbitrarily small sampling rate.

Condition 5 is a condition from Toth and Eltinge (2011) requiring that the selection probabilities are independent of the given partition. The partitioning set Q^{n_ν} is a function of an algorithm applied to the selected data, so this condition on both the algorithm and the sample design limits the influence any selected unit can have on the resulting partition.

Condition 6 and 7 are both conditions on the partitioning algorithm. The first requires that the number of observations in each partitioning box grows at a certain rate relative to the sample size, while the second requires the l-norms of the partitioning boxes, defined by 5 and 6, shrink toward zero as the sample size grows.

Proposition 2.1 (Toth and Eltinge, 2011)**.** *Let* $\{(Y_i, \mathbf{X}_i)\}_{i\in U_\nu}$, *be a sequence of finite populations, indexed by* ν *and distributed i.i.d. from the super-population model* ξ *and let* S_ν *denote a random sample from* U_ν *selected using the sample design. Given* Q^{n_ν}, *the collection of partitioning boxes created from the algorithm applied to the sample data,* S_ν, *define*

$$\tilde{h}_{n_\nu}(\mathbf{x}) = (\tilde{\#}B^{n_\nu}(\mathbf{x}))^{-1}\sum_{i\in S_\nu}\pi_i^{-1}y_i\,\mathbb{I}_{\{\mathbf{x}_i\in B^{n_\nu}(\mathbf{x})\}}. \tag{2.6}$$

If $\lim_{\nu\to\infty}N_\nu^{-1}\sum_{i=1}^{N_\nu}Y_i^2<\infty$ *and Conditions 2 through 7 are satisfied with* ξ *-probability 1, then*

$$\lim_{\nu\to\infty}E_{\xi p}\left[\left|\tilde{h}_{n_\nu}(\mathbf{x})-h(\mathbf{x})\right|^2\right]=0.$$

Notice that the right hand side of equation (2.6) is the Hájek estimator of the mean of the box containing $\mathbf{x}$ given by (2.1). Since the set of boxes partition the data, each observation falls into exactly one box and the model predicted value of Y for an observation with auxiliary variables $\mathbf{X}=\mathbf{x}$ is simply

$$\tilde{h}_{n_\nu}(\mathbf{x}) = \tilde{\mu}(\mathbf{x}). \tag{2.7}$$

Therefore, Proposition 2.1 tells us that a regression tree model that estimates the mean of Y in each end node is an L^2 - consistent estimator of the function $E_\xi[Y\,|\,\mathbf{x}]$, so

$$\tilde{\mu}(\mathbf{x})\to_{L^2}E_\xi[Y\,|\,\mathbf{x}]. \tag{2.8}$$

We will rely on this result in the following section to show that the proposed random forest estimator is consistent as well as the following corollary.

Corollary 2.1. *For a given tree* j, *if Conditions 1 through 7 are satisfied, then*

$$\left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1}\sum_{i\in S}\pi_i^{-1}y_i^2\,\mathbb{I}_{\{\mathbf{x}_i\in B_j^{n_v}(\mathbf{x})\}}\to_p E_\xi[Y^2\mid\mathbf{x}].$$

Proof in Appendix.

3. Design consistent forest models

Random forest model estimates are obtained by averaging the estimates of M tree models. This requires using a procedure for producing several *different* tree models using the same data; it does not improve the estimate to average over the same model. The tree models used in the forest are typically obtained by estimating tree models on bootstrapped samples of the original data and using a random subset of the predictor variables at each split (Breiman, 2001). However, a bootstrap sample of a dataset is not always easy or possible to produce for a general informative sample design (Mashreghi, Haziza and Léger, 2016). The typical approach of ignoring the sample design during estimation is likely to lead to a biased random forest model when applied to survey data.

That is, given an estimator $\hat{m}_n$ of the regression function m and a point $\mathbf{x}$,

$$\text{Bias}(\hat{m}_n(\mathbf{x})) := \mathbb{E}[\hat{m}_n(\mathbf{x})]-m(\mathbf{x}), \tag{3.1}$$

where the expectation is taken with respect to the joint distribution of the super-population model and the sample design.

We now propose a forest model that is design consistent for a family of informative sample designs provided that the sample design, super-population distribution, and recursive partitioning procedure satisfy Conditions 1 through 7 in Section 2. In order to obtain different regression tree models from a given sample, at each step of the recursive partitioning we select the variable completely at random from the d possible predictor variables and the splitting point at random from the observed support of the selected variable. This algorithm is outlined in Figure 3.1.

Figure 3.1 Recursive partitioning algorithm to produce random tree models.

Recursive Partitioning Algorithm
1. Let $n_{\text{end}} = \max(5,\ \text{floor}\ \{10^{-7}n\})$.
2. If the dataset contains at least $2n_{\text{end}}$ observations continue to the next step; otherwise stop.
3. Among the auxiliary variables x_l, $l=1,\ldots,d_1$, randomly choose a variable on which to split the data.
4. Split the data into two sets S_L and S_R by randomly selecting a value of the selected variable x_l that results in each sub-dataset containing at least n_{end} observations.
5. Apply the algorithm beginning at step 2 to each of the two resulting subsets S_L and S_R.

Note: Notice that n_{end} is defined so that the end-nodes of each tree satisfies Conditions 6 and 7.

Note that n_{end} is defined in such a way as to satisfy Condition 6, because it is linear in n and thus dominates $\sqrt{n}$, but still allows for a relatively small number of observations. This is important because in practice a small number of observations is effective in getting accurate estimates.

3.1 Extended notation for forests

Because we are interested in forests, which require a set of trees, we extend some of the notation and functions used in Section 2 to facilitate this discussion. For instance, rather than one set of partitioning boxes for instance v, we will have one for each of the M trees in the model. Denote the partition for the j-th tree by $Q_j^{n_v} = \{B_{j1}^{n_v}, \ldots, B_{jq_j}^{n_v}\}$. Note that the number of partitioning boxes q, created using the sample S_v, can be different for the different trees in the forest model, so depends on j. Let $\mathcal{Q}^{n_v} = \{Q_j^{n_v}\}_{j=1}^{M}$ be the set of all the partitions making up the forest model.

The function $B_j^{n_v}(\mathbf{x})$ will denote the box in the j-tree that contains the value $\mathbf{x}$, while $\#B_j^{n_v}(\mathbf{x})$ and $\tilde{\#}B_j^{n_v}(\mathbf{x})$ are the number of observed sample units and estimated number of population units in box $B_j^{n_v}(\mathbf{x})$ respectively. Likewise, the estimated mean of the observations in the box containing the value $\mathbf{x}$, defined by equation (2.1), for the j-th tree is denoted $\tilde{\mu}_j(\mathbf{x})$.

3.2 Weights for averaging estimates

Notice that the algorithm and therefore the structure of each tree depends only on the observed values of the modeling variables $\{\mathbf{X}_i\}_{i=1}^{n}$, resulting in a k-nearest neighbor estimate of Y based on the closeness of a random sub-sample of modeling variables. The forest model is then an average over M k-nearest neighbor estimates. However, because the trees are built based on randomly selected splits, the homogeneity of Y will likely vary across boxes, resulting in more or less informative boxes. So while the simple average of the random tree estimates given by (1.1) leads to an asymptotically unbiased estimator, it will also be rather inefficient.

In order to improve the efficiency of the forest estimator, we use a weighted average, with the goal of giving more weight to estimates from tree models with greater predictive accuracy. Let $\{\tilde{h}_j^{n_v}\}_{j=1}^{M}$, denote the set of regression tree models. Then a weighted forest model would have the form

$$\mathcal{F}_w(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x})\, \tilde{h}_j^{n_v}(\mathbf{x}), \tag{3.2}$$

where $\lambda_j(\mathbf{x})$ is a weight that depends on the end-node of tree j that $\mathbf{x}$ belongs to.

Methods for using a weighted average of tree estimates to produce a forest estimate have been considered (Gajowniczek, Grzegorczyk, Ząbkowski and Bajaj, 2020; Shahhosseini and Hu, 2020; Winham, Freimuth and Biernacka, 2013) but these involve using a weight based on the fit of each tree only. In testing several different approaches, we found using a weight that depends on the final end node produced the best results. However, this approach induces bias in the estimates which needs to be adjusted for.

For our proposed method, we weight each tree estimate using a weight that is inversely proportional to the estimate of one plus the end node variance, $V_{B_j^{n_v}(\mathbf{x})} = \mathrm{Var}_{\xi}(Y \mid \mathbf{X} \in B_j^{n_v}(\mathbf{x}))$ similar to resulting weights used in some adaptive methods (Williams, Neilley, Koval and McDonald, 2016). If we knew the true mean, $\mu_j(\mathbf{x}) = E[Y \mid B_j^{n_v}(\mathbf{x})]$, of the Y-values for the observations in box $B_j^{n_v}(\mathbf{x})$, then a design consistent estimator of $V_{B_j^{n_v}(\mathbf{x})}$ is given by

$$\left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \mu_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}.$$

However, because the true $\mu_j(\mathbf{x})$ is unknown, we use the estimator

$$\tilde{V}_{B_j^{n_v}(\mathbf{x})} = \left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \tilde{\mu}_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}, \tag{3.3}$$

where the estimated value $\tilde{\mu}_j(\mathbf{x})$ replaces the true mean.

Given $\mathbf{x} \in B_j^{n_v}(\mathbf{x})$, then the weight for tree j in the forest is set to

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}. \tag{3.4}$$

so that the weights $\lambda_j(\mathbf{x}) \propto (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}$ and they sum to 1.

The forest estimated value of y given the $\mathbf{x}$ is

$$\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j^{n_v}(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{\mu}_j(\mathbf{x}). \tag{3.5}$$

noting again the equivalence between the j-th tree estimate and the estimated mean of the end-node that contains $\mathbf{x}$ for tree j. For a given set of sample data, each $\lambda_j(\mathbf{x})$ and $\tilde{\mu}_j(\mathbf{x})$ are functions of the random partitioning process, so can be seen as M independent observations of the random vector, $(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$. Therefore, the expression given by (3.5) can be seen as a sum of products of the components of these M random vectors.

3.3 Estimate of bias from weights

Using this weighted average does increase the efficiency of the estimator, but also makes the estimator potentially biased under the sample design. In particular, if we explore the expectation of the weighted forest model with respect to the selection probability and the randomness of the recursive partitioning algorithm (Figure 3.1) denoted by E_{p^*}, then we get

$$\begin{aligned} E_{p^*}\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j(\mathbf{x})\right] &= \sum_{j=1}^{M} E_{p^*}\left[\lambda_j(\mathbf{x})\right] E_{p^*}\left[\tilde{h}_j(\mathbf{x})\right] + \sum_{j=1}^{M} \mathrm{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{h}_j(\mathbf{x})\right) \\ &= \tilde{h}^* E_p\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x})\right] + \sum_{j=1}^{M} \mathrm{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{\mu}_j(\mathbf{x})\right), \end{aligned}$$

where $E_{p^*}[\tilde{h}_j] = \tilde{h}^*$ for each j.

Therefore,

$$E_{p^*}\left[\sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{h}_j(\mathbf{x})\right] = E_\xi[Y \mid \mathbf{x}] + \sum_{j=1}^{M}\mathrm{cov}_{p^*}(\lambda_j(\mathbf{x}),\,\tilde{\mu}_j(\mathbf{x})),$$

because $\tilde{h}^*(\mathbf{x}) \to E_\xi[Y \mid \mathbf{X} = \mathbf{x}] = E_\xi[Y \mid \mathbf{x}]$ by Proposition 2.1 and $\sum_{j=1}^{M}\lambda_j(\mathbf{x}) = 1$ by design.

Since each $\lambda_j(\mathbf{x})$ and $\tilde{\mu}_j(\mathbf{x})$ are observations of random variables $\lambda(\mathbf{x})$ and $\tilde{\mu}(x)$, for a given value $\mathbf{x}$, the bias term is

$$\sum_{j=1}^{M}\mathrm{cov}(\lambda_j(\mathbf{x}),\,\tilde{\mu}_j(\mathbf{x})) = M\mathrm{cov}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})). \tag{3.6}$$

In order to correct for this bias for a fixed sample, we estimate $\mathrm{cov}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$ using the M observations by

$$\widehat{\mathrm{cov}}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})) \;=\; (M-1)^{-1}\sum_{j=1}^{M}(\lambda_j(\mathbf{x}) - \bar{\lambda})\,(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}), \tag{3.7}$$

where $\bar{\lambda} = M^{-1}\sum_{j=1}^{M}\lambda_j(\mathbf{x})$ and $\bar{\mu} = M^{-1}\sum_{j=1}^{M}\tilde{\mu}_j(\mathbf{x})$. Therefore the proposed forest estimator for the function $h(\mathbf{x})$ is

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{\mu}_j(\mathbf{x}) - M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})). \tag{3.8}$$

The following result is the main theoretical result of the article. It states that the proposed forest estimator, $\mathcal{F}_{n_v}(\mathbf{x})$, defined by equation (3.8) is asymptotically design-unbiased which converges in probability to $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{X} = \mathbf{x}]$. This result provides theoretical justification for using this random-forest estimation method on data collected from an informative sample design.

Proposition 3.1. *For a fixed $M > 0$, if Conditions 1 through 7 are satisfied for each tree in the forest, then*

$$\mathcal{F}_{n_v}(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}],$$

for all $\mathbf{x}$ as $n_v \to \infty$.

4. Relative performance of the estimators

In order to understand how the proposed random forest method compares to the typical i.i.d. random forest of Breiman (2001), we test the two methods over repeated samples of two publicly available datasets using two different sample designs, simple random sampling (SRS) and sampling with probability proportional to size (PPS). We evaluate the efficiency and bias of the predictions of each methods empirically and compare them to the standard Hájek estimator. The proposed random forest approach given by the algorithm in Figure 3.1, was tested using the algorithm available in the R-package *rpms* (Toth, 2024),

and for the method (RF), proposed by Breiman (2001), we use the algorithm available in the R-package *randomForest* (Liaw and Wiener, 2002).

For the finite populations, we use the two datasets described below. In each dataset description, we also identify the variable of interest, the predictor variables, and the variable used as the measure of size for the PPS sample design.

API The California Academic Performance Index (API) dataset available in the *survey* package (Lumley, 2020) contains data on 5,973 schools including the school average score on the API standardized test as well as demographic and administrative data about the school and the neighborhood it serves. We treat the school's average test score in the 2000 academic year as the variable of interest with five school level predictor variables including 1.) the average education level attained by the parents of the students in the school, 2.) the percentage of students that are English language learners, 3.) the percentage of students enrolled in a subsidized lunch program, 4.) the percentage of teachers with full qualifications, and 5.) whether or not the school was eligible for an awards program. To compare the methods on an informative PPS design, we sample proportionally by the size of the school's enrollment.

CEx A subset of the Consumer Expenditure Survey public use interview data file from the US Bureau Labor of Statistics which is available in the *rpms* package. This dataset contains information from 2015 on 45,308 households with total expenditures greater than \$0. We consider the total household expenditures for the current quarter as the variable of interest with five predictor variables including 1.) whether or not the household lives in a home which they own (with or with a mortgage), rent, or is part of student housing; 2.) the region in which the household is located; 3.) whether or not the household lives in an urban location; 4.) whether or not a member of the household currently earns a wage; and 5.) the age of the person identified as the primary earner of the household. To compare the methods using an informative PPS design, we sample proportionally to the size (number of residents) of the household.

Treating each dataset as a finite population, we take $D = 500$ repeated samples of size n from the population where $n = 600$ for API and $n = 1{,}000$ for CEx. For each random sample, we fit the random forest models with 500 trees to the sample data using the default settings of both algorithms and all the available predictor variables. The default settings require that each end node of every tree contains at least 5 observation. Using these models we predict the values of the variable of interest for every unit in the finite population and the predicted values are compared to the true values.

Specifically, for each sample s_l, $l = 1, \ldots, D$, we find the estimated model $\tilde{h}^{(l)}(\mathbf{x})$ and calculated the empirical mean error

$$b_l = \frac{1}{N} \sum_{i=1}^{N} (\tilde{h}^{(l)}(\mathbf{x}_i) - y_i), \tag{4.1}$$

the empirical mean relative error, $b_l / \bar{y}$, where $\bar{y} = N^{-1}\sum_{i=1}^{N} y_i$, and the empirical mean squared error

$$m_l = \frac{1}{N}\sum_{i=1}^{N}(\tilde{h}^{(l)}(\mathbf{x}_i) - y_i)^2. \tag{4.2}$$

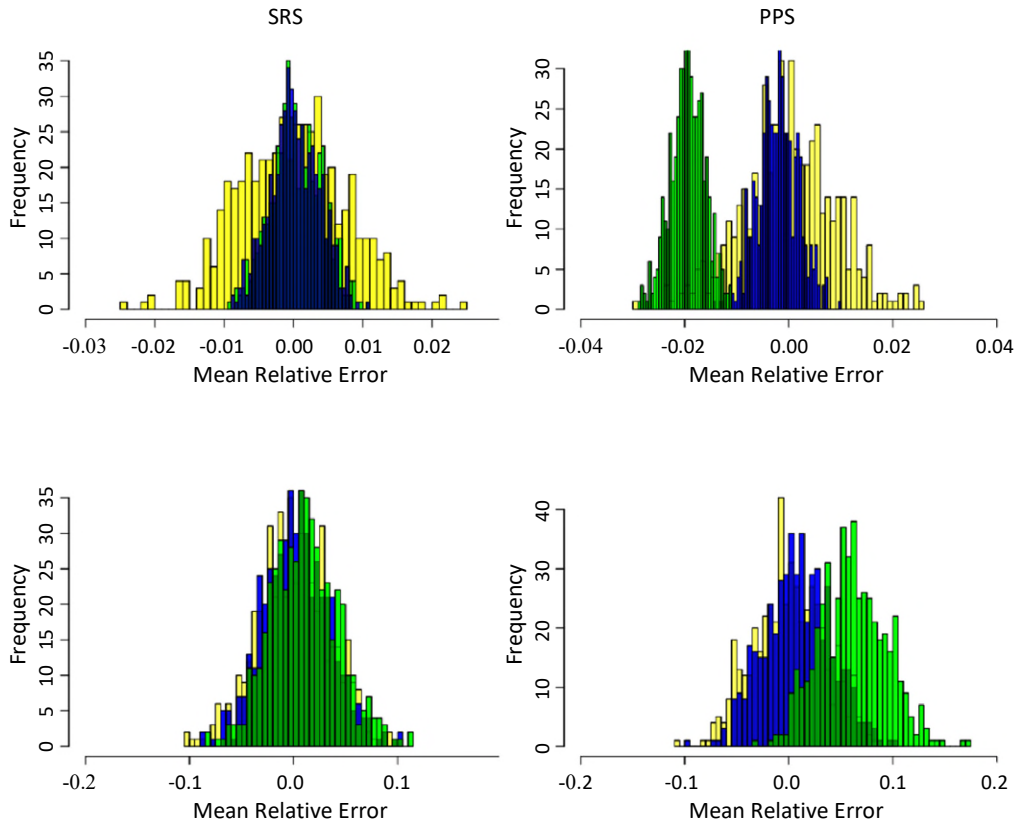
Notice the empirical mean error is an estimate of the average bias, which is defined as

$$\text{ABias}(\hat{m}_n(\mathbf{x})) := \mathbb{E}_{\xi}[\mathbb{E}_{\xi_p}[\hat{m}_n(\mathbf{x})] - m(\mathbf{x})], \tag{4.3}$$

where the first expectation is taken with respect to the joint distribution of the population and sample distribution and the second is with respect to the population distribution.

Since on of the biggest risk of ignoring an informative sample design when modeling a dataset is the introduction of bias in the model, it important to assess the potential bias of each of the estimators. The empirical distributions of the empirical relative mean errors over repeated samples for two data sets using two sample designs for the two forest algorithms as well as the Hájek estimator are shown in Figure 4.1.

Figure 4.1 Distribution of the mean relative errors of the three estimators over 500 repeated samples from the two datasets.



Note: The yellow histogram is the distribution of the RME for the Hájek estimator, the green is the unweighted random forest and the blue is the weighted forest method. The top graphs are the distributions of the mean relative errors using the API dataset and the bottom two using the CEx dataset.

When an SRS design is used, the two distribution of relative error given on the left side of Figure 4.1, show that all three estimators produce relatively unbiased estimates as their errors are centered very close to zero. In addition, the distribution of the errors of the two forest models have a smaller range than the Hájek estimator which shows that using these models leads to an increase in efficiency. This gain of efficiency is especially seen in the distributions using the API data.

Since the RF algorithm ignores the design weights, one would expect that there could be more bias from estimates of values using this model compared to values obtained from the Hájek estimator or the forest model using the algorithm in the *rpms* package, when the sample design is informative. The plotted distributions of the mean relative errors over repeated PPS samples, shown on the right hand side of Figure 4.1, confirm this. Under repeated PPS samples, the Hájek estimator still appears unbiased and the distribution of the relative errors are still wider than both forest models. Though the distribution of the relative errors of the RPMS model (in blue) appears to be centered close to zero, the relative errors of the RF model ignoring the weights is centered close to -2% for the API dataset and around 6% for the CEx dataset. This suggests that ignoring the weights leads to much more bias than the proposed method under the PPS design.

Table 4.1 contains the average relative mean errors and average mean squared error $\bar{m} = D^{-1}\sum_{l=1}^{D} m_l$ over the 500 random samples for each of the three models, the two datasets, and the two sample designs. The relative mean error statistic is presented as a percentage, $(\bar{b} / \bar{y})$ 100% and the mean square error statistic is given relative to that of the Hájek estimator, $\bar{m} / \bar{m}_H$, where $\bar{m}_H$ is the average mean squared errors of the Hájek estimator over the 500 samples.

Table 4.1
Averages over 500 random samples comparing the prediction error using the Hájek estimator and the two random forest methods on two datasets and for two sample designs.

	API N = 5,973 n = 600				CEx N = 45,308 n = 1,000			
	% Rel. Error		RMSE		% Rel. Error		RMSE	
Method	SRS	PPS	SRS	PPS	SRS	PPS	SRS	PPS
Hájek	-0.010	-0.007	1.000	1.000	-0.206	0.066	1.000	1.000
RF	0.043	-1.945	0.209	0.218	0.806	6.111	0.862	0.865
RPMS	0.021	-0.210	0.204	0.204	-0.056	0.878	0.844	0.844

Note: The percent relative error is the mean error of the estimated values for the full population relative to the population mean of the variable of interest, multiplied by 100. The relative RMSE is the mean over the 500 samples of the calculated mean squared error of the estimated values for the full population, relative to that of the Hájek estimator.
API = Academic Performance Index; CEx = Consumer Expenditure Survey; PPS = Probability proportional to size; RPMS = Recursive Partitioning for Modeling Survey Data; SRS = Simple random samples.

One can see from the results in Table 4.1 that relative mean squared error of the two estimators using the forest modeling methods are smaller than that of the Hájek estimator for both datasets under the SRS and PPS sample designs. However, the RF procedure that ignores the sample weights produces biased estimates under the PPS design for both datasets while both the proposed random forest modeling procedure and the Hájek estimator provide relatively unbiased estimates.

4.1 Demonstration of consistency

So far, we have been examining the efficiency and bias of our forest estimator compared to the usual i.i.d. random forest algorithm and the standard Hájek mean estimator by focusing on the difference between the model estimates and the true value of the variable of interest y using two real datasets as our finite population. An estimator $\tilde{h}(\mathbf{x})$ is consistent if

$$E_{\xi p}[(\tilde{h}(\mathbf{x}) - E_{\xi}[Y \mid \mathbf{x}])^2] \to 0 \text{ as } \nu \to \infty, \tag{4.4}$$

which requires knowing the true mean function $h(\mathbf{x}) = E_{\xi}[Y \mid \mathbf{x}]$ and letting the size of the sample and population go to ∞. When using a real dataset as the finite population, you don't know $h(\mathbf{x})$. Therefore, to explore consistency we use simulated data $(Y, \mathbf{X})_{i=1}^{N}$, where the values are obtained through random draws from a known distribution and we study the behavior of the estimators for a sequence of sample sizes.

For each randomly generated observation i we generate a random vector

$$\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, V_{i1}, V_{i2}, V_{i3})$$

of 6 independent random variables. Variables X_{i1} and X_{i2} both follow a uniform distribution $U(-10, 20)$, and $X_{i3} \sim U(-100, 200)$ while V_{i1} through V_{i3} are categorical random variables with equal probability among categories. Both V_{i1} and V_{i2} take one of 5 categories and V_{i3} takes one of 14 categories. These are the auxiliary variables available to the analyst for every unit in the population and can be used in the model for the variable of interest Y.

Because random forests are known to be very flexible, nonparametric models, rather than testing this method on a set of standard parametric models, we show the results for data that follows the mean model $y = \mu(x) + \epsilon$ where

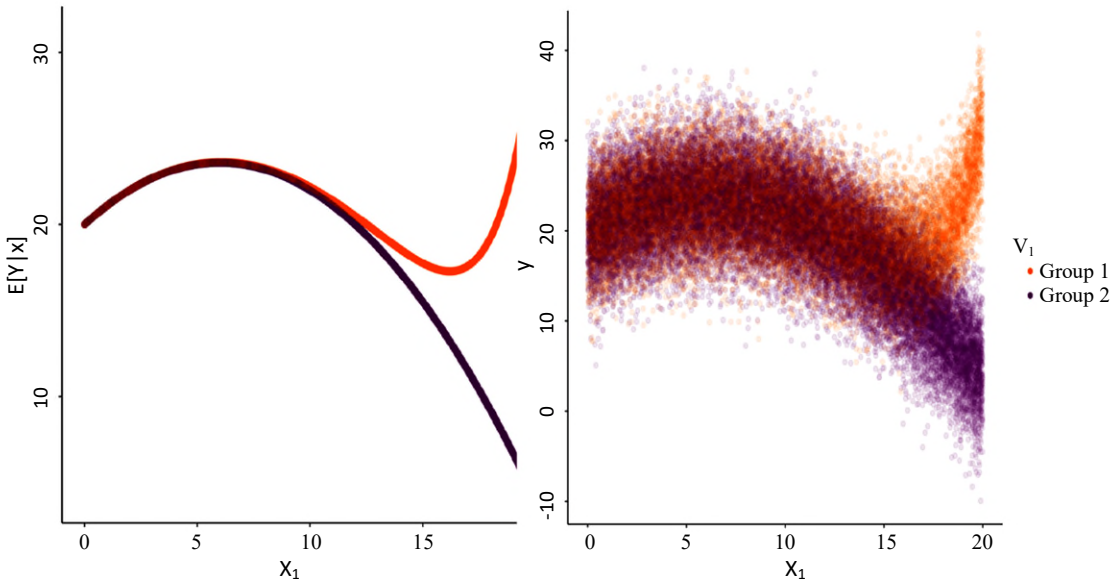
$$\mu(\mathbf{x}) = 0.2X_1(X_1 - 12) + 0.5\exp\{(X_1 - 15)\}\, \mathbb{I}_{\{V_1 \in \{A, B\}\}}.$$

The left side of Figure 4.2 shows the mean function, $\mu(\mathbf{x})$ and the right side shows a graph of the population values that were randomly generated from the model.

We also generate a variable Z, that we use for the size variable in order to test the methods under a PPS sample design. The values of the size variable Z are independently generated from the model $Z = \frac{1}{2}\mu(\mathbf{x}) + 5\eta$, where η has a chi-squared distribution with 5 degrees of freedom. The correlation between the size variable and Y is of 0.663, and so, in this example, the PPS design is informative.

For this simulation, we generate random finite populations with 1, 2, 4, 8, 16, and 32 thousand units. We then draw 500 repeated samples from each of these six finite populations. For each random sample we sample 5% of the units, which is 50, 100, 200, 400, 800, and 1,600 observations respectively. Again we use the sample data to estimate the forest model, use the forest model to predict the values of Y for the non-sampled units given the values of $\mathbf{X}$ in the population, and then use these values to estimate the population mean.

Figure 4.2 The values of $E[Y \mid \mathbf{x}]$ with respect to the variable X_1.



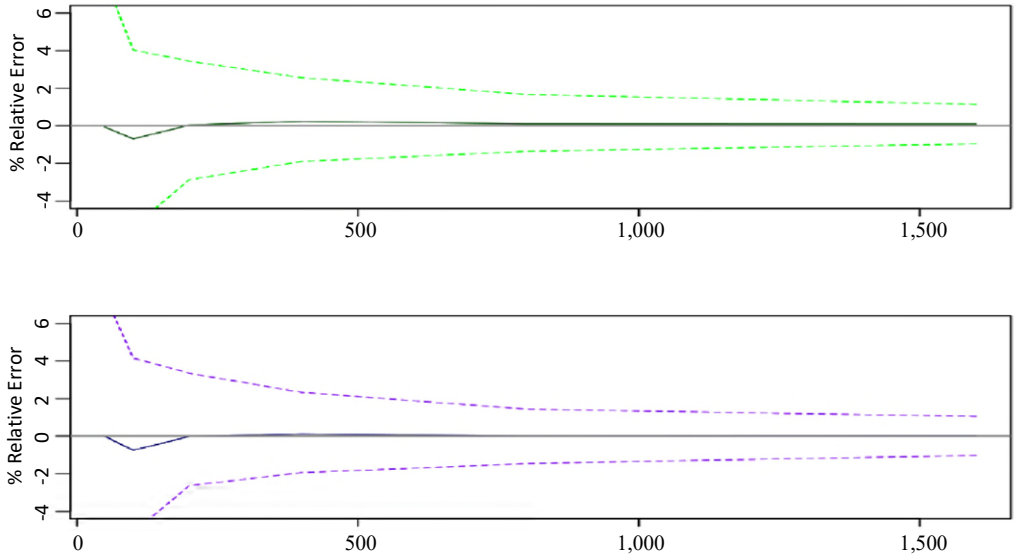
Note: The color denotes the values for the two groups of observations which is based on the value of the categorical variable V_1. Group 1 consists of observations where $V_1 \in \{A, B\}$ and Group 2 are all other observations.

By drawing random samples of increasing size and comparing our estimator to the true mean function, we evaluate the behavior of the accuracy and the variance of our estimator with respect to sample size. For instance, for a consistent estimator, one would expect the empirical confidence interval of average differences between the estimated values and values generated from the true mean function to contain zero. In addition, as the sample size increases, the variance of the average difference should decrease toward zero.

We can see that this is the case with the SRS design shown in Figure 4.3. shows the average over the 250 samples of the mean relative errors of the estimated population values as the sample size increases from 50 to 1,600. The distribution of the mean errors over the 500 repeated samples is centered right around 0 for every sample size for both forest modeling methods. In addition, the variance of the mean errors goes to zero as the sample size increases at about the same rate for both modeling methods.

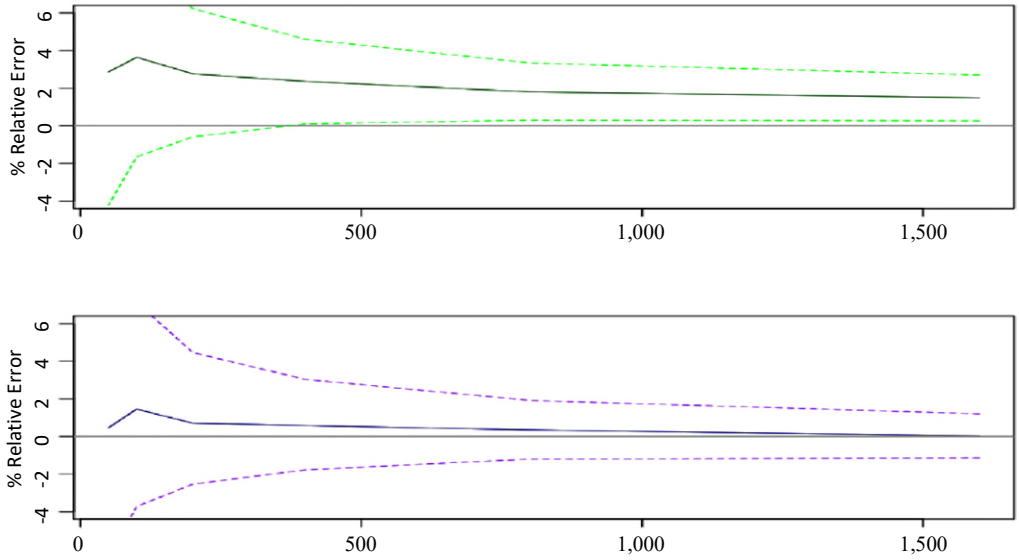
As we might expect, the story gets more interesting when the sample is drawn using a PPS design and the size variable is related to the variable of interest. In Figure 4.4, we see that the averages of the percent relative errors over the 250 repeated samples are longer as near zero for either method. Zero is contained within the middle 95% of values of percent relative errors when the sample size is below 800 for both methods because of the large variances in these values at small sample sizes. This interval no longer contains zero for the standard random forest algorithm as deviation of errors decrease with the increasing sample sizes. However, the proposed random forest method contains zero for every sample size for the proposed method.

Figure 4.3 Percent relative error by sample size for the regular (top) and the design consistent (bottom) random forest algorithms over repeated simple random samples.



Note: The solid line is the mean percent relative error over all the samples while the dashed lines give the 2.5 and the 97.5 percentile values.

Figure 4.4 Percent relative error by sample size for the regular (top) and the design consistent (bottom) random forest algorithms over repeated probability proportional to size samples.



Note: The solid line is the mean percent relative error over all the samples while the dashed lines give the 2.5 and the 97.5 percentile values.

These simulation results confirm the main result of the paper. That is, the proposed algorithm satisfying certain conditions that has been shown theoretically to be asymptotically design unbiased and consistent, has demonstrated over repeated samples to converge toward the true mean and be relatively unbiased.

5. Conclusions

Traditionally, complex survey data are collected to estimate finite population quantities. However, with the rise of machine learning methods, there is now greater interest in employing survey data in predictive problems and so the adaptation of machine learning methods to handle unequal probability data is now a vibrant area of research. In this paper, we present a new algorithm for estimating random forest models. This method which relies on independent random trees and a weighting procedure based on the weighted variability of the y-values is more appropriate for survey and other data collected from an informative design. We provide a set of conditions under which we show the method is design consistent for the conditional expectation of the variable of interest. The theoretical asymptotic unbiasedness and consistency of this algorithm is demonstrated through a simulation. The simulation studies are performed using real and generated data; we show that in practice the proposed method greatly reduces the bias of a random forest algorithm under an informative sample designs. In contrast, the estimates from the usual random forest method, which does not account for the sample design, are not unbiased under repeated informative samples. The estimates of both methods exhibited fairly similar mean squared errors. To ensure the individual trees are independent, our algorithm constructs truly random trees where a random variable and a random cut point are selected for each split.

Nalenz et al. (2024)'s approach of adjusting for an informative sample design is straight forward and interesting but it came out while this article was in review and so we have not compared it to our random trees method. Rather than avoid bootstrap sampling Nalenz et al. (2024) use a Hájek bootstrap. In their application this works very well as the outliers occur in the over sampled part of the population (units with low weights) and thus these units are down-weighted by the algorithm. However, in a general dataset, outlying values can also be associated with highly weighted survey units. Future work should compare the methods and, in the spirit of model aggregation, should consider blending the two approaches.

Acknowledgements

The authors would like to thank a lot of people later.

Appendix

Proofs and minor results

Proof of Corollary 2.1. Condition 1 requires the variable Y to have a finite fourth moment, then the random variable Y^2 has a finite second moment. Therefore, by taking Y^2 as the variable of interest in Proposition 2.1, we get a consistent estimator of $E_\xi[Y^2 \mid \mathbf{x}]$.

Lemma 6.1. *For a given tree j, if Conditions 1 through 7 are satisfied, then*

$$\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x}) < \infty, \text{ as } \nu \to \infty.$$

Proof of Lemma 6.1.

$$\begin{aligned}
\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} &= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i - \tilde{\mu}_j(\mathbf{x})\right)^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i^2 - 2y_i\tilde{\mu}_j(\mathbf{x}) + \tilde{\mu}_j^2(\mathbf{x})\right) \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&\quad - 2\tilde{\mu}_j(\mathbf{x}) \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} + \tilde{\mu}_j^2(\mathbf{x}) \\
&= \underbrace{\left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}}}_{I} \quad - \quad \underbrace{\tilde{\mu}_j^2(\mathbf{x})}_{II}
\end{aligned}$$

by equation (2.1).

By Proposition 2.1, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$, so the term $II \to_p E_\xi^2[Y \mid \mathbf{x}]$ for each $j = 1 \ldots M$ and by Lemma 2.1,

$$I = \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \to_p E_\xi[Y^2 \mid \mathbf{x}],$$

for each $j = 1 \ldots M$. Therefore, $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$, and by Condition 1, both the quantity I and II are finite with ξ probability 1.

Lemma 6.2. *If Conditions 1 through 7 are satisfied for each tree in a given forest with $M > 0$ trees, then*

$$\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$$

for all $\mathbf{x}$ *and all* $j = 1 \ldots M$, *as* $\nu \to \infty$.

Proof of Lemma 6.2. Because, for each $j = 1, \ldots, M$, the random variable $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \geq 0$, by equation (3.3), the function

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}$$

is continuous for each j. Also, by Lemma 6.1, $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$ for each j. Therefore, by the Continuous Mapping Theorem,

$$\lambda_j(\mathbf{x}) \to_p \frac{(\mathrm{Var}(Y \mid \mathbf{x})) + 1)^{-1}}{\sum_{j=1}^{M} (\mathrm{Var}(Y \mid \mathbf{x}) + 1)^{-1}} = \frac{1}{M}.$$

This lemma states that because the variance converge to $V[Y \mid \mathbf{x}]$ and so all end-node weights converge to $1/M$ asymptotically. However, note that for finite n, the weights will differ substantially depending on the efficiency gain from the random splits resulting in the end-node. This is the adaptiveness that has been added to the procedure and where all the real work is being done.

Lemma 6.3. *For a fixed $M > 0$, if Conditions 1 through 7 are satisfied for each tree in the forest of M trees, then*

$$M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0,$$

for all $\mathbf{x}$ *as* $\nu \to \infty$.

Proof of Lemma 6.3. Since $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by Lemma 6.2, $\bar{\lambda} = M^{-1}\sum_{j=1}^{M}\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by the Continuous Mapping Theorem. Likewise, $\bar{\mu} = M^{-1}\sum_{j=1}^{M}\tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ since each $\tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ by Proposition 2.1. Therefore, each of the terms of equation (3.7),

$$(\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}) \to_p 0.$$

Once again applying the Continuous Mapping Theorem to

$$M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) = \frac{1}{(1 - M^{-1})}\sum_{j=1}^{M}(\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}),$$

gives the result.

Proof of Proposition 3.1. Since each $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by Lemma 6.2, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$ by Proposition 2.1 and $M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0$, by Lemma 6.3, we can apply the Continuous Mapping Theorem to get that

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{\mu}_j(\mathbf{x}) - M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p \sum_{j=1}^{M}\frac{1}{M}E_\xi[Y \mid \mathbf{x}] + 0 = E_\xi[Y \mid \mathbf{x}].$$

References

Arlot, S., and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9).

Bilton, P., Jones, G., Ganesh, S. and Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, 115, 53-66.

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Buskirk, T.D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Surv Pract*, 11, 2709.

Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 1-18.

Earp, M., Toth, D. Phipps, P. and Oslund, C. (2018). Assessing nonresponse in a longitudinal establishment survey using regression trees. *Journal of Official Statistics*, 34(2), 463-481.

Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T. and Bajaj, C. (2020). Weighted random forests to improve arrhythmia classification. *Electronics*, 9(1), 99.

Gelman, A., King, G. and Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443), 846-857.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361-74.

Hong, H.G., and He, X. (2010). Prediction of functional status for the elderly based on a new ordinal regression model. *Journal of the American Statistical Association*, 105(491), 930-941.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

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

Krebs, M.A., Reeves, M.C. and Baggett, L.S. (2019). Predicting understory vegetation structure in selected western forests of the United States using fia inventory data. *Forest Ecology and Management*, 448, 509-527.

Kshirsagar, V., Wieczorek, J., Ramanathan, S. and Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach. *arXiv preprint arXiv:1711.06813*, 2017.

Lavallée, P., and Beaumont, J.-F. (2015). Why we should put some weight on weights. *Survey Methods: Insights from the Field (SMIF)*.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18-22. URL: https//CRAN.R-project.org/doc/Rnews/.

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(466), 546-556.

Loh, W.-Y. (2008). Classification and regression tree methods. *Encyclopedia of Statistics in Quality and Reliability*, 1, 315-323.

Lumley, T. (2020). survey: analysis of complex survey samples, 2020. R package version 4.0.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

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

Morgan, J.N., and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415-434.

Nalenz, M., Rodemann, J. and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 113(6), 3379-3398.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 772-794.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146, 72-83.

Shahhosseini, M., and Hu, G. (2020). Improved weighted random forest for classification problems. *International Online Conference on Intelligent Decision Science*, 42-56. Springer.

Toth, D. (2024). *rpms: Recursive Partitioning for Modeling Survey Data*. R package version 1.0.0.

Toth, D., and Eltinge, J. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106, 1626-1636.

Wieczorek, J. (2023). Design-based conformal prediction. *Survey Methodology*, 49, 2, 443-473. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00007-eng.pdf.

Williams, J.K., Neilley, P.P., Koval, J.P. and McDonald, J. (2016). Adaptable regression method for ensemble consensus forecasting. *Thirtieth AAAI Conference on Artificial Intelligence*.

Winham, S.J., Freimuth, R.R. and Biernacka, J.M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6), 496-505.

Yang, D.K., and Toth, D.S. (2022). Analyzing the association of objective burden measures to perceived burden with regression trees. *Journal of Official Statistics*, 38(4), 1125-1144.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 209-234
Statistics Canada, Catalogue No. 12-001-X

Adaptive cluster sampling, a quasi Bayesian approach

Glen Meeden and Muhammad Nouman Qureshi[1]

Abstract

Adaptive cluster sampling designs were proposed as a method that could be used when sampling rare populations whose units tend to appear in clusters. The resulting estimator is not based on any model assumptions and is design unbiased. It can have smaller variance than the standard estimator which does not incorporate the fact that one is dealing with a rare population. Here we will demonstrate that, when adaptive cluster sampling is appropriate, its estimator does not take into account all the available information in the design. We present a quasi Bayesian approach which incorporates the information which is now ignored. We will see that the resulting estimator is a significant improvement over the current methods.

Key Words: Adaptive cluster sampling; Bayesian inference; Finite population sampling; Prior information.

1. Introduction

Consider the problem of estimating the total number of a plant or animal species that live in a specified geographical area where the area has been partitioned into a collection of equally sized squares. Furthermore, assume that the species of interest is rare in this area so that most of the squares will contain none of the species. In addition assume that the few squares that do contain some of the species tend to cluster together in just a few neighborhoods of adjacent squares.

For this problem Thompson (1990) introduced the notion of adaptive cluster sampling, ACS. An initial simple random sample of squares is taken and the number of the species in each selected square is observed. Most of these observed counts will typically be zero. But whenever the count in a square is greater than zero, the adjacent squares, those to the left, right, above and below are added to the sample. When any of these squares have a count greater than zero then all of its unobserved adjacent squares are also observed. This process is continued until we obtain a set of contiguous nonempty squares surrounded by empty squares. A set of contiguous nonempty squares is called a network and the surrounding empty squares its edges. By definition an empty square is a network of size one. For this adaptive cluster sampling plan the usual estimator of the population total based on the counts in all the observed squares will be biased upwards. Thompson (1990) developed an unbiased estimator for the population total along with an estimator of its variance. More detail along with more references can be found in Thompson (2012). Many field researchers have adopted various versions of adaptive cluster sampling and it is used in a variety of disciplines. Turk and Borkowski (2005) describes several such examples. Latpate, Kshirsagar, Gupta and Chandra (2021) discuss some modifications of the standard ACS approach.

In the Bayesian approach to survey sampling prior information about the population of interest is incorporated in a prior distribution. After units in a sample have been observed inferences about the population are based on the posterior distribution of the unobserved units given the observed units. Moreover, this posterior distribution does not depend on how the units in the sample were selected. The

1. Glen Meeden, Emeritus, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: gmeeden@umn.edu; Muhammad Nouman Qureshi, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: qures089@umn.edu.

details of this approach have been described by Basu (Ghosh, 1988). Three Bayesian approaches to adaptive cluster sampling are given in Rapley and Welsh (2008), Pacifici, Reich, Dorazio and Conroy (2016) and Goncalves and Moura (2016). In these approaches the authors construct a Bayesian model for possible populations consistent with the assumptions underlying adaptive cluster sampling. Nolau, Goncalves and Pereira (2022) introduce a Bayesian model which includes auxiliary variables which could include additional information about the count in a square.

Given an ACS sample, our goal is to find a point estimator and an upper bound for the total number of the species in the population which have good frequentist properties. Since we are considering rare species we will assume that the ratio of the number of units with counts greater than zero to the total number of units in the population is small. We will break the problem into two parts. First we specify a prior distribution for the number of squares or units in the population with counts greater than zero, say θ an unknown parameter. This prior will reflect our assumption that the number of such units is small. Given the sample and our prior we have *a posterior* for θ. So our first step is to simulate a possible value for θ, say $\hat{\theta}$. Then conditional on $\hat{\theta}$, we find an estimate for the total of all counts greater than zero by using a distribution which assumes exchangeability among all the observed and unobserved counts greater than zero. This distribution does not arise from a prior distribution but is specified after the ACS sample has been observed. This is not a standard Bayesian procedure because this second "posterior" distribution does not follow from any prior distribution defined on the unknown finite population. This explains the terminology "quasi Bayesian" in our title. Two other recent examples where inferences are based on pseudo posterior distributions are given in Si, Pliiai and Gelman (2015) and Savitsky and Toth (2014). This makes sense when there is information in the sampling design which cannot be incorporated into a prior distribution. But we then combine these two distributions to simulate complete copies of the unknown population. We will see that the resulting point and interval estimators of the population total have better frequentist properties than the standard ACS estimators.

In Section 2 we briefly review the adaptive cluster sampling approach and outline our approach to the problem. In Section 3 we explain our approach in detail and present our estimators. We developed our estimator by doing simulations on a set of six populations where ACS sampling would be appropriate. In Section 4 we describe these six populations. In Section 5 we present simulations to compare our approach to the standard ACS approach. This is done for the six populations in Section 4 and six new populations which were not used in the development of our method. In Section 6 we discuss some possible extensions when more prior information is available about the population of interest. Section 7 contains some concluding remarks.

2. Adaptive cluster sampling

2.1 The basic setup

We begin by introducing some notation. We assume that the unknown population is a rectangular area consisting of N_r by N_c squares or units. So $N = N_r \times N_c$ is the population size. For integers (i, j) where $1 \le i \le N_r$ and $1 \le j \le N_c$ let $y_{i,j}$, denote the number of the species in the $(i\text{th}, j\text{th})$ square. Note $y_{i,j}$ is a

non negative integer. Let Y denote the matrix of the $y_{i,j}$ values. Given a square, the neighbors of a square are those squares just above and below it and those squares just to the right and left of it, with obvious modifications for squares on the boundary of the population.

In adaptive cluster sampling for each square in the initial random sample with a y value greater than zero all its neighbors are also observed and if any of them have a y value greater than zero then their neighbors are also observed and so on. This process is continued until only zero values are observed. For a given square the resulting set of squares found this way with y values greater than zero is called a **network**. Hence a network consists of a set of non zero squares with the property that if one of them appears in the sample then all of them will. The set of squares with a y value of zero that were observed in the process are called the edges of a network. As we noted in the introduction the network for a square whose y value is zero is just itself. For square (i, j) let $\Psi_{i,j}$ be all squares in its network.

Suppose now an initial simple random sample without replacement of size n_1 is taken. For $k = 1, \ldots, n_1$ let Ψ_{i_k,j_k} denote the network of the square which appears on the k th draw of the sample. Note that if two squares, which are in the same network, are in the first sample then their networks are identical and they each will be included when estimating the population total. For each k we let m_k be the number of squares in Ψ_{i_k,j_k} and $\bar{y}_k^*$ the mean of the counts for the units that appear in Ψ_{i_k,j_k}. For ACS (adapted cluster) sampling Thompson (1990) gave an unbiased estimator of the population total and an unbiased estimator of its variance. Given an ACS sample this estimate, $\hat{T}_{ac}$ and its estimated variance, $\hat{v}_{ac}$ are given by

$$\hat{T}_{ac} = N\frac{\sum_{k=1}^{n_1}\bar{y}_k^*}{n_1} \quad \text{and} \quad \hat{v}_{ac} = \frac{N(N-n_1)}{n_1}\frac{\sum_{i=k}^{n_1}(\bar{y}_k^* - \hat{T}_{ac}/N)^2}{n_1 - 1}. \tag{2.1}$$

We see that the ACS estimators, point and interval, depend only on the network means. This means that the variability of counts within a given network plays no role. In effect it is assuming that all the counts within a given network are the same. In addition, the fact that there were just a few squares with counts greater than zero in the population seems to play no explicit role in the inference stage after the ACS sample has been selected.

There is an alternative way to think about the ACS estimator in the above equation which is described in Dryver and Chao (2007). They consider a second but related version of the population. To form this second population they proceed as follows. For each network in the population we replace each y value by the average of all the y values in that network. If a network contains just one square then its y value is unchanged. But if a network contains more than one square then each of its y values are changed to the mean of all the y values of the squares making up the network.

Clearly this alternative population has the same total as the original population. Moreover, the observed y_k values in the second population, for the units in the initial sample, are identical to the $\bar{y}_k^*$ in equation (2.1). This makes it clear that the ACS estimator is an unbiased estimator but it also makes it clear that at the inferential stage the ACS estimator does not make use of the fact that we are sampling from a rare population.

Finally, some might consider it surprising that $\hat{T}_{ac}$ depends on N and that it includes the networks with a mean of zero. But a similar thing happens when estimating a domain total when the domain size is unknown and one has a random sample from the entire population. See for example the discussion on domain estimation in Cochran (1977).

2.2 A new approach

The ACS approach is based on two basic assumptions; there are only a few squares with counts greater than zero and these squares tend to be grouped in clusters. Although the sampling design is based on these two assumptions, as we have just noted at the end of previous section, the ACS estimator never seems to make use of the information that the proportion of squares with counts greater than zero is small at the inference stage. One should be able to do better if one incorporates this information when constructing an estimator.

Let D_b be the set of all the units in the population with counts greater than zero and let θ be the number units in D_b. For us θ is an unknown parameter and it will play an important role in what follows. Let T_b be the total of all the units in D_b. Of course T_b is also the total of all the units in the population. But we introduce this notation to emphasize the fact that our approach focuses on D_b. We break the problem into two parts. In the first part we find an estimate for θ using the information contained in the initial random sample of size n_1. Given this estimate and all the counts in the observed networks we then find an estimate for T_b.

As far as we know the notion of rare has never been explicitly defined in the literature. In some sense it is the analog for a few items of the Sorites paradox for many items. One version of which is "How many stones are needed to be considered a pile?". If we remove a single stone from a pile it should still be considered a pile but if we repeat this enough times we will no longer have a pile. Similarly if a species is rare in our population adding one positive count to a zero square would not change our perception that it is rare. But if we add enough positive counts then the species would no longer be rare. Let K_θ be the largest integer less than or equal to $N/10$. We will begin by assuming that our parameter space for θ is the set of integers

$$\Theta = \{i\colon i = 0, i = 1, \ldots, i = K_\theta\}. \tag{2.2}$$

One could argue that this choice is somewhat arbitrary but it is consistent with the notion of rarity and it holds for many of the examples considered in the literature. Later we will see that we can relax this assumption.

We do not make any other assumptions about the population except that the squares or units with counts greater than zero tend to appear in clumps and form networks. Where the networks are located in the population plays no role when ACS sampling is used. We make no explicit assumption about the range of possible counts. Anyone who wants to use ACS sampling and assumes that the number of positive counts is rare, as we described in the previous paragraph, could use our approach.

To develop our estimator we considered six possible populations. Three of them have appeared in the literature and the other three we constructed. They will be described in detail in Section 4 when we discuss our simulation results. Our estimator was found by trial and error by doing simulations from these six fixed populations.

It needs to be emphasized that finding sensible point and interval estimators for T_b is not an easy problem when θ is small. In such cases an ACS sample may contain zero or one or two counts greater than zero. In such cases we believe, finding sensible estimators is only possible when one has additional prior information. But here we are assuming that no such information is in hand. In the spirit of our objective approach we decided to ignore such samples in our simulation studies.

In the next section we will present our estimators and will explain the underlying logic and intuition that led to them.

3. A new estimator

Recall that D_b is the subset of Y consisting of all the squares with a y value greater than zero. If T is the population total of Y then T is equal to T_b, the total of the units belonging to D_b. Our goal is to estimate T_b. Remember that θ denotes the size of D_b, the number of $y_{i,j}$'s in the population which are greater than zero. For us θ is an unknown parameter and we take as its parameter space, Θ, the set of integers which are greater than or equal to zero and less than or equal to K_θ.

Let X be the number of units that belong to D_b in the first simple random sample without replacement of size n_1. Then X has a hypergeometric distribution which depends on N, n_1 and the unknown parameter θ. Given $X = x$ we can use the resulting likelihood function when estimating θ. Let y_b be the values of all the units in the ACS sample with values greater than zero. Let n_b be the number of units in y_b. Note that $n_b \geq x$. Let $y_{b'}$ be the remaining members of D_b which were not observed in the ACS sample. Note that if $\theta = n_b$ then $y_{b'}$ is the empty set.

From the Bayesian perspective one can break the problem of estimating T_b into two stages. First one simulates a possible value for θ from its posterior. Then given this value, one simulates possible counts for the $\theta - n_b \geq 1$ set of possible values for the unobserved values making up $y_{b'}$. Combining this simulated set of values with the observed y_b we have generated one complete copy of D_b and its corresponding value of T_b. Repeating these two steps many times we can use the resulting totals, in a Bayesian way, to find a point estimate and upper bound for the total of the units belonging to D_b.

We will now present our prior distribution for θ and explain how we simulate complete copies of D_b after the ACS sample has been observed.

3.1 Estimating the number of counts greater than zero

Given $X = x$ one natural estimate of θ is the maximum likelihood estimate. But one sees in simulations that the likelihood function tends to give too much weight to larger values of θ for the smaller values of x

we are likely to observe with ACS sampling. An alternative would be the standard Bayesian approach where one could specify a prior distribution over Θ. One possible non informative prior would be the uniform distribution over Θ, defined in equation (2.2). Its prior expectation is approximately $N/20$ which could be a good default guess for the value of θ. However, simulations demonstrate that it has the same weakness as the maximum likelihood estimator. They both tend to overestimate θ for ACS samples.

Our goal was to find a good default distribution for θ that will work well for a variety of populations where ACS sampling would be used. Rather than putting our distribution on Θ we considered distributions on Θ/N, the population proportion of $y_{i,j}$'s greater than zero. Any such distribution can then be thought of as a distribution on θ.

Recall that the beta distribution with parameters $\alpha > 0$ and $\beta > 0$ is defined on the unit interval and its density function is given by

$$f_{\alpha,\beta}(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1} \quad \text{for} \quad 0 \le z \le 1. \tag{3.1}$$

If for any choice of the parameters α and β we take the values of the above function, for all the members of θ in Θ, and then divided them by their sum the resulting set of numbers is a probability distribution defined on Θ. After much experimentation on our six test populations we chose as our prior distribution on Θ the re-normalized beta density with $\alpha = 1$ and $\beta = 90$.

To help understand this choice consider the case when $N = 400$. The prior mean for our choice is 3.92 while the choice of $\alpha = \beta = 1$ has a prior mean of 20. When the true value of θ/N is around 0.02 or 0.03 the later choice will give an over estimate of the size of θ. Our choice will help to decrease the upward bias of the likelihood function for the initial random sample. This will be discussed in more detail when we present the simulations in Section 5.

This is related to a similar problem with the ACS estimator $\hat{T}_{ac}$ (defined in equation (2.1)) that we saw in our simulations. Under the ACS design an important property of $\hat{T}_{ac}$ is that it is an unbiased estimator of the population total. If n_1, the size of the original random sample, is small and θ is small then there is a non-trivial probability that one observes samples where each count is zero. For such samples, $\hat{T}_{ac}$ gives an estimate of zero for the population total. This will be an underestimate except when the population total is in fact zero. To compensate for this underestimate, we saw that $\hat{T}_{ac}$ overestimates when the sample contains "lots" of units with counts greater than zero. It gets closest to the true value of θ when the number of units in the sample, with counts greater than zero, is close to $\theta/2$.

We need to emphasize that our estimator of θ does not depend on the fact that our basic units are squares and that the population is a rectangle. In addition it is not based on any assumptions about the form or shape of the networks or the values of the counts within a network. It depends only on the fact that we have a random sample from the basic units of the population. It should work whenever the basic units are approximately of the same size and the notion of neighbor can be defined in a sensible fashion. The overall form of the configuration of the population is immaterial. In addition it is not based on any assumptions about the values in y_b, the counts in the ACS sample greater than zero.

3.2 Estimating the population total

Given a value for θ, say $\hat{\theta}$, generated from our posterior distribution, and n_b the number of observed counts in y_b we need to define a distribution which simulates possible values for the remaining $\hat{\theta} - n_b \geq 1$ units whose counts must be greater than zero. We are assuming that little is known about the shape of the networks and how the size of a network may affect its y values. In this case a simple assumption is to assume that y_b, all the counts greater than zero in the full ACS sample, is approximately a "representative" sample of the values in D_b. Under this assumption a sensible way to simulate the unobserved members of D_b is to use Polya sampling.

Specifically, place n_b balls into an urn where each ball represents one of the counts belonging to y_b. Give each of the balls the weight $w > 0$. Pick a ball at random from the urn. Return it to the urn along with another ball which is assigned the count of the selected ball. This new ball is given the weight one. Now another ball is selected from the urn, which now has $n_b + 1$ balls, with probability proportional to their weights. The selected ball is returned to the urn along with another ball whose count is equal to that of the selected ball. This new ball is given a weight of one and the urn now contains $n_b + 2$ balls. This process is continued until the urn contains $\hat{\theta}$ balls. Under this distribution the expected value to the total of all the counts in the simulated urn does not depend on w and is $\hat{\theta}\bar{y}_b$, where $\bar{y}_b$ is the mean of the counts in y_b. The variance, however, does depend on w and decreases as w increases. It is demonstrated in Meeden (1999), (equation 2.5), that the variance is given by

$$\text{Var}(T_b \mid y_b, n_b, \hat{\theta}, w) = \hat{\theta}(\hat{\theta} - n_b)\frac{\text{var}(y_b)}{n_b}\frac{n_b - 1}{1 + n_b w}\frac{\hat{\theta} + n_b w - n_b}{\hat{\theta}} \tag{3.2}$$

where $\text{var}(y_b)$ is the sample variance of the counts in y_b.

Now we need to specify a value for w, the weight of each ball in the urn at the beginning of this process. Note that w only appears in the last two fractions in equation (3.2). It is easy to check that the value of the product to these two fractions is one if we take as our value of w

$$w^* = \frac{n_b(\hat{\theta} - n_b + 1) - 2\hat{\theta}}{n_b(\hat{\theta} - n_b + 1)}. \tag{3.3}$$

It is easy to check that $w^* > 0$ when $n_b > 2$.

With this choice of w^* and for a fixed y_b and fixed value of $\hat{\theta}$ our conditional variance is given by

$$\text{Var}(T_b \mid y_b, n_b, \hat{\theta}, w^*) = \hat{\theta}(\hat{\theta} - n_b)\frac{\text{var}(y_b)}{n_b}. \tag{3.4}$$

This reflects our assumption that we are viewing the values in y_b as exchangeable and arising from something approximating a random sample. Now if n_b is reasonably large, say around 20, and y_b is approximately a random sample then the above should be a reasonably good estimate of variance. But for smaller values of n_b, say less than five, it will not work well. In the following we will describe a way to handle this problem.

But first let $\hat{T}_{ab}$ denote our quasi or approximate Bayes estimator of T_b, the total number of the species in the population, that is based on our two stage simulation procedure. Let $\overline{y}_b$ be the mean of all the units in y_b. Then our estimate of T_b is

$$\hat{T}_{ab} = E(T_b) = E(E(T_b \mid \theta)) = E(\theta \overline{y}_b) = \hat{\theta}_{1.90} \overline{y}_b \tag{3.5}$$

where $\hat{\theta}_{1.90}$ is the mean of our posterior distribution for θ.

To find the variance of $\hat{T}_{ab}$ we use the well known formula that a variance can be written as the variance of a conditional expectation plus the expectation of a conditional variance. So for a fixed ACS sample we have that

$$\begin{aligned} \mathrm{Var}(\hat{T}_{ab}) &= \mathrm{Var}\Big(E(\hat{T}_{ab} \mid \theta)\Big) + E\Big(\mathrm{Var}(\hat{T}_{ab} \mid \theta)\Big) \\ &= \mathrm{Var}(\theta \overline{y}_b) + E\left(\theta(\theta - n_b)\frac{\mathrm{var}(y_b)}{n_b}\right) \\ &= \overline{y}_b^2 V(\theta) + \Big(E(\theta^2) - n_b E(\theta)\Big)\frac{\mathrm{var}(y_b)}{n_b}. \end{aligned} \tag{3.6}$$

We still need to find a way to use the variance of our estimator to produce a good interval estimator of the total number of the species in the population. In ACS sampling the initial random sample will consist of mostly zero counts. Recall x is the number of counts greater than zero in the initial sample of size n_1 and n_b is the number of units in the full ACS sample with counts greater than zero. Note that x is less than or equal to n_b. A very naive upper bound would be our point estimate plus the product of 1.96 and the square root of $\mathrm{Var}(\hat{T}_{ab})$. But, as we have already noted, it is no surprise that this works poorly because in ACS sampling the values of x and n_b can be quite small.

Consider a case where a population contains one or two networks of size one or two with counts much larger than the rest of the counts in the population. For such populations, when x is small, these networks are unlikely to be observed and our estimate of variance will be too small. To help protect against this possibility we need to increase the naive upper bound given just above, especially when n_b is small. To this end we let

$$\lambda = 10^{(2.5/n_b)}. \tag{3.7}$$

Note that λ will decrease as n_b increases. We are not claiming that this is an optimal choice for adjusting upwards our upper bound. We just found that it seemed to work well for our test populations.

When calculating upper bounds for our estimate we will use

$$\hat{T}_{ab} + \lambda \times 1.96 \sqrt{\mathrm{Var}(\hat{T}_{ab})} \tag{3.8}$$

as our approximate 95% upper confidence bound. Because of the small sample sizes in ACS sampling assuming that $\hat{T}_{ab}$ has, approximately, a normal distribution might seem surprising. But we will see in the

simulations that this seems to work reasonably well for most cases where ACS sampling is appropriate because of our choice for λ.

The second part of our two stage simulation process is clearly not Bayesian since our choices for w^* and λ depend on n_b which comes from the observed data . It does not arise from some prior distribution although the fact that the initial ACS was a random sample is important information for us. As we noted in the Introduction two other examples where "posterior" distributions are defined without a prior distribution and which are based in part on the sampling design can be found in Si et al. (2015) and Savitsky and Toth (2014). This seems to make sense when there is available prior information that cannot be incorporated into a prior distribution.

A sensible lower bound for T_b is just the sum of all the ACS sample counts greater than zero. For the ACS estimator we will use the same lower bound and for its upper bound we will use its estimate plus the product of 1.96 and its estimated variance. Again assuming that the ACS estimator has, approximately, a normal distribution.

One might object that our assumption that y_b is approximately a "representative" sample of the values in D_b is too strong. But recall that at the end of Section 2.1 we saw that the only information used in the ACS estimator is the mean of the counts in a network and by considering the alternative population of Dryver and Chao (2007), we saw that the ACS estimator is essentially assuming exchangeability among the network means. We believe that all the observed counts can contain some additional information and that our assumption that the observed sample is approximately a representative sample is no stronger than assuming the exchangeability among the network means.

The way we found our estimator was by doing simulations on a set of six populations where ACS sampling would be appropriate. Three of them had appeared in the literature and the other three we constructed. We then used simulation studies to see how different choices of a prior and an estimated variance would work. What we have described above is the best that we have found. We found others that worked almost as well but these are our best choices at this time. In the next section we describe these six populations.

4. The populations

In our simulations we used three different populations that have appeared in the literature. One is the first example discussed in Thompson (1990). He presents it as a typical example where ACS sampling could be used. He gives no further details so presumable it was constructed by him. The population has 400 units with three networks whose sizes are 6, 11 and 4 and whose means are 6.0, 9.73 and 11.75 respectively. The mean of all the counts greater than zero is 9.4 and the largest count is 39 which appears in the network of size 11. We denote this population by thmp. Gattone, Mohamed and Di Battista (2016) describe two samples

of African buffalo and African hartebeest taken in 2010. We denote these two populations by afrbuf and afrhart.

We also constructed three more populations. To construct a population we considered a grid of points on the surface of the earth. Assume that their latitudes and longitudes are equally spaced, although the successive differences in the two directions need not be the same. Depending on the topography of the location of the points, their altitudes, measured in meters, can exhibit the clumping behavior which is the underlying assumption of ACS sampling.

To find the altitude on a grid we used the function, *elevation*, in the R (R Core Team, 2023) package rgbif (Chamberlain, Ram, Mcglinn and Barve, 2019). To get the final set of "counts" we did three things. First we rounded every altitude in the set to its closest integer value. Next we choose an $\epsilon > 0,$ but close to zero, and found the $1-\epsilon$ percent quantile of our set of integer values, say q_ϵ. We set every count less than q_ϵ to zero. Finally, we subtracted some integer, less than $q_\epsilon,$ from every count greater than zero. The number of resulting counts greater than zero would then be small enough to represent counts of a rare species. For example, if we set $\epsilon = 0.05$ the resulting set of counts will have five percent of it's values greater than zero. So in the resulting population, the counts or the $y_{i,j}$ values, are either zero if their rounded altitude falls below a certain level or the difference between their rounded altitude and some constant. Depending on the topography of an area, the resulting grid of "counts" can exhibit the clumping behavior that makes ACS sampling a sensible choice. This is a flexible method that allows one to find many realistic populations to use in simulation studies of ACS sampling.

Next we describe the three grids we used to construct the three populations used in our simulation studies. For the first we chose a grid in Paris, France, the second a grid at Niagara Falls, on the border between Canada and the United States and the third a grid near Devil's Tower in the western United States. For the first two we used a 23 by 23 grid while for the third we used a 26 by 45 grid. We denote these three populations by paris, nfalls and devt. Summary information for the six populations are give in Table 4.1. Note the proportion of counts greater than zero range from 0.038 to 0.084. Three dimensional plots of the populations are given in the six figures on the next two pages.

Table 4.1
Summary information for the populations used in the simulations.

Population	N	N_{ntw}	θ	θ/N	$y_{\max}$	T_b	T_b/θ
afrbuf	391	5	15	0.038	99	334	22.3
nfalls	529	5	20	0.038	59	368	18.4
thmp	400	3	21	0.053	39	190	9.4
devt	1,170	10	85	0.073	34	868	10.2
paris	529	6	43	0.081	63	1,112	25.9
afrhart	391	9	33	0.084	20	171	5.18

Notes: Recall that N is the population size, θ is the number of units greater than zero and T_b is their sum. We let N_{ntw} be the number of networks in the population and $y_{\max}$ be the maximum y value in the population.

Figure 4.1 A three dimensional plot of the population afrbuf.

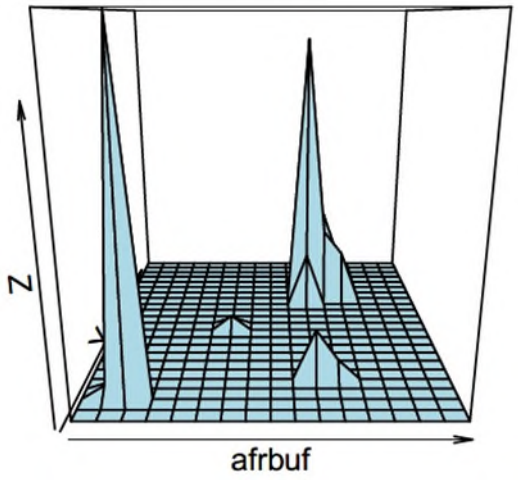


Figure 4.2 A three dimensional plot of the population nfalls.

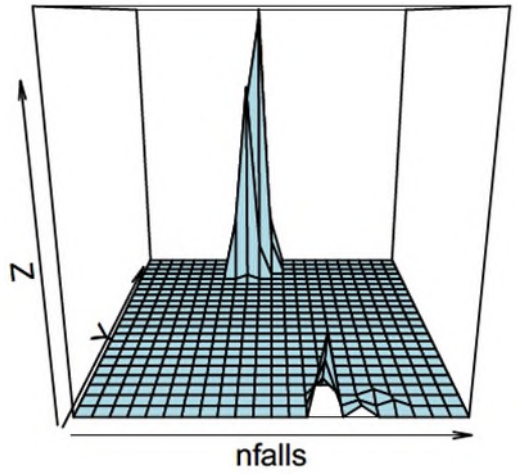


Figure 4.3 A three dimensional plot of the population thmp.

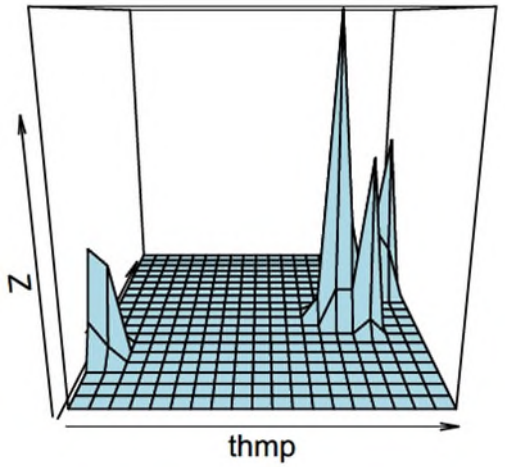


Figure 4.4 A three dimensional plot of the population devt.

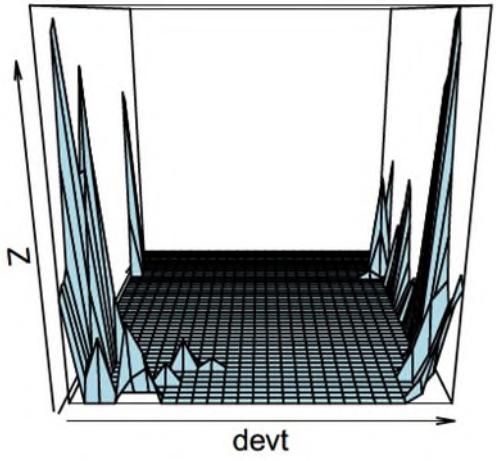


Figure 4.5 A three dimensional plot of the population paris.

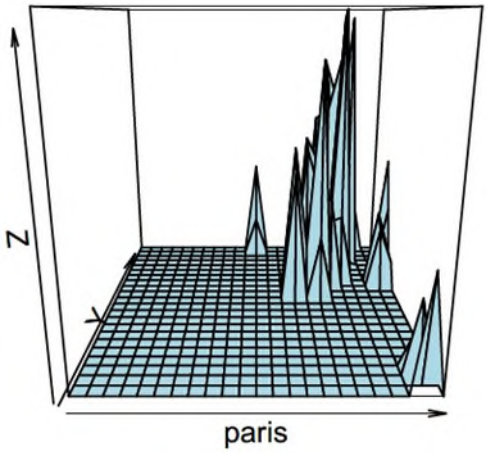
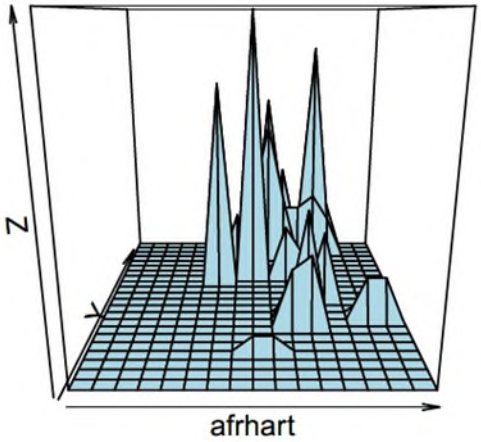


Figure 4.6 A three dimensional plot of the population afrhart.



5. The simulations

5.1 Doing the simulations

Near the end of Section 2.1 we explained why we would only consider ACS samples which had at least three counts greater than zero. For this reason, in our simulations, we will only consider samples where $n_b > 2$. In other words, our frequency of coverage is a conditional one; it is conditional on seeing at least three counts greater than zero. So the results given in the Tables 5.1 through 5.8 and 5.10 through 5.13 for the adaptive cluster sampling estimator and our quasi Bayes estimator, ACS and BAY respectively, are conditional. For each method the tables give the average value of an estimator, Est, its average relative bias, Rbias, its average absolute error, ABerr, the average lower bound of its interval estimate, Lowbd, the average length of its interval, Len, and the frequency which its upper bound was larger than T_b.

In our adaptive cluster sampling simulations we used two different initial sample sizes; ten percent and twenty percent of the population size. Here we will present just the ten percent results. How the two methods compare is essentially the same for the two different initial sample sizes but of course they both do better for the larger initial sample size.

Finally, someone might be concerned about what would happen if the true size of D_b was slightly bigger than $N/10$ and hence violating our definition of a rare species. In practice it is unlikely that the observed n_b would be bigger than this bound. But to allow for this possibility, when doing the simulations, we defined our prior on the integers between $0.01\,N$ and $0.15\,N$. The results hardly differ from what happens if the set of possible integers are the non negative integers less than or equal to $0.1\,N$. This happens because as we move away from n_b the posterior decreases very quickly and points far out in the tail contribute little probability. So even though we developed our "posterior" with the smaller upper bound in mind it works almost as well with the much larger and imprecise upper bound. Hence to use our method one does not need a good guess for the upper bound of θ.

5.2 Results for the six populations

There are several things to note about the simulation results given in Tables 5.1 through 5.6. For all the populations the BAY estimator has smaller average absolute error than the ACS estimator, sometimes dramatically so. Overall the BAY upper bounds perform better than the ACS upper bounds. The population afrbuf is the only case where the ACS upper bound is superior. The BAY upper bound is too large. For population nfalls the two methods behave about the same. For populations thmp, devt and paris the BAY upper bound is clearly the best. The only population where its frequency of coverage falls below 0.90 is population afrhart. In this case its frequency of coverage is only 0.892 but its upper bound was much smaller than the ACS bound which had a frequency of coverage of 0.861. Some of the information in the tables is presented graphically in Figure 5.1.

Table 5.1
Population afrbuf.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	489	0.464	187.2	195	974	0.98
BAY	334	-0.000	69.3	195	1,644	0.98

Notes: For 1,000 simple random samples of size 39, there were 660 samples with at least three counts greater than zero. The average total number of positive values observed was 6.96.

Table 5.2
Population nfalls.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	448	0.217	198.9	319	706	0.924
BAY	447	0.214	126.5	319	724	0.998

Notes: For 1,000 simple random samples of size 53, there were 819 samples with at least three counts greater than zero. The average total number of positive values observed was 13.57.

Table 5.3
Population thmp.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	211	0.112	87.9	118	341	0.921
BAY	157	-0.173	45.7	118	274	1.000

Notes: For 1,000 simple random samples of size 40, there were 887 samples with at least three counts greater than zero. The average total number of positive values observed was 12.98.

Table 5.4
Population devt.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	874	0.007	224	723	708	0.942
BAY	857	-0.012	89	723	548	0.972

Notes: For 1,000 simple random samples of size 117, there were 1,000 samples with at least three counts greater than zero. The average total number of positive values observed was 65.98.

Table 5.5
Population paris.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	1,119	0.006	389.5	970	1,119	0.909
BAY	1,114	0.002	69.2	969	603	0.984

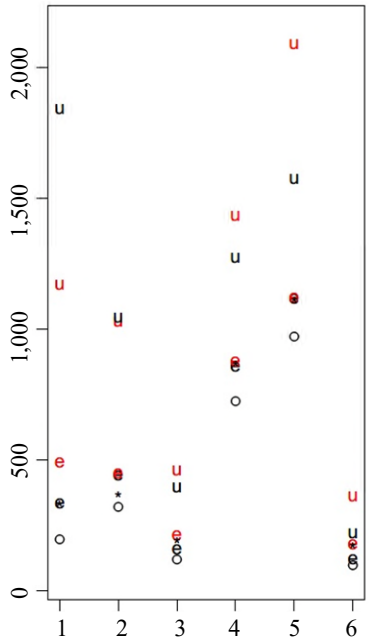
Notes: For 1,000 simple random samples of size 53, there were 983 samples with at least three counts greater than zero. The average total number of positive values observed was 34.19.

Table 5.6
Population afrhart.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	175	0.021	78.4	96	261	0.861
BAY	119	-0.305	52.4	96	120	0.892

Notes: For 951 simple random samples of size 39, there were 968 samples with at least three counts greater than zero. The average total number of positive values observed was 18.74.

Figure 5.1 The above represents, graphically, some of the information in Tables 5.1 through 5.6 for the six populations afrbuf(1), nfalls(2), thmp(3), devt(4), paris(5) and aftrhart(6).



Notes: The number in parentheses is their location on the horizontal axis. The true population total is denoted by $\star$. The ACS estimator and its upper bound are denoted by e and u. For the BAY estimator these quantities are in the color black and along with the common lower bound, o.

For the six populations in Table 4.1 the one for which the BAY estimator had the most extreme negative bias, -0.305, was population afrhart. Looking more closely at afrhart we see it has 33 units greater than zero in 9 networks with an average value of 5.18. The three largest values in the population are 17, 15 and 20. These latter two each appear in a network of size one while the first appears in the largest network which has 16 units. Most of the rest of the values in the population are 5 or less. The mean of these remaining 30 units is 3.97. With the choice of $\alpha = 1$ and $\beta = 90$ the average value of our estimator of $\theta = 33$ was 23.5. This helps to explain the large negative bias of our estimator for this population. One way to improve its performance would be to increase the estimate of θ by a different choice of the prior distribution. More generally, having good prior information about the size of θ can lead to improved results.

It is not surprising that the ACS estimator is biased upwards. This is because we are ignoring all ACS samples with less than three counts greater than zero. If they were included then the ACS estimator would always be unbiased. On the other hand our BAY estimator can be both biased upwards or downwards. It can be biased upwards when we over estimate the size of D_b. It can be biased downwards when we underestimate the size of D_b or when there is a very small network whose counts are much larger than the counts in the remaining networks. But this latter case is also a problem for the ACS estimator. We will discuss this in more detail in Section 5.3.

We see from Table 4.1 that population paris has 43 units greater than zero in 6 networks with an average value of 25.9. Checking the networks we find that just one network contains 31 units whose average value

is 30.0. We see from Table 5.5 that on the average we observed 39.4 units. This means the ACS sample almost always contained the units in this network. This helps to explain our excellent results for this population.

A similar thing happens with population devt. In this population there are 10 networks containing 85 units with counts greater than zero. The total of these units is 868. The two biggest networks contain 31 and 23 units respectively and their respective means are 11.3 and 12.8. The next biggest network contains 13 units and all the rest have 5 or less units. From Table 5.4 we see that the average number of units in the final ACS samples was 65. This means that most of the ACS samples contained the two largest networks. This is not surprising, but the fact that these two largest networks are good representative samples of D_b explains why our estimators perform so well for this population.

From Table 4.1 we see that for our six populations the ratio θ/N, the proportion of units with positive counts in the population, ranges from 0.038 to 0.084. Recall that given an ACS sample, the adjusted factor λ, defined in equation (3.7), was introduced to increase our estimate of variance when the number of counts greater than zero in the ACS sample was quite small. For populations afrbuf and nfalls, the populations with the smallest values of the ratio θ/N, the average value of λ was 1.42 and 1.20 respectively while for the two largest populations, devt and paris, these averages were 1.01 and 1.02 respectively. This shows that our choice of the λ is working as expected.

5.3 Six more populations

As we have mentioned, we settled on our estimator by simulation studies on these six populations. They were chosen because they represent a variety of situations where ACS sampling would be used. A reader might be worried that our estimator was too dependent on the populations we used in our study even though they are quite different. In an attempt to show that this is not the case we now present six new populations that were not used in the development of our estimator. We constructed the first two populations on a grid of 400 squares with just two networks. The first network had three members with values 50, 60 and 70. The second network had twelve members with the values 14, 15 and 16 each appearing four times. Let ref1 denote this population. Let ref2 denote the population where 14, 15 and 16 appear in the smaller network and 50, 60 and 70 each appear four times in the larger network. Tables 5.7 and 5.8 give the results for 1,000 samples of size 40 for the two populations. The same set of samples was generated for both populations.

We see that in the 789 samples with at least two counts greater than 0 the average number of observed counts was 11.67. This means that in the majority of these samples only the larger network was observed. Note however that our results for population ref1 are quite good, even though such samples only contain the small counts. Our average estimate of θ was 15.93, a slight overestimate. Because we have observed the larger network this helps to compensate for the fact that only the units with the smaller count size are in the sample. While for population ref2, having all the larger counts in the larger network, means that our estimator is biased upwards but less so than the ACS estimate.

Table 5.7
Population ref1.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	464	0.290	269.6	223	892	1
BAY	330	-0.084	129.8	223	704	1

Notes: For 1,000 simple random samples of size 40, there were 789 samples with at least three counts greater than zero. The average total number of positive values observed was 11.67.

Table 5.8
Population ref2.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	957	0.251	444	653	1,574	0.885
BAY	865	0.131	244	653	962	1.000

Notes: For 1,000 simple random samples of size 40, there were 789 samples with at least three counts greater than zero. The average total number of positive values observed was 11.67.

Reasoning similarly to the discussion of population ref2 just given we can explain the upward bias of our estimator for population nfalls in Table 5.2. This population has seven networks containing a total of $\theta = 20$ units. Here one network contains 13 of the units and it contains all the largest units as well. The average number of observed counts greater than zero was 13.57 so that the network containing 13 units was almost always in the sample. The average value of our estimator for θ was 19.5, which is quite good. But it cannot compensate for the fact that the ACS samples almost always include the larger counts.

Remember that $\hat{T}_{ac}$ would be unbiased if we used all possible samples and not just those with at least three counts greater than zero. When all the observed counts are zero then it will estimate zero. This event will happen with positive probability and will be an underestimate except when the population total is in fact zero. To compensate for this underestimate, $\hat{T}_{ac}$ overestimates when the sample contains "lots" of neighborhoods with counts greater than zero. This also helps to explain why BAY does better than ACS.

To see this we will look more closely at the two populations ref1 and ref2. First consider the three cases where only the second network was either observed once or twice or three times. For population ref1, whose total is 360, the corresponding estimates BAY(ACS) ranged from 226(150) to 257(450). Next consider the three cases where the first network was observed just once and the second network was observed either once or twice or three times. For these three cases the estimates ranged 446(750) to 500(1,050). For population ref2 whose population total is 765 the corresponding values of the estimates for the first three cases ranged from 903(600) to 1,029(1,800) and for the second set of cases the corresponding values of the estimates ranged from were 949(750) to 1,063(1,950). Note that the range of values for the ACS estimator is much larger than the range for the BAY estimator. This also explains why in all our simulations $\hat{T}_{ac}$ is biased upwards. It is because we are ignoring all the samples with less than three counts greater than zero.

We will now describe the remaining four new populations, mich, rome, afrdeer and sunspt. Summary information about the four populations are given in Table 5.9. Plots of the four populations are given in Figures 5.2 through 5.5.

Table 5.9
Summary information for the new populations rome, mich, afrdeer and sunspt.

Population	N	N_{ntw}	θ	θ/N	$y_{\max}$	T_b	T_b/θ
rome	1,600	16	48	0.03	495	7,927	165.1
mich	400	10	20	0.05	3,577	24,740	1,237
afrdeer	391	13	76	0.19	140	1,309	16.6
sunspt	500	14	27	0.05	98	434	16.1

Notes: Recall that N is the population size, θ is the number of units greater than zero and T_b is their sum. We let N_{ntw} be the number of networks in the population and $y_{\max}$ be the maximum y value in the population.

Figure 5.2 A three dimensional plot of the population rome.

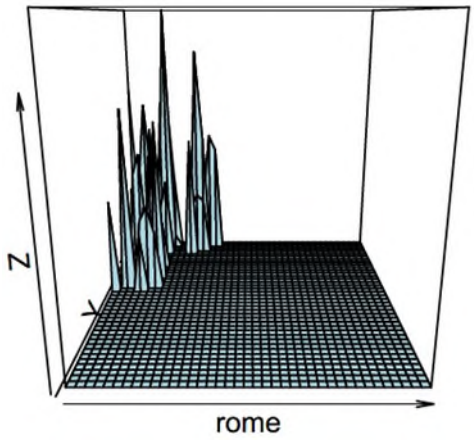


Figure 5.3 A three dimensional plot of the population mich.

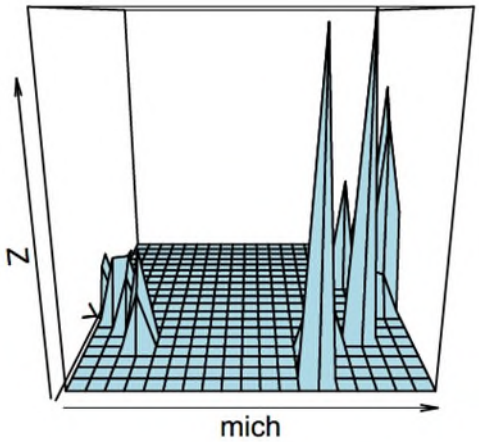


Figure 5.4 A three dimensional plot of the population afrdeer.

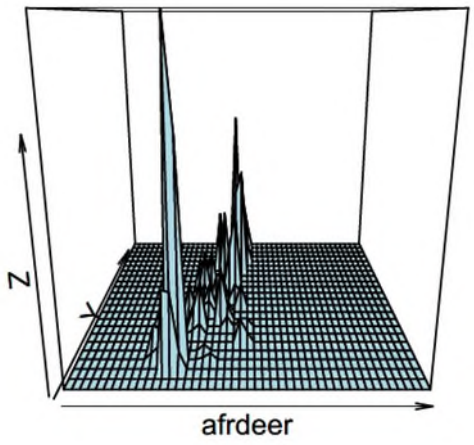
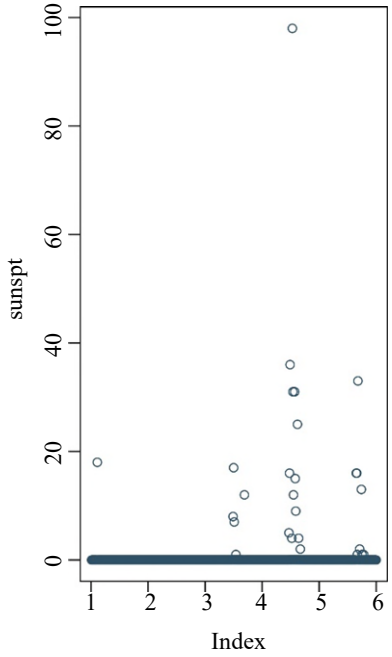


Figure 5.5 A plot of the population sunspt.



The first population is based on a 40 by 40 grid in Rome, Italy. The second is based on a 20 by 20 grid in the state of Michigan in the United States. We denote these two populations by rome and mich respectively

For population rome we see that there are 16 networks which contain 48 counts greater than zero. Four of the networks, that range in size from 6 to 11, contain 34 of the 48 counts greater than zero. The means of these networks range from 139.3 to 219. The mean of these four means is 176.8 which is quite close to 165.1, the mean of all the counts greater than zero. The remaining 12 networks contain either one or two values and most of them quite small. The two largest are 335 and 301. This explains the excellent behavior

of our estimator BAY in Table 5.10. With high probability the ACS sample will contain one of the four larger networks.

Table 5.10
Population rome.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	8,206	0.035	2,793	4,044	11,107	0.923
BAY	7,459	-0.059	1,731	4,044	12,673	1.000

Notes: For 1,000 simple random samples of size 160, there were 986 samples with at least three counts greater than zero. The average total number of positive values observed was 22.81.

For population mich we see that there are 10 networks containing 20 counts greater than zero. Of these counts, 11 are in 6 networks where the maximum count is 695 and the six smallest range from 105 to 190. The largest network contains 5 counts and its mean is 969.2. The largest count is 3,077 and is in a network of size one. So although there are just a few units with counts greater than zero the actual counts can be quite large. We include this example to see what could happen when this is the case. The large negative bias of the BAY estimator in Table 5.11 occurs because the largest count is a network of size one. Even with this bias it still has significantly smaller average absolute error than the ACS estimator. Both of their upper bounds perform poorly however. It is difficult to get a sensible upper bound when there is one network with one or two values which are significantly larger than the rest of the counts in the population. Both will underestimate if the big counts are not included in the sample and overestimate when they are. Unless there is additional prior information we believe that it is very difficult to find sensible estimates without approximate exchangeability across the counts in the networks when we only observe a very small number of counts greater than zero.

Table 5.11
Population mich.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	32,657	0.320	14,333	8,065	63,114	0.964
BAY	16,805	-0.321	8,602	8,065	64,281	1.000

Notes: For 1,000 simple random samples of size 40, there were 686 samples with at least three counts greater than zero. The average total number of positive values observed was 6.49.

For the next new population we consider a third population given in Gattone et al. (2016). We let afrdeer denote this population and summary information is given in Table 5.9. The majority of the units belong to the three largest networks of sizes 13, 17 and 21. Their respective means are 29.3, 8.8 and 25.8. The next largest remaining network is of size 5 and the remaining values are mostly small. The largest value in these other networks is 34. Note that for this population the true value of θ is 76 so that the ratio $\theta/N = 76/391 = 0.19$ is greater than 0.15 the upper bound used in defining our prior.

So the natural question is how does our estimator work in this case? The results of this simulation are given in Table 5.12. We see that as in the other examples we do reasonably well and better than the ACS estimator. This happens because the ACS samples tend to include the larger networks and the smaller networks tend to have smaller counts.

Table 5.12
Population afrdeer.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	1,307	-0.001	406	962	1,304	0.917
BAY	1,213	-0.073	214	962	1,269	0.970

Notes: For 1,000 simple random samples of size 39, there were 1,000 samples with at least three counts greater than zero. The average total number of positive values observed was 49.76.

At the end of Section 3.1 we noted that our posterior distribution for θ does not depend on the fact that our basic units are squares and that the population is a rectangle. But it is important that the basic units be approximately the same size and that neighbors can be defined in a sensible manner. To demonstrate this point we now consider a population defined on a set of successive intervals of equal length. The neighbors of an interval are just the two adjacent intervals to the left and to the right except the end points which have only one neighbor. To get the counts we use the population of sunspots that is available in R (R Core Team, 2023). These are the monthly mean relative sunspot numbers from 1,749 to 1,983 and they have the clumping behavior which should make ACS sampling appropriate. We took the first 500 values of sunspots and subtracted 141 from each one. The resulting negative numbers were set to zero. Summary information for this new population, sunspt, is given in Table 5.9. In sunspt there are 14 networks: 8 of size 1, 1 of size 2, 3 of size 3 and 2 of size 4. so theta = 27. The range of the 27 counts greater than zero range from 1 to 98. The next largest count is 36. The largest count appears in a network of size four with a mean of 36.25 which is quite a bit larger than $T_b/\theta =$ 16.1. We see in Table 5.13 that the behavior of the two estimators is very similar to what we have seen in our two dimensional examples. This example demonstrates that our approach could be useful when studying longitudinal data which satisfies our assumptions of rareness and clumping.

Table 5.13
Population sunspt.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
ACS	497	0.144	217.8	140	930	0.952
BAY	304	-0.299	159.3	140	1,369	0.999

Notes: For 1,000 simple random samples of size 50, there were 874 samples with at least three counts greater than zero. The average total number of positive values observed was 7.29.

Three dimensional plots of the populations mich, rome and afrdeer are given in Figures 5.2, 5.3 and 5.4. A plot of sunspt is given in Figure 5.5.

6. R code for calculating our estimator

Here we argued for a particular quasi Bayes estimator when little is known about the population of interest except that the species of interest is rare. But in fact we have defined a whole family of possible estimators by considering different choices of the parameters in the beta density and different choices for the definition of the parameter space Θ. Any of these possible estimators is very simple to compute in R (R Core Team, 2023). What follows below is an R function which calculates one of these estimates for the population total, along with an upper bound and the lower bound. One just needs to specify the parameter space Θ, the beta distribution which one uses to define their prior distribution on Θ and some of the information from a full ACS sample. One does not need to know the edges of the observed networks.

More formally here is what you need to specify;

- n, the size of the initial random sample,
- nbiginrs, the number of counts greater than zero in the initial random sample,
- bigsmp, all the counts greater than zero in the full ACS sample,
- bds, used to define the parameter space Θ,
- alp and bet, the parameters for the beta distribution used to define our prior distribution,
- N, the population size.

Next we give the code for the function

```
qbay<-function(n,nbiginrs,bigsmp,bds,alp,bet,N)
  {
     klw<-floor(bds[1]*N)
     kup<-ceiling(bds[2]*N)
     theta<-klw:kup
     nbigsmp<-length(bigsmp)
     dtheta<-theta[theta>=nbigsmp]
     ntheta<-length(dtheta)
     llike<-lchoose(dtheta,nbiginrs) + lchoose(N-dtheta,n-nbiginrs)
     lprior<-log(dbeta(dtheta/N,alp,bet))
     dum<-lprior + llike
     post<-rep(0,ntheta)
     for(i in 1:ntheta){
       post[i]<-1/sum(exp(dum-dum[i]))
      }
     pstmnth<-sum(dtheta*post)
     est<-pstmnth*mean(bigsmp)#the point estimate in equation (5.6)
     pst2ndmnth<-sum(dtheta^2*post)
     pstvrth<-pst2ndmnth - pstmnth^2
```

```
    nbig<-length(bigsmp)
    mnbg<-mean(bigsmp)
    lwbd<-nbig*mnbg
    vr1<-(mnbg)^2*pstvrth
    nb<-length(bigsmp)
    d1<-var(bigsmp)
    d2<-sum(post*(dtheta-nb)^2)
    vr2<-d1*d2
    vr<-vr1+vr2              #the variance in equation (5.7)
    lam<-10^(2.5/nbig)       #the adjustment factor in equation (5.8)
    upbd<-est+sqrt(vr)*lam*1.96
    ans<-c(est,lwbd,upbd)
    return(ans)
}
```

This function allows one to explore a better choice of a quasi Bayes estimator when one has some additional prior information about the population. Note that one can also easily change our adjustment factor, λ, that appears in equation (3.7). In the next section we will briefly discuss some possible extensions.

7. Possible extensions

Thompson (1990) introduced ACS sampling for biological applications where one was interested in a rare species that tended to appear in clumps. As we noted in Section 2.2 we could not find in the literature any definition of rare. But as our simulations demonstrated our approach can work well for a fairly wide range of rareness. Our prior for the number of counts greater than zero does not depend on how large they are. From one point of view this is sensible because the notion of rareness can depend on the species under study. In some cases one might have good prior information about the number of counts greater than zero. In this case one could select different values of α and β in the prior distribution that better reflects this information.

As far as we know there is no formal definition of clumping in the literature. An extreme form of clumping would be if only one unit contains the total T_b elements of the species of interest. But this is clearly not in the spirit of ACS sampling. It seems to us that ACS sampling presumes that such a large concentrated count is not possible. Instead it assumes that such a large count would tend to spread out into neighboring squares forming a network. As we saw in our description of population thmp at the beginning of Section 4 its three networks are of this form. When the size of the networks tend to be larger and counts within a network are representative of all the counts in the population we will have good results.

When this is not true and there is a large count, compared to the rest, in a small network then both the ACS and BAY upper bounds perform poorly. We saw this for populations afrbuf and mich. Our choice of

the adjustment factor λ in equation (3.7) is based on an implicit clumping assumption about the range of possible values for the counts and on how likely it is to have a small network with extremely large counts. In some cases one could have information about what is a "big" count for the population and the proportion on such "big" counts in the population. Depending on the size of the counts in the ACS sample one could let the adjustment factor, λ, depend on the value of the counts in the sample and the prior information. One could also weaken our assumption of approximate exchangeability and use prior information to replace the mean of the observed counts by a different estimator. Prior information could improve our estimator but will result in poorer estimators when it is incorrect. How to make use of additional prior knowledge in a more formal manner needs further study.

Rapley and Welsh (2008) developed interesting models for the study of the type of populations where ACS sampling is used. In particular they modeled how networks and edges could be formed. Our approach is much simpler because we are ignoring the networks and only consider the number of nonempty units in the population. Our suggested prior distribution for the size of the D_b is very similar to a prior that is used in a slightly different context in their model. Here we have been interested in showing that one can improve on the standard ACS approach without making any model assumptions. It would be interesting to compare our approach to Bayesian approaches to the problem. How much better a Bayesian approach can do than what we have presented here when one has in hand good prior information is a question that needs further study.

Rapley and Welsh also noted that the sampling could be done sequentially. That is, a unit is selected at random from the population. If its count is greater than zero then one observes all the units in its network and its edges. At each step we only selected one unit at random from the remaining unobserved units. We continue in this way until our stopping rule tells us the sampling is finished. Our approach can be extended to such sequential sampling plans. Assuming that the order in which the sample was taken is known then the form of the likelihood function would change but one could use the same prior distribution for θ. From a theoretical point of view sequential sampling makes a lot of sense but it is not clear if it would be practical in many of the problems where ACS sampling would be used.

8. Concluding remarks

Adaptive cluster sampling was introduced to improve sampling efficiency for populations with a special type of structure. It is appropriate when the statistician knows that the population has just a few cells with a y value greater than zero and that they tend to appear in clusters. It is an interesting approach which has been widely adopted when studying biological populations in the field. We showed however that in its focus on finding an estimator which is design unbiased it has ignored some of the available information. Here we introduced a quasi Bayesian approach to the problem that makes use of this information. We showed that our point estimator and our 95% upper confidence bound for the population total gave much better results than the standard approach.

Acknowledgements

We wish to thank Gabriel Mersy for his help with using the *R* package rgbif.

References

Chamberlain, S., Ram, K., Mcglinn, D. and Barve, V. (2019). *rgbif: Interface to the Global Biodiversity Information Facility API*. https://CRAN.R-project.org/package=rgbif.

Cochran, W. (1977). *Sampling Techniques* (third ed.). New York: John Wiley & Sons, Inc.

Dryver, A., and Chao, C. (2007). Ratio estimators in adaptive cluster sampling. *Environmetrics*, 18, 607-620.

Gattone, S., Mohamed, E. and Di Battista, T. (2016). Adaptive cluster sampling with clusters selected without replacement and stopping rule. *Environmental and Ecological Statistics*, 23, 453-468.

Ghosh, J.K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu*. New York: Springer-Verlag.

Goncalves, K.C.M., and Moura, F.A.S. (2016). A mixture model for rare and clustered populations under adaptive cluster sampling. *Bayesian Analysis*, 11, 519-544.

Latpate, R., Kshirsagar, J., Gupta, V.K. and Chandra, G. (2021). Adaptive cluster sampling. *Advanced Samping Methods*, 125-156. Springer Singapore.

Meeden, G. (1999). Interval estimators for the population mean for skewed distributions with a small sample size. *Journal of Applied Statistics*, 26, 81-96.

Nolau, I., Goncalves, K.C.M. and Pereira, J.B.M. (2022). Model-based inference for rare and clustered populations from adaptive cluster sampling using auxiliary variables. *Journal of Survey Statistics and Methodology*, 10, 439-465.

Pacifici, K., Reich, B., Dorazio, R. and Conroy, M. (2016). Occupancy estimation for rare species, using a spatially-adaptive design. *Methods in Ecology and Evolution*, 7, 285-293.

R Core Team (2023). *R: A Language and Evironment for Statistical Computing*. R Foundation for Statistical Computing. http://R-project.org.

Rapley, V.E., and Welsh, A.H. (2008). Model-based inferences from adaptive cluster sampling. *Bayesian Analysis*, 3, 717-736.

Savitsky, T., and Toth, D. (2014). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10, 1677-1708.

Si, Y., Pliiai, N. and Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10, 605-625.

Thompson, S. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.

Thompson, S. (2012). *Sampling* (third ed.). Hoboken, New Jersey: Wiley.

Turk, P., and Borkowski, J. (2005). A review of adaptive cluster sampling: 1990-2003. *Ecological and Environmental Statistics*, 12, 55-94.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 235-259
Statistics Canada, Catalogue No. 12-001-X

Inference from sampling with response probabilities estimated via calibration

Caren Hasler[1]

Abstract

A solution to control for nonresponse bias consists of multiplying the design weights of respondents by the inverse of estimated response probabilities to compensate for the nonrespondents. Maximum likelihood and calibration are two approaches that can be applied to obtain estimated response probabilities. We consider a common framework in which these approaches can be compared. We develop an asymptotic study of the behavior of the resulting estimator when calibration is applied. A logistic regression model for the response probabilities is postulated. Missing at random and unclustered data are supposed. Three main contributions of this work are: 1) we show that the estimators with the response probabilities estimated via calibration are asymptotically equivalent to unbiased estimators and that a gain in efficiency is obtained when estimating the response probabilities via calibration as compared to the estimator with the true response probabilities, 2) we show that the estimators with the response probabilities estimated via calibration are doubly robust to model misspecification and explain why double robustness is not guaranteed when maximum likelihood is applied, and 3) we highlight problems related to response probabilities estimation, namely existence of a solution to the estimating equations, problems of convergence, and extreme weights. We present the results of a simulation study in order to illustrate these elements.

Key Words: Maximum likelihood estimation; Nonresponse; Two-phase estimation; Weighting adjustment.

1. Introduction

Under complete response the Horvitz-Thompson (HT) estimator is unbiased (Horvitz and Thompson, 1952). With nonresponse, however, this estimator is unavailable. Nonresponse can be seen as a second phase of the survey, where the mechanism that yields the nonresponse, called the *response mechanism*, is unknown (Oh and Scheuren, 1983; Särndal and Swensson, 1987). If the response probabilities were known, a two-phase estimator with response probabilities as inclusion probabilities of the second phase would be unbiased. Unfortunately, the response probabilities are unknown in practice. A solution to control for nonresponse bias is to postulate a model for the response probabilities, estimate these probabilities based on the postulated model, and use the estimated response probabilities in a two-phase estimator. The resulting estimator is called *two-phase Nonresponse Weighting Adjusted (NWA) estimator* or *empirical double expansion estimator*. Särndal and Lundström (2005) and Haziza and Beaumont (2017) provide overviews of some NWA estimators and weighting systems adjusted for nonresponse.

Two general approaches to NWA estimators are Maximum Likelihood Estimation (MLE) and calibration (Deville and Särndal, 1992). In the first approach, a model such as the logistic regression model is postulated (Cassel, Särndal and Wretman, 1983; Ekholm and Laaksonen, 1991). The parameters of the model are estimated via MLE and fitted response probabilities are obtained based on the estimated parameters. In the second approach, calibration weights are found so that the resulting NWA estimator of some auxiliary variables is equal to its full sample HT estimator (calibration at the full sample level) or its population total

1. Caren Hasler, Institut de Statistique, Université de Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel. E-mail: caren.hasler@unine.ch.

(calibration at the population level). The calibration weights can be viewed as the design weights times the inverse of the estimated response probabilities. To the best of our knowledge, the first author to suggest the use of what would later be called calibration weighting to estimate the response probabilities is Folsom (1991), shortly followed by Deville and Dupont (1993) and Dupont (1993). Lundström and Särndal (1999) further study point and variance estimators for both levels of calibration, sample and population.

The first approach is studied in depth in Kim and Kim (2007), which presents asymptotic properties of the NWA estimator under a general response model. Two main results of their paper are: 1) the NWA estimator with response probabilities estimated via MLE is asymptotically equivalent to an unbiased estimator and 2) a gain in efficiency is obtained when estimating the response probabilities via MLE as compared to the estimator with the true response probabilities. The second result is also shown by Beaumont (2005) under the logistic response model.

The second approach can be divided into two levels: calibration at the sample level and calibration at the population level. The NWA estimator obtained when the response probabilities are estimated via calibration at the sample level is a particular case of the propensity-score-adjustment estimator of Kim and Riddles (2012). These authors develop the asymptotic properties of this estimator in a theoretical framework different from that considered in Kim and Kim (2007). This estimator is also considered in Iannacchione, Milne and Folsom (1991) which focuses on practical aspects of NWA estimation with calibration at the sample level. It does not provide any theory.

The main goal of both approaches is to reduce the nonresponse bias and, if possible, the variance of population estimators. The second approach, calibration, also ensures consistency between estimated and known population totals. This is not the case of the first approach, MLE. However, the second approach, i.e., direct estimation of the response probabilities via calibration, called *one-step approach*, is sometimes criticized as it tends to yield biased estimates when the response model is misspecified (Haziza and Lesage, 2016). An alternative consists of first estimating the response probability via MLE and then applying calibration to ensure consistency between estimated and known totals. This alternative is called *two-step approach*. The reader may refer to Haziza and Lesage (2016) and Haziza and Beaumont (2017), page 222, for a discussion of the one- and two-step approaches.

In this paper, we study MLE and the one-step approach to calibration for nonresponse weighting adjustment. We build on Kim and Kim (2007) and develop asymptotic properties of the NWA estimator under the second approach, calibration at both the sample and the population levels. For the first time, a common theoretical framework is considered for both approaches to NWA estimation, namely MLE and calibration. This allows us to compare the asymptotic behavior of the resulting NWA estimators in terms of bias and variance under common assumptions. We postulate a logistic regression model for the response probabilities. We suppose that the data are missing at random (see Rubin, 1976, for a detailed definition) and unclustered. Two main theoretical results are 1) the NWA estimators with the response probabilities estimated via calibration are asymptotically equivalent to unbiased estimators and 2) a gain in efficiency is obtained when estimating the response probabilities via calibration as compared to the estimator with the true response probabilities. These results are valid for both levels of calibration.

Another main contribution of this work is the study of the double-robustness of the NWA estimators. Indeed, both approaches assume, implicitly or explicitly, two models: 1) a model that links the variable of interest and the auxiliary variables, called *superpopulation model*, and 2) a model for the response probabilities, called *response model*. We show that the NWA estimators with response probabilities estimated via calibration are doubly robust. That is, these estimators are consistent even if one of the two aforementioned models is misspecified. We also explain why double robustness of the NWA estimator with response probabilities estimated via MLE is not guaranteed. To the best of our knowledge, only Kott and Liao (2012) discusses double robustness of NWA estimation via calibration in probability sample surveys. In their article, the emphasis is put on an exponential form for the response probabilities. Finally, one last main contribution of this work is a discussion about problems of convergence and extreme weights. Indeed, it may happen that the estimating equations used to obtain estimated response probabilities do not admit a solution or that a solution to the estimating equations exists but the resulting weights, that is, the inverse of the estimated response probabilities, are very large. We illustrate this phenomenon. Results of a simulation study confirm the theoretical results and practical considerations presented. A longer version of this paper with additional technical and practical elements is available on ArXiv at

https://doi.org/10.48550/arXiv.2202.03897.

The current paper is organized as follows: Section 2 contains pieces of notation and important concepts. In Section 3, we present both approaches to response probabilities estimation. We describe some asymptotic properties of the NWA estimators of interest in Section 4 with some technical elements left in the Appendix of the longer version of this article (Hasler, 2023). We discuss double robustness to model misspecification in Section 5. In Section 6, we present the variance and variance estimation of the NWA estimators of interest. Section 7 contains the results of a simulation study. A discussion closes the paper in Section 8.

2. Framework

Consider a finite population $U = \{1, 2, \ldots, i, \ldots, N\}$ of size N. A vector of v auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iv})$ is attached to a generic unit i. We suppose that the first auxiliary variable is constant and equal to 1. The parameter of interest is the population total

$$Y = \sum_{i \in U} y_i,$$

for some unknown variable of interest y. A sample s of size n is selected from U according to a non-informative probabilistic sampling design $p(\cdot)$ with the aim of observing y_i for $i \in s$. A random sample S is a random variable such that $\Pr(S = s) = p(s)$. The random sample is also defined via an indicator variable $(a_i \mid i \in U)^\top$ where a_i is 1 if unit i is in the sample and 0 otherwise. Consider

$$\pi_i = \Pr(i \in S) = \sum_{s \subset U;\, s \ni i} p(s),$$

the first-order inclusion probability of unit i and suppose that $\pi_i > 0$ for all $i \in U$. Let $\mathrm{E}_p(\cdot)$ and $\mathrm{V}_p(\cdot)$ denote the expectation and variance computed with respect to the sampling design $p(\cdot)$. Under complete response, the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952)

$$\hat{Y}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} \tag{2.1}$$

is design-unbiased for Y, i.e., $\mathrm{E}_p(\hat{Y}) = Y$.

Under nonresponse, each sampled unit $i \in S$ is classified as either *respondent* or *nonrespondent* depending on whether y_i is observed or missing. Consider the response indicator vector $\left(r_i \mid i \in S\right)^\top$ where r_i takes value 1 if y_i is observed and 0 if it is missing and $p_i = \Pr(r_i = 1 \mid i \in S)$ the response probability of a sampled unit i. The set of respondents is written $S_r = \left\{i \in S \mid r_i = 1\right\}$ and its size n_r. In the presence of nonresponse, the HT estimator in (2.1) is unavailable and the total Y could potentially be estimated via the *two-phase* (or *double expansion*) estimator

$$\hat{Y}_p = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}, \tag{2.2}$$

provided that p_i is known and strictly positive for all $i \in S$. This estimator is unbiased since

$$\mathrm{E}_p\left\{E_q\left(\hat{Y}_p \mid S\right)\right\} = Y,$$

where $q(\cdot \mid S)$ is the probability distribution of S_r given a sample S and subscript q indicates that the expectation is computed with respect to probability distribution $q(\cdot \mid S)$. The response probabilities are unknown in practice. To address this issue, a model for the response probabilities, called the *response model*, is postulated. The response probabilities are estimated via this model, which yields estimated response probabilities $\hat{p}_i$, and the *NWA estimator* (or *empirical double expansion estimator*)

$$\hat{Y}_{\hat{p}} = \sum_{i \in S_r} \frac{y_i}{\pi_i \hat{p}_i} \tag{2.3}$$

is used. The response probabilities are estimated via $\hat{p}_i = f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})$ for some model $f(\mathbf{x}_i; \boldsymbol{\lambda})$ and estimator $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$. A commonly used model for the response probabilities is the logistic regression model

$$p_i = f(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\lambda})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\lambda})} = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\lambda})}, \tag{2.4}$$

where $\boldsymbol{\lambda}$ is a parameter vector to be estimated. Two available estimation methods are maximum likelihood and calibration, see Section 3. Note that there are ways to use calibration weighting to adjust for nonresponse other than through an assumed logistic response model. For instance, other methods use a linear or logit function that bounds the probabilities of response between 0 and 1. More details can be found in Deville

and Särndal (1992), Deville, Särndal and Sautory (1993), and Haziza and Beaumont (2017), among others. In the current work, we focus on the logistic regression model in (2.4).

Some required assumptions on the response mechanism are:

(R1): The units respond independently of one another, i.e.,

$$\Pr(i, j \in S_r \mid i, j \in S) = p_i p_j.$$

(R2): The response probabilities are bounded below, i.e., there exists a constant $c > 0$ such that $p_i > c$ for all $i \in U$.

(R3): The response probabilities are $p_i = f(\mathbf{x}_i, \boldsymbol{\lambda}^0)$ as defined in (2.4) for some true unknown parameter vector $\boldsymbol{\lambda}^0$.

Assumption (R3) implies that the data are missing at random (see Rubin, 1976, for a detailed definition). This means that

$$\Pr(i \in S_r \mid i \in S, \mathbf{x}_i, y_i) = \Pr(i \in S_r \mid i \in S, \mathbf{x}_i).$$

That is, the propensity to respond is independent from the variable of interest when the auxiliary variables are taken into account. This assumption may fail in practice when the propensity to respond still depends on the variable of interest when all available auxiliary information has been taken into account. If this is the case, one may use generalized calibration (Deville, 2002; Kott, 2006; Lesage, Haziza and D'Haultfoeuille, 2019; Ranalli, Matei and Neri, 2023) to estimate the response probabilities instead of the approaches presented in Section 3.

3. Estimation

We consider two approaches to obtain NWA estimators: MLE and calibration (Deville and Särndal, 1992). Kim and Kim (2007) study NWA estimators via MLE of the response probabilities under a general response model. For the logistic regression model, the maximum likelihood estimator of $\boldsymbol{\lambda}^0$ is the solution $\hat{\boldsymbol{\lambda}}^{\text{mle}}$ to the estimating equation

$$Q^{\text{mle}}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S} k_i \left\{ r_i - f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}}) \right\} \mathbf{x}_i = 0. \tag{3.1}$$

When $k_i = 1$, the solution is the usual maximum likelihood estimator. When $k_i = 1/\pi_i$, we obtain a survey weighted estimating equation, which is often called *pseudo-maximum likelihood*. The idea is that one first unbiasedly estimates the population likelihood estimating equation and then maximizes it. Other choices of k_i are possible. We focus on the common two aforementioned choices. An efficiency gain of the NWA estimator in (2.3) as compared to the two-phase estimator in (2.2) with true response probabilities is claimed

when $k_i = 1$ (Beaumont, 2005; Kim and Kim, 2007). This choice yields the best estimate of $\boldsymbol{\lambda}^0$ and the best estimate of the response probabilities. The efficiency of the NWA estimator may, however, be improved upon with other choices of k_i, such as $k_i = 1/\pi_i$, for example. There is only very limited available literature on this choice. Kott (2012) discusses this choice and the impact on the efficiency of the NWA estimator for the case of response homogeneity groups. No general theory or guidelines about the choice of k_i have been suggested yet in the literature. This goes beyond the scope of this paper.

Two levels of calibration are possible: calibration at the sample level and calibration at the population level. In the first case, the calibration estimator of $\boldsymbol{\lambda}^0$ is the solution $\hat{\boldsymbol{\lambda}}^{\text{cal},S}$ to the estimating equation

$$Q^{\text{cal},S}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})} - \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i} = 0. \tag{3.2}$$

Estimating equation (3.2) is suggested in Iannacchione et al. (1991). In the second case, the calibration estimator of $\boldsymbol{\lambda}^0$ is the solution $\hat{\boldsymbol{\lambda}}^{\text{cal},U}$ to the estimating equation

$$Q^{\text{cal},U}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})} - \sum_{i \in U} \mathbf{x}_i = 0. \tag{3.3}$$

Both estimating equations (3.2) and (3.3) can be solved using a software for calibration in the complete response case, such as function *calib* of R package *sampling* (Tillé and Matei, 2021).

When calibrating at the sample level, the aim is to find response probabilities so that the NWA estimated total of some auxiliary variables is equal to their HT estimator. When calibrating at the population level, the aim is to find response probabilities so that the NWA estimated total of some auxiliary variables is equal to their true total. Hence, the first approach attempts to correct the nonresponse error. The second approach attempts to correct both the nonresponse and sampling error.

Both approaches, MLE and calibration, are here applied to estimate the response probabilities used in the NWA estimator in (2.3). They differ, however, in spirit and required information in the estimation process. The spirit of MLE is to maximize the likelihood that the postulated response model generated the data at hand. The focus is the estimation of the response probabilities with no explicit parameter of interest in mind. Moreover, MLE does not explicitly assume a superpopulation model, i.e., a model that links the variable of interest and the auxiliary variables. We will see in Section 4, however, that MLE assumes an implicit superpopulation model. The spirit of calibration is to estimate the total of some auxiliary variables as precisely as possible. As a result, the nonresponse bias of the total of the variable of interest is as small as possible when the variable of interest and the auxiliary variables are correlated. Calibration thus focuses on a particular parameter of interest, the total, and explicitly states a superpopulation model, a linear regression model.

Both approaches also differ in the required information in the estimation process. MLE requires to know the values $\mathbf{x}_i$ for all sampled units $i \in S$. Calibration at the sample level via estimation equation (3.2)

requires to know the values $\mathbf{x}_i$ for all respondent units $i \in S_r$ and the HT estimator of $\mathbf{x}_i$ at the sample level. Calibration at the population level via estimating equation (3.3) requires to know the values $\mathbf{x}_i$ for all respondent units $i \in S_r$ and the population total of $\mathbf{x}_i$. For MLE and calibration at the sample level, no information is needed about the $\mathbf{x}_i$ out of the sample.

We compare four NWA estimators: 1) $\hat{Y}_{\hat{p}}^{\text{mle},1}$ obtained with response probabilities estimated via equation (3.1) with $k_i = 1$, 2) $\hat{Y}_{\hat{p}}^{\text{mle},1/\pi}$ obtained with response probabilities estimated via equation (3.1) with $k_i = 1/\pi_i$, 3) $\hat{Y}_{\hat{p}}^{\text{cal},S}$ obtained with response probabilities estimated via equation (3.2), and 4) $\hat{Y}_{\hat{p}}^{\text{cal},U}$ obtained with response probabilities estimated via equation (3.3).

4. Asymptotics I

4.1 Theoretical framework

In this section, we build on the results and assumptions of Kim and Kim (2007) to obtain some asymptotic properties of the NWA estimators obtained via calibration. We use the asymptotic framework of Isaki and Fuller (1982). Consider a sequence U_N of embedded finite populations of size N where N grows to infinity. Consider a sequence of samples s_N selected from U_N with sampling design $p_N(\cdot)$. The first- and second-order inclusion probabilities associated with $p_N(\cdot)$ for some generic units i and j are $\pi_{N,i}$ and $\pi_{N,ij}$, respectively. In what follows, we will omit the subscript N whenever possible to simplify notation. We consider the following common regularity conditions on the sequence of sampling designs to ensure consistent estimation of the HT estimator and its variance estimator.

(D1): As $N \to +\infty$, we have $n/N \to \pi^* \in (0,1)$,

(D2): For all N, $\pi_i > \lambda_1 > 0$ for all $i \in U$,

(D3): For all N, $\pi_{ij} > \lambda_2 > 0$ for all $i, j \in U$,

(D4): $\limsup_{N \to +\infty} n \max_{i,j \in U, i \neq j} \left| \pi_{ij} - \pi_i \pi_j \right| < +\infty.$

where lim sup is the limit superior. It is defined as the limit of the sequence of supremums. In the case of (D4), we can write

$$\limsup_{N \to +\infty} n \max_{i,j \in U, i \neq j} \left| \pi_{ij} - \pi_i \pi_j \right| = \lim_{N \to +\infty} \sup \left\{ u_k \middle| k \geq N \right\},$$

where

$$u_k = n_k \max_{i,j \in U_k, i \neq j} \left| \pi_{k,ij} - \pi_{k,i} \pi_{k,j} \right|,$$

and n_k is the size of s_k. Assumption (D4) states that the dependence between sample inclusion indicators is small enough (Breidt and Opsomer, 2017). Intuitively, if we regard

$$n \max_{i,j \in U, i \neq j} \left| \pi_{ij} - \pi_i \pi_j \right|$$

as a measure of dependence between the sample inclusion indicators, this measure should not increase to infinity. For instance, this assumption is satisfied for simple random sampling without replacement, Bernoulli sampling, and any stratified sampling that is not highly stratified. This assumption is not satisfied for cluster sampling or for highly stratified sampling designs. The next section summarizes the results of Kim and Kim (2007) about the asymptotics of the NWA estimator when Maximum Likelihood is applied to obtain estimated response probabilities. The two sections that follow extend these results for the case in which calibration is used. In this section, the reference probability distribution for the convergence is the one jointly defined by the sampling mechanism and the response mechanism.

4.2 Maximum likelihood

From Theorem 1 of Kim and Kim (2007), we have that under the regularity conditions (D1)-(D4), Assumptions (R2)-(R3) about the response mechanism, and additional regularity conditions stated in the Appendix of the longer version of this article (Hasler, 2023), the NWA estimator $\hat{Y}_{\hat{p}}^{\text{mle}}$ satisfies

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\text{mle}} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\text{mle}} + O_p(n^{-1}),$$

where

$$\begin{aligned}
\hat{Y}_{\hat{p},l}^{\text{mle}} &= \sum_{i\in S}\frac{1}{\pi_i}\left\{k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\text{mle}} + \frac{r_i}{p_i}\left(y_i - k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\text{mle}}\right)\right\},\\
\boldsymbol{\gamma}_n^{\text{mle}} &= \left\{\sum_{i\in S} k_i p_i(1-p_i)\,\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1}\sum_{i\in S}\frac{1-p_i}{\pi_i}\,\mathbf{x}_i y_i.
\end{aligned}$$

Remark 1. *The NWA estimator $\hat{Y}_{\hat{p}}^{\text{mle}}$ behaves asymptotically like the linearized estimator $\hat{Y}_{\hat{p},l}^{\text{mle}}$, which is unbiased for the population total Y.*

Remark 2. *If there exists a vector $\boldsymbol{\beta}$ such that $y_i = k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\beta}$ for all $i \in S$ then*

$$\hat{Y}_{\hat{p},l}^{\text{mle}} = \sum_{i\in S}\frac{y_i}{\pi_i}.$$

This means that $\hat{Y}_{\hat{p}}^{\text{mle}}$ is asymptotically equivalent to the full sample unknown HT estimator in this case. When estimating the response probability via MLE, see equation (3.1), we implicitly assume a super-population model, i.e., y_i is a linear combination of $k_i\pi_i p_i \mathbf{x}_i$.

4.3 Calibration at the sample level

Result 1. *Let the sequence of sampling designs satisfy Assumptions (D1)-(D4), the response mechanism satisfy Assumptions (R2)-(R3), and the sequence of finite populations satisfy regularity conditions stated in the Appendix of the longer version of this article (Hasler, 2023). The NWA estimator $\hat{Y}_{\hat{p}}^{\text{cal},S}$ satisfies*

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\text{cal},S} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\text{cal},S} + O_p(n^{-1}),$$

where

$$\begin{aligned}\hat{Y}_{\hat{p},l}^{\text{cal},S} &= \sum_{i\in S}\frac{1}{\pi_i}\left\{\mathbf{x}_i^\top\boldsymbol{\gamma}_S + \frac{r_i}{p_i}\left(y_i - \mathbf{x}_i^\top\boldsymbol{\gamma}_S\right)\right\},\\ \boldsymbol{\gamma}_S &= \left(\sum_{i\in S}\frac{1-p_i}{\pi_i}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sum_{i\in S}\frac{1-p_i}{\pi_i}\mathbf{x}_i y_i.\end{aligned}$$

The proof is given in the Appendix of the longer version of this article (Hasler, 2023).

Remark 3. *The NWA estimator* $\hat{Y}_{\hat{p}}^{\text{cal},S}$ *behaves asymptotically like the linearized estimator* $\hat{Y}_{\hat{p},l}^{\text{cal},S}$, *which is unbiased for the population total* Y.

Remark 4. *If there exists a vector* $\boldsymbol{\beta}$ *such that* $y_i = \mathbf{x}_i^\top\boldsymbol{\beta}$ *for all* $i \in S$ *then*

$$\hat{Y}_{\hat{p},l}^{\text{cal},S} = \sum_{i\in S}\frac{y_i}{\pi_i}.$$

This means that $\hat{Y}_{\hat{p}}^{\text{cal},S}$ *is asymptotically equivalent to the full sample unknown Horvitz-Thompson estimator in this case. When calibrating at the sample level via equation (3.2), we assume a superpopulation model, i.e.,* y_i *is a linear combination of* $\mathbf{x}_i$.

4.4 Calibration at the population level

Result 2. *Let the sequence of sampling designs satisfy Assumptions (D1)-(D4), the response mechanism satisfy Assumptions (R2)-(R3), and the sequence of finite populations satisfy regularity conditions stated in the Appendix of the longer version of this article (Hasler, 2023). The NWA estimator* $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *satisfies*

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\text{cal},U} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\text{cal},U} + O_p(n^{-1}),$$

where

$$\begin{aligned}\hat{Y}_{\hat{p},l}^{\text{cal},U} &= \sum_{i\in U}\left\{\mathbf{x}_i^\top\boldsymbol{\gamma}_U + \frac{a_i}{\pi_i}\frac{r_i}{p_i}\left(y_i - \mathbf{x}_i^\top\boldsymbol{\gamma}_U\right)\right\},\\ \boldsymbol{\gamma}_U &= \left\{\sum_{i\in U}(1-p_i)\,\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1}\sum_{i\in U}(1-p_i)\,\mathbf{x}_i y_i.\end{aligned}$$

The proof is given in the Appendix of the longer version of this article (Hasler, 2023).

Remark 5. *The NWA estimator* $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *behaves asymptotically like the linearized estimator* $\hat{Y}_{\hat{p},l}^{\text{cal},U}$, *which is unbiased for the population total* Y.

Remark 6. *If there exists a vector* $\boldsymbol{\beta}$ *such that* $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ *for all* $i \in U$ *then*

$$\hat{Y}_{\hat{p},l}^{\text{cal},U} = \sum_{i \in U} y_i.$$

This means that $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *is asymptotically equivalent to the unknown population total in that case. When calibrating at the population level via equation (3.3), we assume a superpopulation model, i.e.,* y_i *is a linear combination of* $\mathbf{x}_i$.

5. Asymptotics II: Double robustness

The results in Section 4 rely on Assumption (R3). That is, these results are valid if the response model is correctly satisfied. In this section, we show that the NWA estimators obtained with calibration may still be consistent when the response model is misspecified provided that a superpopulation model, i.e., a model that links the variable of interest to the auxiliary variables, is correctly specified. We say in this case that the resulting NWA estimators are doubly robust because consistency is maintained even when one of the two models, response model or superpopulation model, is misspecified. This is formalized by the results below. For the first result, two required assumptions about the response mechanism and estimated response probabilities are:

(R4): The data are MAR.

(R5): The estimated response probabilities are bounded below, i.e., there exists a constant $c_1 > 0$ such that $\hat{p}_i > c_1$ for all $i \in S$ and all N.

Result 3. *Consider the superpopulation model* $\xi: y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ *where* $\mathrm{E}_\xi(\varepsilon_i) = 0$, $\mathrm{E}_\xi(\varepsilon_i \varepsilon_j) = \sigma^2 \leq +\infty$ *if* $i = j$ *and 0 otherwise, and subscript* ξ *means that the expectation and variance are computed with respect to model* ξ. *Suppose that assumptions (D1)-(D4), (R2), (R4), (R5) are satisfied. Then*

$$\frac{\hat{Y}_{\hat{p}}^{\text{cal},U} - Y}{N} = o_{\mathbb{P}}(1),$$

$$\frac{\hat{Y}_{\hat{p}}^{\text{cal},S} - Y}{N} = o_{\mathbb{P}}(1).$$

Subscript $\mathbb{P}$ *means that the reference probability distribution is that determined by the superpopulation model, the sampling design, and the response mechanism.*

The proof is given in the Appendix of the longer version of this article (Hasler, 2023). This result states that when the response probabilities are obtained via calibration, the resulting NWA estimators are consistent estimators of the true total. Result 3 holds even when the response model in Assumption (R3) is misspecified.

Result 4. *Let the sequence of sampling designs satisfy Assumptions (D1)-(D4), the response mechanism satisfy Assumptions (R1)-(R3), and the sequence of finite populations satisfy Assumptions (P1)-(P6) in the Appendix of Hasler (2023). Then*

$$\frac{\hat{Y}_{\hat{p}}^{\text{cal},U} - Y}{N} = o_p(1),$$
$$\frac{\hat{Y}_{\hat{p}}^{\text{cal},S} - Y}{N} = o_p(1).$$

The proof is given in the Appendix of the longer version of this article (Hasler, 2023). This result states that when the response probabilities are obtained via calibration, the resulting NWA estimators are consistent estimators of the true total when the response model is correctly specified. Result 4 holds even when the superpopulation model stated in Result 3 is misspecified. Note that the probability distribution in Result 4 is that determined by the sampling design and the response mechanism. The two quantities in Result 4 are therefore also $o_{\mathbb{P}}(1)$.

From Results 3 and 4, we conclude that the NWA estimators obtained with calibration are doubly robust. That is, these estimators remain consistent even when one of the two models, superpopulation model or response model, is misspecified. When the response probabilities are estimated via MLE, however, consistency of the resulting NWA estimator is not guaranteed under the assumptions stated in the results. Indeed, when the response probabilities are obtained via MLE from equation (3.1), the resulting weights may not be calibrated. This plays a central role in the proof of Result 3. As a result, if the double robustness of the NWA estimator obtained with MLE holds, further assumptions are required. This goes beyond the scope of this paper.

6. Variance and variance estimation

We suppose throughout this section that Assumption (R1) holds. Under nonresponse, we can write the variance of a generic estimator $\hat{Y}_g$ as

$$\mathrm{V}(\hat{Y}_g) = \mathrm{V}_{\text{sam}}(\hat{Y}_g) + \mathrm{V}_{\text{nr}}(\hat{Y}_g),$$

where the two terms are the *sampling variance* and the *nonresponse variance*, respectively, and are given by

$$\begin{aligned} \mathrm{V}_{\mathrm{sam}}(\hat{Y}_g) &= \mathrm{V}_p\left\{\mathrm{E}_q(\hat{Y}_g \mid S)\right\}, \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_g) &= \mathrm{E}_p\left\{\mathrm{V}_q(\hat{Y}_g \mid S)\right\}. \end{aligned}$$

Based on this decomposition, the variance of the estimator with the true response probabilities can be written as

$$\mathrm{V}(\hat{Y}_p) = \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right) + \mathrm{E}_p\left(\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}y_i^2\right).$$

Based on the same decomposition, Kim and Kim (2007), page 507, write the variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{mle}}$ as

$$\mathrm{V}\left(\hat{Y}_{\hat{p},l}^{\mathrm{mle}}\right) = \mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{mle}}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{mle}}),$$

where

$$\begin{aligned} \mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{mle}}) &= \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right), \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{mle}}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\mathrm{mle}})^2\right\}. \end{aligned}$$

The first term is the variance of the full sample HT estimator. The second term vanishes if there exists a vector $\boldsymbol{\beta}$ such that $y_i = k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\beta}$. This agrees with Remark 2 in Section 4 saying that $\hat{Y}_{\hat{p},l}^{\mathrm{mle}}$ matches the full sample HT estimator when this relationship holds.

A similar decomposition holds for the case when calibration is applied. More details can be found in Hasler (2023). The variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ can be written

$$\mathrm{V}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) = \mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}),$$

where

$$\begin{aligned} \mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) &= \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right), \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_S)^2\right\}. \end{aligned}$$

The first term is the variance of the full sample HT estimator. The second term vanishes if there exists a vector $\boldsymbol{\beta}$ such that $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. This agrees with Remark 4 saying that $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ matches the full sample HT estimator when this relationship holds.

Similarly, the variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ can be written

$$\mathrm{V}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) = \mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}),$$

where

$$\begin{aligned}
\mathrm{V}_{\mathrm{sam}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) &= \mathrm{V}_p\left\{\sum_{i\in S}\frac{1}{\pi_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U)\right\},\\
\mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U)^2\right\}.
\end{aligned}$$

The first term is the variance of the full sample HT estimator of the differences $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$. Both the first and second terms vanish if there exists a vector $\boldsymbol{\beta}$ such that $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. This agrees with Remark 6 saying that $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ matches the true population total, which has zero variance, when this relationship holds.

Remark 7. *The sampling variance of the linearized estimators $\hat{Y}_{\hat{p},l}^{\mathrm{mle}}$ and $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ is equal to the sampling variance of $\hat{Y}_p$. Their nonresponse variance is no greater than that of $\hat{Y}_p$. This means that the NWA estimators $\hat{Y}_{\hat{p}}^{\mathrm{mle}}$ and $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$ are asymptotically equivalent to estimators that are at least as efficient as the estimator with the true response probabilities. This was shown in Kim and Kim (2007) for $\hat{Y}_{\hat{p}}^{\mathrm{mle}}$, see page 505.*

We expect the sampling variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ to be smaller than the sampling variance of $\hat{Y}_p$ provided that the residuals $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$ have less variation than the y_i's. The nonresponse variance of $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ is no greater than that of $\hat{Y}_p$. Thus, $\hat{Y}_{\hat{p}}^{\mathrm{cal},U}$ is asymptotically equivalent to an estimator that is at least as efficient as the estimator with the true response probabilities under the condition stated above.

Overall, there seems to be a gain in efficiency when using estimated response probabilities as compared to true response probabilities, at least for large enough populations and samples. A possible explanation is that estimating response probabilities can be viewed as a smoothing of the weights using an appropriate model. Such a smoothing has already been shown to improve the efficiency of the usual Horvitz-Thompson estimator, see Beaumont (2008) for instance.

Remark 8. *Now comparing the variance of the NWA calibration estimators $\hat{Y}_{\hat{p}}^{\mathrm{cal},U}$ and $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$. We expect the sampling variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ to be smaller than the sampling variance of the linearized estimator $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ provided that the residuals $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$ have less variation than the y_i's. Moreover, we expect the nonresponse variance of $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ to be close to that of $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$, since the only difference is that the population coefficient $\boldsymbol{\gamma}_U$ in the nonresponse variance of the former is replaced by a sample estimator $\boldsymbol{\gamma}_S$ in the latter. In practice, this means that we expect a gain in efficiency of the NWA estimator when estimating the response probabilities via calibration at the population level as compared to the sample level.*

We suppose from this point and until the end of the current section that Assumptions (D1)-(D4), (R1)-(R3), and additional regularity conditions stated in the Appendix of the longer version of this article (Hasler, 2023) are satisfied. Using the decomposition of the variance above, the following estimator may be used for the variance of the NWA estimator $\hat{Y}_{\hat{p}}^{\text{mle}}$, see Kim and Kim (2007), page 507,

$$\hat{\text{V}}(\hat{Y}_{\hat{p}}^{\text{mle}}) = \hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{mle}}) + \hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{mle}}),$$

where

$$\begin{aligned}
\hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{mle}}) &= \sum_{i\in S_r} \frac{1-\pi_i}{\pi_i^2}\frac{y_i^2}{\hat{p}_i} + \sum_{i,j\in S_r, i\neq j} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\frac{y_i}{\hat{p}_i}\frac{y_j}{\hat{p}_j}, \\
\hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{mle}}) &= \sum_{i\in S_r} \frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}(y_i - k_i\pi_i\hat{p}_i\mathbf{x}_i^\top\hat{\boldsymbol{\gamma}}_n^{\text{mle}})^2, \\
\hat{\boldsymbol{\gamma}}_n^{\text{mle}} &= \left\{\sum_{i\in S_r} k_i(1-\hat{p}_i)\,\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1} \sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{x}_i y_i.
\end{aligned}$$

We consider the same approach to derive a variance estimator of NWA estimators $\hat{Y}_{\hat{p}}^{\text{cal},S}$ and $\hat{Y}_{\hat{p}}^{\text{cal},U}$. We obtain

$$\hat{\text{V}}(\hat{Y}_{\hat{p}}^{\text{cal},S}) = \hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{cal},S}) + \hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{cal},S}),$$

where

$$\begin{aligned}
\hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{cal},S}) &= \sum_{i\in S_r} \frac{1-\pi_i}{\pi_i^2}\frac{y_i^2}{\hat{p}_i} + \sum_{i,j\in S_r; i\neq j} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\frac{y_i}{\hat{p}_i}\frac{y_j}{\hat{p}_j}, \\
\hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{cal},S}) &= \sum_{i\in S_r} \frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}\left(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\gamma}}_n^{\text{cal}}\right)^2 \\
\hat{\boldsymbol{\gamma}}_n^{\text{cal}} &= \left(\sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{x}_i y_i.
\end{aligned}$$

Similarly, we obtain

$$\hat{\text{V}}(\hat{Y}_{\hat{p}}^{\text{cal},U}) = \hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{cal},U}) + \hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{cal},U}),$$

where

$$\begin{aligned}
\hat{\text{V}}_{\text{sam}}(\hat{Y}_{\hat{p},l}^{\text{cal},U}) &= \sum_{i\in S_r} \frac{1-\pi_i}{\pi_i^2}\frac{e_i^2}{\hat{p}_i} + \sum_{i,j\in S_r; i\neq j} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\frac{e_i}{\hat{p}_i}\frac{e_j}{\hat{p}_j}, \\
e_i &= y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\gamma}}_n^{\text{cal}}, \\
\hat{\text{V}}_{\text{nr}}(\hat{Y}_{\hat{p},l}^{\text{cal},U}) &= \sum_{i\in S_r} \frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}e_i^2.
\end{aligned}$$

More details can be found in Hasler (2023).

7. Simulation study

7.1 Simulation settings

Five different populations are considered and obtained as follows. For each population, we generate $N = 2{,}000$ population units. The auxiliary variables are the same across all five populations and are $\mathbf{x}_i = (1, x_i)^\top$ where x_i are observations of independent and identically distributed (iid) uniform random variables with parameters, i.e., bounds, 0 and 100. The values of the variables of interest are obtained as follows:

$$
\begin{aligned}
y_{1i} &= 1{,}000 + 20x + \varepsilon_{1i}, \\
y_{2i} &= 1{,}500 + 500\exp(-10 + 0.1x) + \varepsilon_{2i}, \\
y_{3i} &= \begin{cases} 1 & \text{with probability } \phi_i, \\ 0 & \text{otherwise,} \end{cases} \quad \text{where} \quad \phi_i = \begin{cases} 0.8 & \text{if } x_i > 75, \\ 0.2 & \text{otherwise,} \end{cases} \\
y_{4i} &= 1{,}000 + \varepsilon_{4i}, \\
y_{5i} &= 1{,}000 + 20x + \varepsilon_{5i},
\end{aligned}
$$

where $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{4i},$ and ε_{5i} are observations of iid random normal distributions with mean 0 and standard deviation 750, 100, 750, and 50, respectively. In population 1, there is a linear relationship between x and y_1 with a correlation of approximately 0.6. In population 2, there is a non-linear relationship between x and y_2. In population 3, y_3 is categorical and the values are obtained from independent Bernoulli random variables with parameter 0.8 for large values of x and 0.2 for small values of x. In population 4, there is no relationship between x and y_4. In population 5, there is a very strong linear relationship between x and y_5 with a correlation of approximately 0.99.

Two vectors of response probabilities are created as follows

$$
\begin{aligned}
p_{1i} &= \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\lambda})}, \\
p_{2i} &= \begin{cases} 1 - a_1 (x_i - k_1)^2 + h_1 & \text{if } a_1 (x_i - k_1)^2 + h_1 > 0.01, \\ 0.9 & \text{otherwise,} \end{cases}
\end{aligned}
$$

where $a_1 = -0.0005$, $k_1 = 25.79116$, $h_1 = 0.9$, and $\boldsymbol{\lambda} = (-2, 0.04)^\top$. Both vectors are constructed so that they yield a population mean response rate of approximately 50%. Note that depending on the selected sample, the sample mean response rate may be larger or smaller than 50% as units are not necessarily selected uniformly across all values of x. For the first vector of response probabilities, the logistic regression model in equation (2.4) is correctly specified. For the second vector of response probabilities, this model is misspecified. For both vectors, large values of x tend to have large response probabilities. Figure 7.1 shows the five populations and Figure 7.2 the response probabilities as a function of the values of x.

Two sampling designs are considered: 1) simple random sampling without replacement where $n = 200$ units are selected; 2) stratified sampling where two strata are considered. The first stratum contains the units with a x-value smaller than the median value of x, the second stratum contains those units with a x-value larger the median. Forty units are selected from the first stratum using simple random sampling. The sampling fraction in the first stratum is 4%. One hundred and sixty units are selected from the second stratum using simple random sampling. The sampling fraction in the second stratum is 16%.

Ten thousand simulations are run as explained in what follows for each population, each sampling design, and each vector of response probabilities. This results in 20 scenarios. A sample of size $n = 200$ is selected according to the sampling design. A set of respondents is generated with Poisson sampling design with the vector of response probabilities. Function `optim` is used to solve the estimating equations to obtain the parameters of the response model as presented in Section 3. The function minimizes the maximum of the absolute relative value of the left-hand-side of estimating equations (3.1), (3.2), and (3.3) over the auxiliary variables. We define that the algorithm converges if this maximum is less than 0.01. The initial value of the parameter vector is set to $(0, 0)$ so that the initial response probabilities are all 1/2. When comparing the performance of the NWA estimators and their variance estimators, only those simulation runs for which the algorithm converges are used for computing comparison measures of a given estimator. The total Y is estimated via seven estimators listed below.

1. $\hat{Y}$ (HT): the Horvitz-Thompson estimator. Note that this estimator is unavailable in practice with nonresponse. It serves here as a comparison point.
2. $\hat{Y}_p(p)$: estimator with the true response probabilities in (2.2). This estimator is unavailable in practice. It serves here as a comparison point.
3. $\hat{Y}_{\text{naive}}$ (naive): estimator that ignores nonresponse, that is $\hat{Y}_{\text{naive}} = \frac{n}{n_r} \sum_{i \in S_r} \frac{y_i}{\pi_i}$.
4. $\hat{Y}_{\hat{p}}^{\text{mle},1}$ (mle, 1): NWA estimator with response probabilities estimated via MLE, equation (3.1), with $k_i = 1$.
5. $\hat{Y}_{\hat{p}}^{\text{mle},1/\pi}$ (mle, $1/\pi$): NWA estimator with response probabilities estimated via MLE, equation (3.1), with $k_i = 1/\pi_i$.
6. $\hat{Y}_{\hat{p}}^{\text{cal},U}$ (cal, U): NWA estimator with response probabilities estimated via calibration at the population level, equation (3.3).
7. $\hat{Y}_{\hat{p}}^{\text{cal},S}$ (cal, S): NWA estimator with response probabilities estimated via calibration at the sample level, equation (3.2).

Figure 7.1 Five populations.

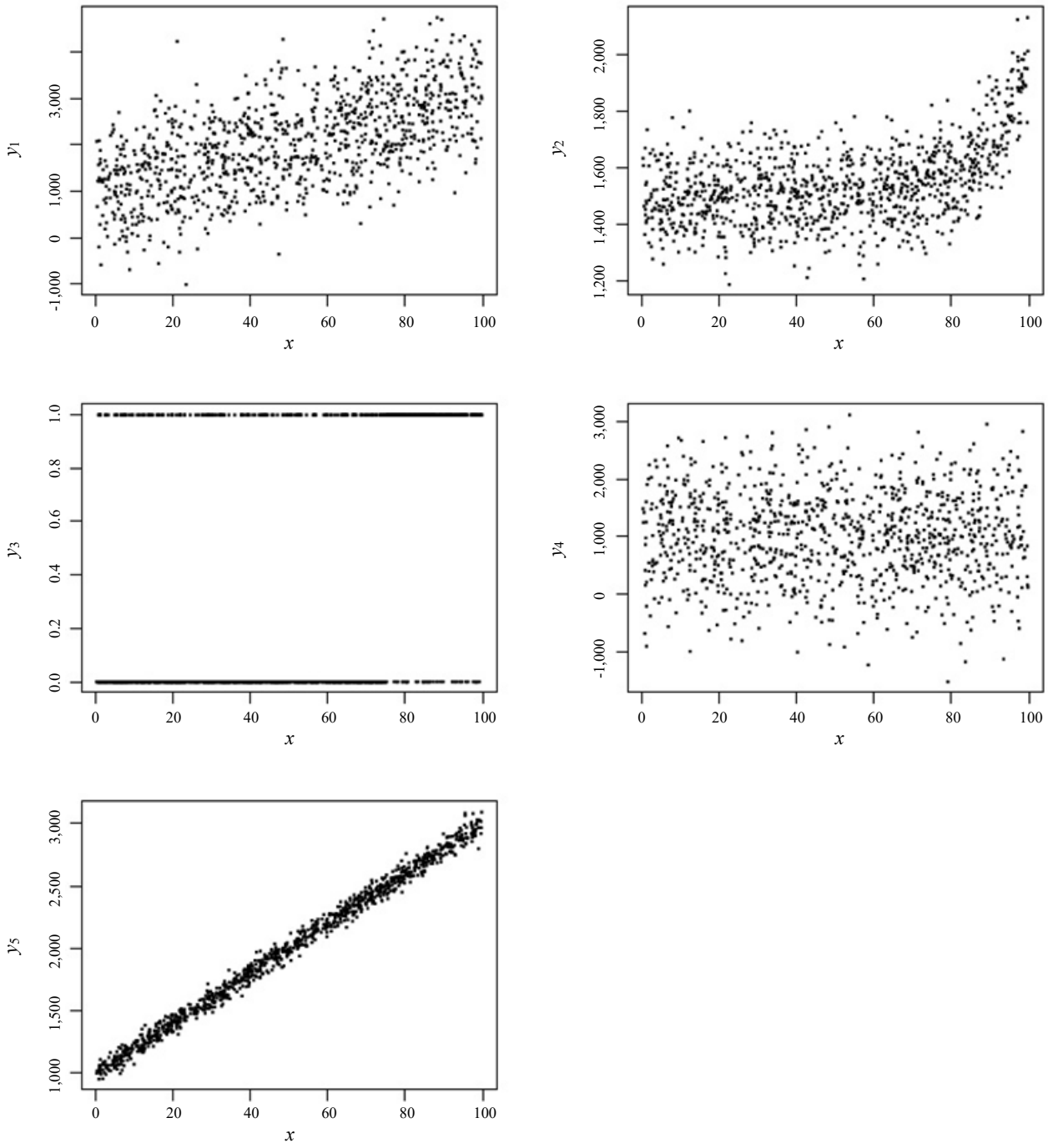
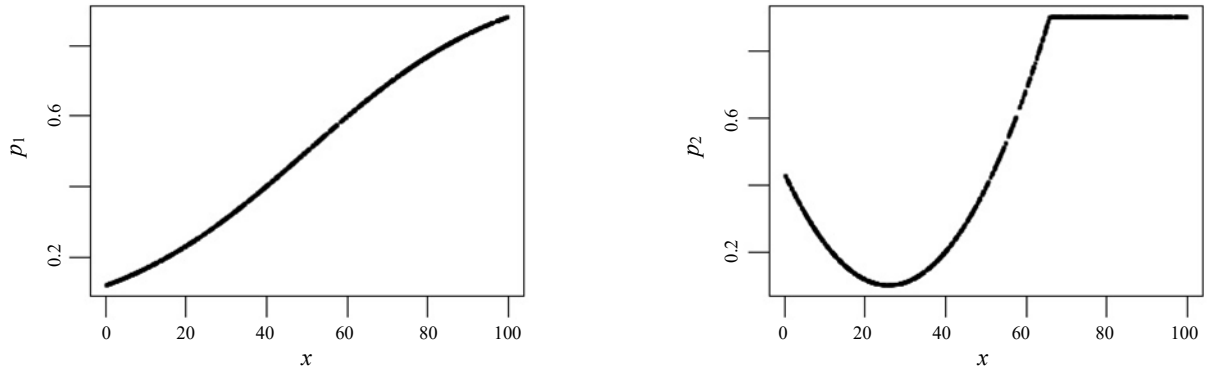


Figure 7.2 Two vectors of response probabilities.



7.2 Performance of the NWA estimators

The performance of the estimators is assessed through the following comparison measures defined for a generic estimator $\hat{Y}_g$:

- Absolute Monte Carlo relative bias $(|\text{RB}|)$ defined as

$$|\text{RB}| = \left|\frac{B}{Y}\right|,$$

where $\text{B} = \hat{Y}_g^{(\cdot)} - Y$, $\hat{Y}_g^{(\cdot)}$ is the mean of the estimator over the L simulation runs (or the L simulation runs for which the optimization algorithm converges if $\hat{Y}_g$ is a NWA estimator),

$$\hat{Y}_g^{(\cdot)} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{Y}_g^{(\ell)},$$

and $\hat{Y}_g^{(\ell)}$ is the estimator $\hat{Y}_g$ obtained at the ℓ th simulation,

- Monte Carlo relative standard deviation (RSd) defined as

$$\text{RSd} = \frac{(\text{VAR})^{1/2}}{Y},$$

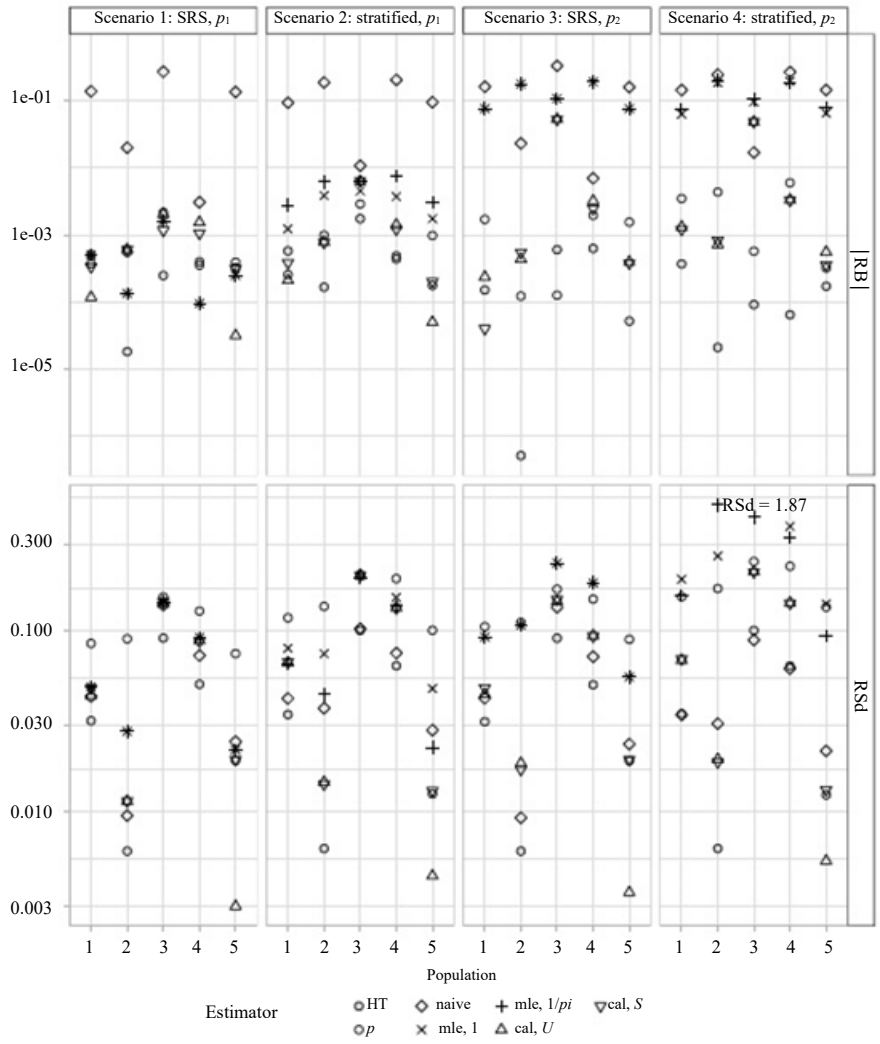
where

$$\text{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L} (\hat{Y}_g^{(\ell)} - \hat{Y}_g^{(\cdot)})^2.$$

The results are presented in Figure 7.3. The y-axes are displayed in logarithmic scales. For the plots of RSd, the maximum value on the y-axis is set to 0.5 for clarity reasons. One estimator has a value larger than 0.5 in scenario 4, population 2. This value is labeled on the graph. In scenarios 1 and 2, when the model for the response probabilities is correctly specified, all four NWA estimators show a RB of the same order of magnitude as the RB of the HT estimator and the estimator with the true response probabilities $\hat{Y}_p$. These last two estimators being unbiased, this result illustrates how the four NWA estimators are nearly unbiased, see Remarks 1, 3, and 5. In scenarios 3 and 4, when the model for the response probabilities is incorrectly specified, the two NWA estimators with response probabilities estimated via calibration show a RB of the same order of magnitude as the RB of the HT estimator and the estimator with the true response probabilities

$\hat{Y}_p$. The two estimators with response probabilities estimated via MLE show a larger RB. This illustrates how calibration may provide a stronger protection against misspecification of the model for the response probabilities as compared to MLE. In all four scenarios, the naive estimator yields the larger RB.

Figure 7.3 |RB| and RSd for seven estimators, five populations, and 4 scenarios.



In scenarios 1 and 2, when the model for the response probabilities is correctly specified, all four NWA estimators show a variance smaller than the variance of the estimator with the true response probabilities $\hat{Y}_p$. This confirms that a gain in efficiency of the total estimator is obtained when estimating the response probabilities via MLE or calibration as compared to using the true response probabilities, see Remark 7. In these two scenarios, all four NWA estimators show a RSd of the same order of magnitude. In scenarios 3 and 4, when the model for the response probabilities is incorrectly specified, the two NWA estimators with response probabilities estimated via calibration show a RSd smaller than the RSd of the two NWA estimators with response probabilities estimated via MLE. This illustrates how calibration may provide a stronger protection against misspecification of the model for the response probabilities as compared to MLE.

7.3 Performance of the variance estimators

The variance of the four NWA estimators is estimated for each simulation run with the formulae of Section 6. The performance of the variance estimators is assessed through the following comparison measures defined for a generic estimator $\hat{Y}_g$:

- Absolute Monte Carlo relative bias $(|\text{RB}|)$ defined as

$$|\text{RB}| = \frac{|B|}{\text{V}_{\text{sim}}(\hat{Y}_g)},$$

  where $\text{V}_{\text{sim}}(\hat{Y}_g)$ is the variance of $\hat{Y}_g$ over the L simulation runs for which the optimization algorithm converges, $\text{B} = \hat{\text{V}}(\hat{Y}_g^{(\cdot)}) - \text{V}_{\text{sim}}(\hat{Y}_g)$, and $\hat{\text{V}}(\hat{Y}_g^{(\cdot)})$ is the mean of $\hat{\text{V}}(\hat{Y}_g)$ over these L simulation runs,
- CR: the actual coverage rate of the 95% confidence interval, i.e., the proportion of simulation runs for which the 95% confidence interval contains the true total Y.

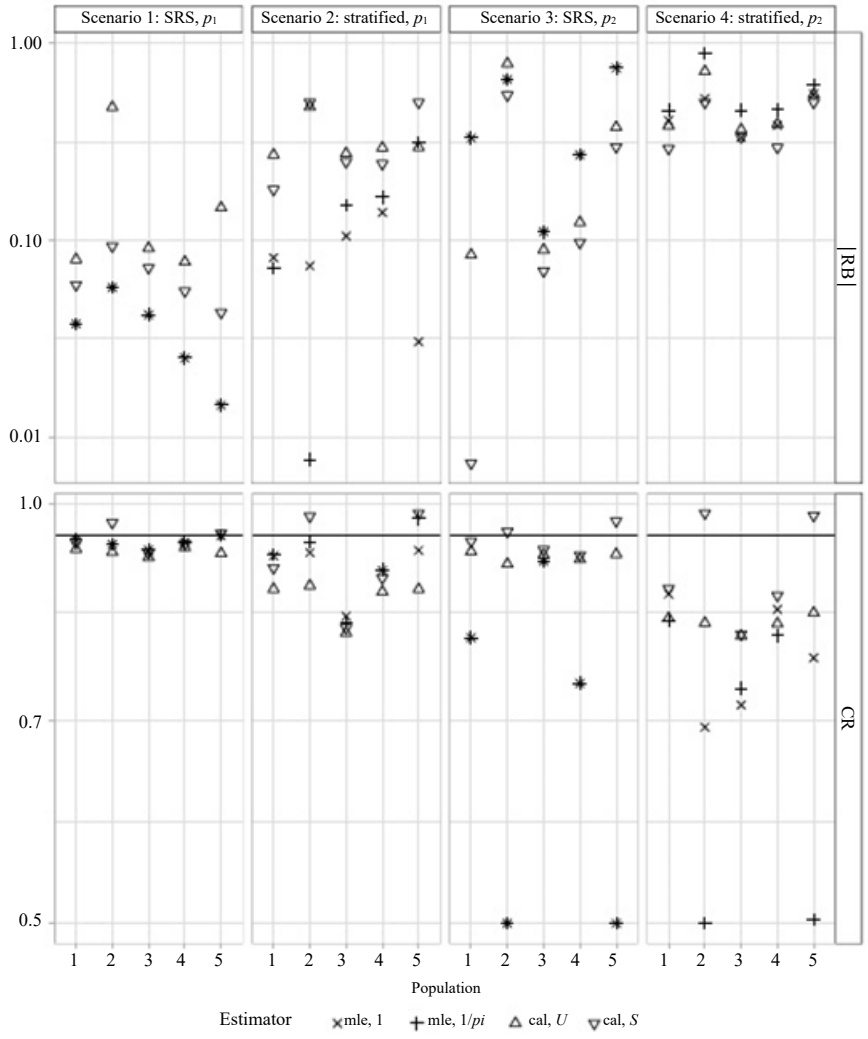
The results are presented in Figure 7.4. The y-axes are displayed in logarithmic scales. To ease the reading of the graphs, four RB larger than 1 were set to 1 and five CR smaller than 0.5 were set to 0.5. In scenarios 1 and 2, when the model for the response probabilities is correctly specified, the RB of the variance estimator with response probabilities estimated via MLE tends to be smaller than the RB of the variance estimator with response probabilities estimated via calibration. In scenarios 3 and 4, when the model for the response probabilities is incorrectly specified, it is the opposite. In scenarios 1 and 2, all four variance estimators yield a CR generally close to the nominal coverage of 95%. In scenarios 3 and 4, the variance estimator with response probabilities estimated via MLE yields very low CR in several cases.

7.4 Weights and convergence

In some cases, the estimating equations used to obtain estimated response probabilities may not admit a solution. In other cases, a solution to the estimating equations exists but the resulting weights, that is, the inverse of the estimated response probabilities, may be very large. Section 6 of Hasler (2023) provides details and explanations. In order to illustrate these problems of convergence and extreme weights, the following three comparison measures are computed for each NWA estimator

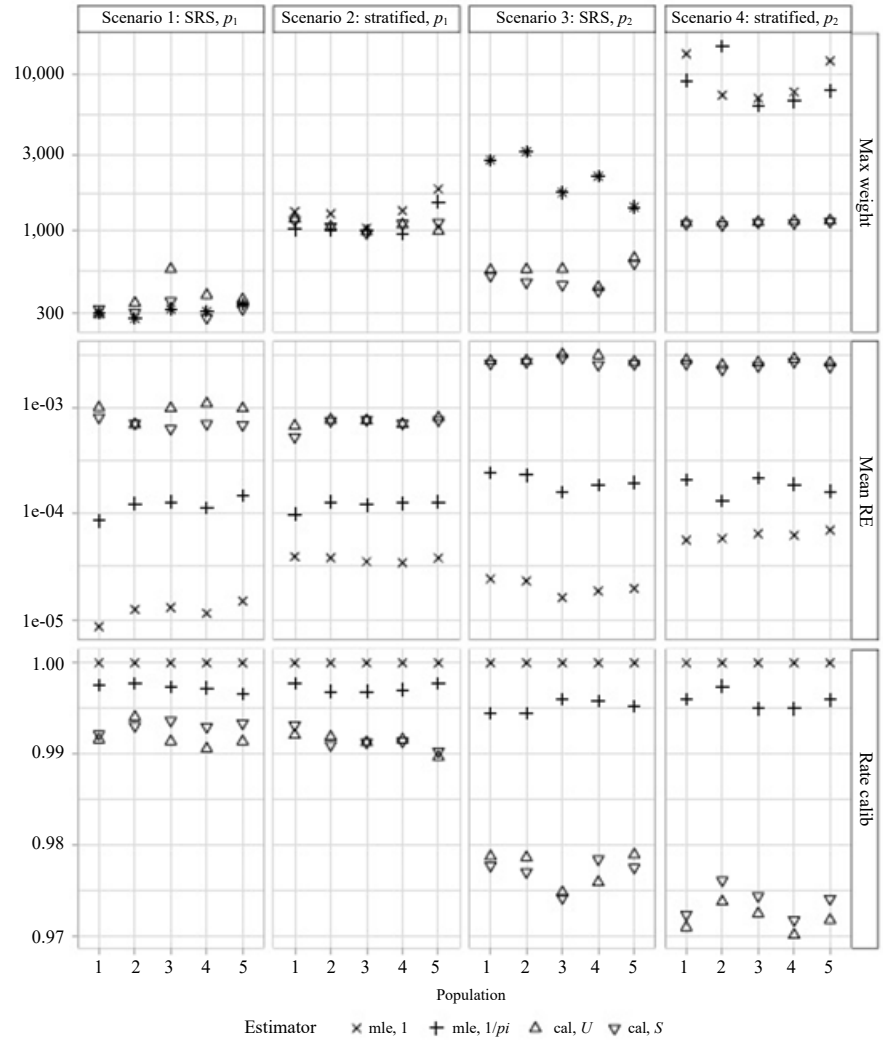
- Maximum weight: the largest final adjusted weight $1/(\pi_i \hat{p}_i)$ over all 10,000 simulations,
- Mean Relative Error (Mean RE): the mean over 10,000 simulations of the maximum of the absolute relative error of the estimating equation,
- Rate calib: the proportion of simulations for which the Mean RE is smaller than the threshold of 0.01. We define that the algorithm converges to a solution when the Mean RE is smaller than this threshold.

Figure 7.4 |RB| and CR for four variance estimators, five populations, and 4 scenarios.



The results are presented in Figure 7.5. The y-axes are displayed in logarithmic scales. One estimator yields a Max weight of more than 400,000 in Scenario 4. To ease the reading of the graphs, this value is set to 15,000. In scenarios 1 and 2, when the model for the response probabilities is correctly specified, all four NWA estimators yield max weights close to one another. No extreme weights is noticeable. In scenario 3 and 4, when the model for the response probabilities is incorrectly specified, very large weights are obtained with MLE, more so in Scenario 4. Calibration may protect against extreme weights when the response model is misspecified. In all four scenarios, the mean RE is smaller with MLE than with calibration. This difference is larger in scenarios 3 and 4, when the model for the response probabilities is incorrectly specified. Moreover, the algorithm yields a mean RE smaller than the threshold of 0.01 more often with MLE than with calibration. This illustrates how the algorithm applied to obtain the response model parameters converges more often to a solution to the estimating equations of MLE than to a solution to the estimating equations of calibration.

Figure 7.5 Max weight, mean relative error, and rate of calibration for four NWA estimators, five populations, and 4 scenarios.



8. Discussion

We build on Kim and Kim (2007) and develop asymptotic properties of the NWA estimator when calibration is applied to estimate the response probabilities. For the first time, a common theoretical framework is considered for both approaches to NWA estimation, namely MLE and calibration. This allows us to compare the asymptotic behavior of four estimators in terms of bias and variance under common assumptions. We postulate a logistic regression model for the response probabilities. We consider two levels of calibration: population and full sample. The main results are 1) the NWA estimators with the response probabilities estimated via calibration are asymptotically equivalent to unbiased estimators, 2) a gain in efficiency is obtained when estimating the response probabilities via calibration as compared to the estimator with the true response probabilities, 3) the NWA estimator with the response probabilities

estimated via calibration at the population level is generally more efficient than the NWA estimator with response probabilities estimated via calibration at the sample level, 4) calibration may better protect against model misspecification than maximum likelihood when applied to estimate the response probabilities used in the NWA estimator, and 5) we explain and illustrate the problems of convergence to a solution to the estimating equations and extreme weights. The paper studies and compares NWA estimators obtained either via MLE or direct calibration (one-step approach). Some authors suggest the two-step approach, i.e., first estimate the response probabilities via MLE in order to bypass the problem of extreme weights and then calibrate to further improve the efficiency of the NWA estimator, see Haziza and Lesage (2016) and Haziza and Beaumont (2017), page 222. This goes beyond the scope of this research and is the subject of future work.

Acknowledgements

This research was supported by the Swiss Federal Statistical Office. The author thanks Pr. Yves Tillé, two referees, an Associate Editor, and the Editor for constructive comments. The views expressed in this article are those of the author solely and do not necessarily reflect those of the aforementioned organization and persons.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95(3), 539-553.

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

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1983). Some uses of statistical models in connexion with the nonresponse problem. In *Incomplete Data in Sample Surveys*, (Eds., W.G. Madow and I. Olkin), 3, 143-160. New York: Academic Press.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Paris. Insee-Méthodes.

Deville, J.-C., and Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, INSEE, Paris, 53-70.

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

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Dupont, F. (1993). Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. *Actes des Journées de Méthodologie Statistique*, INSEE, Paris, 9-42.

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the finish household budget survey. *Journal of Official Statistics*, 3, 325-337.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section*, American Statistical Association, 197-202.

Hasler, C. (2023). Inference from sampling with response probabilities estimated via calibration. Technical report, University of Neuchâtel. Available on ArXiv at DOI: https://doi.org/10.48550/arXiv.2202.03897.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Haziza, D., and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

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.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642.

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

Kim, J.K., and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 35(4), 501-514.

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

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

Kott, P.S. (2012). Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups. *Survey Methodology*, 38, 1, 95-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11689-eng.pdf.

Kott, P.S., and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods*, 6(2), 105-111.

Lesage, E., Haziza, D. and D'Haultfoeuille, X. (2019). A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *Journal of the American Statistical Association*, 114(526), 906-915.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Oh, H.L., and Scheuren, F. (1983). Weighted adjustment for nonresponse. In *Incomplete Data in Sample Survey*, (Eds., W.G. Madow, H. Nisselson and I. Olkin), 2, 143-184. New York: Academic Press.

Ranalli, M., Matei, A. and Neri, A. (2023). Generalised calibration with latent variables for the treatment of unit nonresponse in sample surveys. *Statistical Methods and Applications*, 32(1), 169-195.

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

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

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55(3), 279-294.

Tillé, Y., and Matei, A. (2021). *sampling: Survey Sampling*. R package version 2.9.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 261-285
Statistics Canada, Catalogue No. 12-001-X

Relaxed calibration of survey weights

Nicholas T. Longford[1]

Abstract

Population surveys are nowadays rarely analysed in isolation from any auxiliary information, often in the form of population counts, totals and other summaries. Calibration, or benchmarking, by which the weighted sample totals of auxiliary variables are matched to their (known) population totals, is widely applied. Methods for adjusting the weights to satisfy these constraints involve iterative procedures with unknown finite-sample properties. We develop an alternative method in which the weights are calibrated by minimising a quadratic function, requiring no iterations and yielding a unique solution. The relative priority of each constraint is represented by a tuning parameter. The properties of the weights and of the calibration estimator, as functions of these parameters, are explored analytically and by simulations. A connection of the proposed method with ridge calibration is established.

Key Words: Auxiliary information; Benchmarking; Priority; Ridge regression; Survey sampling; Weight adjustment.

1. Introduction

Calibration, or benchmarking, is generally regarded as an indispensable accompaniment of estimation of population summaries in large-scale surveys that are conducted in an environment in which other data sources provide auxiliary information. Such information has a potential to make estimation more efficient or for it to acquire some other valuable properties or attributes. Calibration has an important role in compensating for imperfections in the sampling design and its implementation, such as deficiencies in the sampling frame and nonresponse.

Calibration has an extensive literature; Deville and Särndal (1992) is widely regarded as a landmark, reinforced by Lundström and Särndal (1999) for its application in the context of modern official statistics. Estevao and Särndal (2006) and Särndal (2007) review subsequent developments. For more recent literature surveys, see Kim and Park (2010), Brick (2013), Wu and Lu (2016) and Lohr and Raghunathan (2017). The monograph of Tillé (2020) contains a comprehensive treatment of the subject. Devaud and Tillé (2019) is an appraisal of the impact of Deville and Särndal (1992) on survey sampling, and official statistics in particular. Davies (2018) reviews a wide range of methods for calibration. In his terminology, hard calibration refers to optimisation while satisfying a set of constraints with no scope for any discrepancy. We prefer soft calibration in which a compromise is sought among the constraints and objectives of weight adjustment and subsequent estimation.

Model-based and model-assisted approaches have found a fertile ground in survey sampling in general, and weight calibration in particular. Ordinary regression and its various generalizations have been widely applied; see Haziza and Beaumont (2017) and references therein. One such approach is motivated by ridge

1. Nicholas T. Longford, SNTL Statistics Research and Consulting, London, UK, 3 Badgers Walk, Whyteleafe CR3 0AS, Surrey, United Kingdom. E-mail: sntlnick@sntl.co.uk.

regression (Hoerl and Kennard, 1970), in which the imperative of satisfying the benchmark constraints is moderated to promote stability of the solution (calibration weights), to avoid large adjustments, unacceptable and extreme values, and reduce the dispersion of the weights. Beaumont and Bocci (2008) relate the original proposal by Chambers (1996) to the likelihood-based approach of Chen, Sitter and Wu (2002).

Calibration is especially challenging when a lot of auxiliary information is available, and many population totals are to be matched. Cardot, Goga and Shehzad (2017) and Vera, Sánchez Zuleta and Rueda (2023) address this problem by projecting the auxiliary data onto a manageable subspace for which established methods can be applied. See also Dagdoug, Goga and Haziza (2023) for a model-assisted approach.

We introduce a method of benchmarking in which the goal of matching the calibration constraints is relaxed to reducing the discrepancies between the sample summaries and their targets in line with specified priorities. The algorithm we develop caters for the same constituency of problems as the established methods of calibration but permits integration of the analyst's (or their client's) priorities in a flexible and transparent manner. The algorithm requires no model-related assumptions but is closely related to ridge calibration, in which a model is implied (Chambers, 1996; Rao and Singh, 2009). The priorities turn out to be very much like the reciprocals of the ridge coefficients. This approach has some commonality with Guggemos and Tillé (2010) who combine hard calibration with penalisation and regard it as a design-based alternative to procedures based on mixed models. Our approach is based entirely on penalisation, although by relating it to ridge calibration we establish a link with linear models.

The algorithm is computationally undemanding and some of its properties are derived analytically. Specification of the priorities as tuning parameters may seem like an additional burden; however, these parameters facilitate a control of the process of calibration that is not available with some established methods.

For a similar computational approach in an unrelated context, namely, balancing in causal inference, see Longford (2024). It replaces the goal of achieving a balance of specified quality for two treatment groups on a set of background variables with the best balance that can be achieved given the analyst's priorities – the relative urgency or importance of reducing the imbalance for each background variable.

The remainder of this section sets up the notation and presents the analytical context of the problem. The next section formulates the problem, derives its solution and relates it to ridge calibration. Section 3 discusses how the tuning parameters are to be set. Section 4 illustrates the method on examples. Section 5 presents a simulation study that demonstrates the efficiency of the proposed estimator and explores the value of auxiliary information and good implementation of the sampling design. The concluding section summarises the method, its strengths and full potential, and discusses some unresolved issues.

1.1 Notation and context

In a population survey with a specified sampling design, we consider an estimator of the population total of a variable y, linear in the vector of its observed values $\mathbf{y} = (y_1, \ldots, y_n)^\top$. For example, when the sample size n is fixed, this may be the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), $\hat{\theta}_{\text{HT}} = \mathbf{w}^\top \mathbf{y}$, where $\mathbf{w} = (w_1, \ldots, w_n)^\top$ is the vector of sampling weights, equal to the reciprocals of the probabilities of inclusion of the subjects in the sample. These probabilities are set by design and, possibly, adjusted after the sample is drawn.

Suppose the population totals t_k are known for one or several variables X_k, $k = 1, \ldots, K$. Denote by $\mathbf{X}^\circ$ their collection. We use $\mathbf{X}^\circ$ also for the $n \times K$ matrix formed by the values of these variables in the sample; $\mathbf{x}_k$ denotes column k of $\mathbf{X}^\circ$. To simplify the discourse, we assume that every variable X_k in $\mathbf{X}^\circ$ is either ordinal (continuous) or binary. A discrete variable with $H \geq 2$ categories is represented in $\mathbf{X}^\circ$ by H binary (dummy) variables. Singularity of $\mathbf{X}^\circ$ will raise no issues. We reserve the subscript 0 for the intercept, $\mathbf{x}_0 = (1, \ldots, 1)^\top$, and denote $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}^\circ)$.

Calibration is defined as a transformation of the weights, $\mathbf{u} = C(\mathbf{w})$, for which the weighted total $\mathbf{x}_k^\top \mathbf{u}$ matches the population total t_k for every variable k. That is, calibration arranges that $\mathbf{X}^{\circ\top} \mathbf{u} = \mathbf{t}^\circ$, where $\mathbf{t}^\circ = (t_1, \ldots, t_K)^\top$. Transformations of the original variables, including interactions (products), may be added to $\mathbf{X}^\circ$ when their population totals are known or are estimated with negligible error. Other population summaries, such as variances and quantiles, can also be matched.

We write $C(\mathbf{w}) = C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)$ to indicate the auxiliary information involved; this is useful when we consider which variables to include in $\mathbf{X}^\circ$. We may qualify the estimator $\hat{\theta}$ similarly, by writing $\hat{\theta}(\mathbf{u})$ and $\hat{\theta}\{C(\mathbf{w})\}$ or, more completely, $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)\}$. The outcomes $\mathbf{y}$ play no role in the search for $\mathbf{u} = C(\mathbf{w})$. Therefore, so long as $\mathbf{y}$ is not inspected until we settle on a particular calibration $C(\mathbf{w})$, the properties of estimator $\hat{\theta}(\mathbf{u})$ can be assessed without any regard for how $\mathbf{u}$ was derived. No concerns about capitalising on chance or data-snooping arise, even if calibration is explored using several matrices $\mathbf{X}^\circ$ and parameters involved in C. Of course, the properties of $\hat{\theta}(\mathbf{u})$ depend on $\mathbf{u}$.

In one perspective, when the estimation error $\hat{\theta} - \theta$ and its stochastic summary, such as the bias or mean squared error (MSE), are the sole concern, calibration has a distinctly cosmetic quality. In another, prevailing in practice, it is essential for the credibility of the estimates, even at the expense of some bias and inflation of MSE. Only when this inflation is substantial, or the weights $C(\mathbf{w})$ are much more dispersed than $\mathbf{w}$, some improvisation is called for. This often happens for relatively large K, when there are many auxiliary variables and constraints associated with them. This problem is usually resolved by removing some variables from $\mathbf{X}^\circ$.

We regard the dichotomy of including or excluding a variable in the calibration process as too rigid, and develop an approach in which the calibration constraints are assigned priorities that reflect the importance or urgency to match the weighted sample total of auxiliary variable X_k with its (population) target t_k. Priority is assigned also to other desirable properties: small alteration of the weights by the adjustment, preference for smaller dispersion of the elements of $\mathbf{u}$ and aversion to a change of the total of the weights.

In brief, the established constraints of exact match (no discrepancy) are replaced by penalties for the discrepancies. These penalties allow for some slack or leeway; they involve user-defined coefficients that quantify the relative priority of the constraints.

Our formulation of the problem leads to quadratic optimisation that has a closed-form solution. It involves no iterations and no large matrices have to be inverted (numerically), even with many margins being matched for large-scale data. For a set of tuning parameters, called *priorities*, the solution is unique, and its dependence on these priorities is easy to explore, dispensing with the need for any asymptotic theory and extensive experimentation with a black-box-like algorithm.

The priorities have a natural interpretation as the importance or urgency of each calibration constraint. More precisely, these constraints are not satisfied exactly, as stipulated in hard calibration, but each discrepancy $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$, an element of $\boldsymbol{\delta}^\circ = \mathbf{X}^{\circ\top}\mathbf{u} - \mathbf{t}^\circ$, may be made negligible, $\delta_k \doteq 0$, by setting the corresponding priority sufficiently high. At the other extreme, zero priority for variable k is equivalent to dropping column $\mathbf{x}_k$ from $\mathbf{X}^\circ$. Such flexibility may be seen as a distraction, imposing the burden of declaring the priorities, and having to justify the choice in a subsequent report. However, it offers an opportunity to incorporate the client's perspective, value judgements, insights and remits in the analysis. Also, one or a few outlying discrepancies $|\delta_k|$ can be reduced by increasing the corresponding priorities, possibly at the expense of increasing some other discrepancies.

2. Unconstrained optimisation

Denote by $\mathbf{0}_K$ the vector of zeros of length K. The subscript K is dropped when the length of the vector is obvious from the context. We use the symbols $\mathbf{1}_n$ and $\mathbf{1}$ similarly for the vector of ones, and $\mathbf{I}$ for the identity matrix. For a population of (finite) size N, we consider a sample of (fixed or random) size $n \ll N$ with the vectors of observations $\mathbf{y}$ and base weights $\mathbf{w}$.

A typical approach to calibration imposes the constraint $\delta_k = 0$ or specifies an upper bound Δ_k on $|\delta_k|$ for each $k = 1, \ldots, K$. These constraints can be replaced by a single upper bound for the sum of squares $\boldsymbol{\delta}^{\circ\top}\boldsymbol{\delta}^\circ = \delta_1^2 + \cdots + \delta_K^2$. For larger K we may distinguish variables k for which the match, $\delta_k = 0$, is more important than for the rest. Further, a separate priority coefficient may be assigned to each variable, or the variables may be partitioned to sets with constant coefficients within these sets. The squares in the sum $\sum_k \delta_k^2$ may be associated with weights, imposing an upper bound on $\boldsymbol{\delta}^{\circ\top}\mathbf{P}^\circ\boldsymbol{\delta}^\circ = \sum_k p_k^\circ \delta_k^2$, with the priorities $p_k^\circ > 0$ set by the analyst; $\mathbf{P}^\circ$ is the diagonal matrix with $p_1^\circ, \ldots, p_K^\circ$ on its diagonal.

These ways of relaxing the calibration constraints and introducing priorities for reducing the discrepancies motivate our proposal. We assign a nonnegative priority p_k° for each variable k and consider first finding the minimum of the function

$$F^\circ(\mathbf{u}; \mathbf{w}) = \sum_{k=1}^{K} p_k^\circ \delta_k^2,$$

subject to the constraints of small deviation of $\mathbf{u}$ from $\mathbf{w}$ and of the weight total matching the population size N; $\mathbf{u}^\top \mathbf{1}_n = N$. When N is not known we replace it by its estimate $\mathbf{w}^\top \mathbf{1}_n$. Parameter p_k° is confounded with the scale of $\mathbf{x}_k$; replacing $\mathbf{x}_k$ with $c^{-1}\mathbf{x}_k$ is compensated by replacing p_k° with $c^2 p_k^\circ$. To avoid the associated ambiguity, we assume that each $\mathbf{x}_k$ is standardised, that is, linearly transformed to have zero mean and unit variance; t_k is adjusted accordingly. As an alternative the values $\mathbf{x}_k$ and t_k may be retained and the corresponding priority rescaled.

We relax the constraint $\mathbf{u}^\top \mathbf{1} = t_0$, where $t_0 = N$ or $t_0 = \mathbf{w}^\top \mathbf{1}$, and incorporate it in the objective function F°. We define priority p_0° by the importance assigned to δ_0^2 attaining a small value. The adapted function F° includes the new term $p_0^\circ \delta_0^2$; that is, its summation is now from $k = 0$ to K. It corresponds to attaching $\mathbf{x}_0 = \mathbf{1}_n$ to $\mathbf{X}^\circ$ as another column, forming the matrix $\mathbf{X}$.

We assign priority S to small dispersion of the weights $\mathbf{u}$, motivated by the desire for small variance of $\hat{\theta}(\mathbf{u}) = \mathbf{u}^\top \mathbf{y}$, and R to the desire for small alteration of the weights, which is indirectly related to bias reduction, preferring $\hat{\theta}(\mathbf{u})$ to remain close to estimator $\hat{\theta}(\mathbf{w})$ which would be unbiased if the weights $\mathbf{w}$ were correct and n fixed. Note that small variance cannot be equated with efficiency.

Instead of $F^\circ(\mathbf{u})$ we find the (unconstrained) minimum of the function

$$\begin{aligned} F(\mathbf{u}; \mathbf{w}) &= \sum_{k=0}^{K} p_k^\circ \delta_k^2 + R(\mathbf{u}-\mathbf{w})^\top (\mathbf{u}-\mathbf{w}) + S\left(\mathbf{u}^\top \mathbf{u} - \frac{1}{n}\mathbf{u}^\top \mathbf{1}\mathbf{1}^\top \mathbf{u}\right). \\ &= (R+S)\,\mathbf{u}^\top \mathbf{u} - \frac{1}{n} S\,\mathbf{u}^\top \mathbf{1}\mathbf{1}^\top \mathbf{u} + \sum_{k=0}^{K} p_k^\circ \delta_k^2 - 2R\,\mathbf{u}^\top \mathbf{w} + R\,\mathbf{w}^\top \mathbf{w}. \end{aligned}$$

Since optimising cF for positive constants c constitutes identical problems, no generality is lost by assuming that $R + S = 1$. With this convention, and by expanding

$$p_k^\circ \delta_k^2 = p_k^\circ \mathbf{u}^\top \mathbf{x}_k \mathbf{x}_k^\top \mathbf{u} - 2 p_k^\circ t_k \mathbf{u}^\top \mathbf{x}_k + p_k^\circ t_k^2$$

for $k = 0, \ldots, K$, we can express F compactly as

$$F(\mathbf{u}) = \mathbf{u}^\top \mathbf{H}\mathbf{u} - 2\mathbf{u}^\top \mathbf{s} + D,$$

where

$$\begin{aligned} \mathbf{H} &= \mathbf{I}_n + \sum_{k=0}^{K} p_k \mathbf{x}_k \mathbf{x}_k^\top = \mathbf{I}_n + \mathbf{X}\mathbf{P}\mathbf{X}^\top \\ \mathbf{s} &= R\mathbf{w} + (1-R)\,\frac{t_0}{n}\,\mathbf{1}_n + \sum_{k=0}^{K} p_k t_k \mathbf{x}_k = \mathbf{w}_R + \mathbf{X}\mathbf{P}\mathbf{t}; \end{aligned} \tag{2.1}$$

$\mathbf{w}_R = R\mathbf{w} + (1-R)\,n^{-1} t_0 \mathbf{1}_n$, $p_0 = p_0^\circ - \frac{1}{n} S$, $p_k = p_k^\circ$ for $k = 1, \ldots, K$, $\mathbf{P} = \mathrm{diag}(\mathbf{p})$, where $\mathbf{p} = (p_0, p_1, \ldots, p_K)$, and $D = \mathbf{t}^\top \mathbf{P}\mathbf{t} + R\mathbf{w}^\top \mathbf{w} + (1-R)\, t_0^2 / n$ is a positive scalar not relevant to what follows. The minimum of $F(\mathbf{u})$ is attained for $\mathbf{u} = \mathbf{H}^{-1}\mathbf{s}$, and the minimum attained is $F(\mathbf{u}) = D - \mathbf{s}^\top \mathbf{H}^{-1}\mathbf{s}$. The calibration estimator of θ is $\hat{\theta}(\mathbf{u}) = \mathbf{y}^\top \mathbf{u} = \mathbf{y}^\top \mathbf{H}^{-1}\mathbf{s}$. Section 3 discusses how the values of the priorities p_k and R should be set.

2.1 Inversion of H

For large sample size n, numerical inversion of $\mathbf{H}$ might seem like a nontrivial computational hurdle. We apply a recursive algorithm that exploits the form of $\mathbf{H}$ as the sum of a matrix that is easy to invert, $\mathbf{I}$, and one, $\mathbf{XPX}^\top$, of (relatively) low rank, at most $K+1 \ll n$. We note in passing that singularity of the matrix $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}^\circ)$ raises no computational issues, although it may have some consequences on the interpretation of the priorities in $\mathbf{p}$. For example, if $\mathbf{x}_k = \mathbf{x}_l$ for some $k \neq l$, then the common variable is more appropriately associated with priority $p_k + p_l$. Also, a conflict arises when $\mathbf{x}_0 = \mathbf{1}_n$ is equal to the total of the set of columns of $\mathbf{X}^\circ$ that correspond to the indicators for a categorical variable (prior to standardisation), but the total of the corresponding targets (population counts) t_k differs from t_0. The algorithm is not affected by such a conflict.

Denote $\mathbf{H}_{-1} = \mathbf{I}_n$ and $\mathbf{H}_k = \mathbf{H}_{k-1} + p_k \mathbf{x}_k \mathbf{x}_k^\top$, $k = 0, \ldots, K$, so that $\mathbf{H} = \mathbf{H}_K$. We have the identity

$$\mathbf{H}_k^{-1} \;=\; \mathbf{H}_{k-1}^{-1} \;-\; \frac{p_k}{1 + p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1} \mathbf{x}_k}\; \mathbf{H}_{k-1}^{-1} \mathbf{x}_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1}.$$

Its validity is easy to check by evaluating the product of the expressions for $\mathbf{H}_k$ and $\mathbf{H}_k^{-1}$. In the recursive evaluation of $\mathbf{u} = \mathbf{H}^{-1}\mathbf{s}$ we do not have to form any matrices $\mathbf{H}_k$ or $\mathbf{H}_k^{-1}$ because we require only the vectors $\mathbf{h}_{kl} = \mathbf{H}_k^{-1}\mathbf{x}_l$ and $\mathbf{h}_{k,\mathrm{w}} = \mathbf{H}_k^{-1}\mathbf{w}$. For the former we have the identities

$$\mathbf{h}_{kl} \;=\; \mathbf{h}_{k-1,l} \;-\; \frac{p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,l}}{1 + p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,k}}\; \mathbf{h}_{k-1,k},$$

and for the latter the same identities, with index l replaced by w. Note that every denominator $1 + p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1} \mathbf{x}_k$ is positive. The solution,

$$\mathbf{u} \;=\; R\mathbf{h}_{K,\mathrm{w}} \;+\; (1-R)\,\frac{t_0}{n}\,\mathbf{h}_{K0} \;+\; \sum_{k=0}^{K} p_k t_k \mathbf{h}_{Kk},$$

is a linear combination of the vectors $\mathbf{h}_{Kk} = \mathbf{H}^{-1}\mathbf{x}_k$ and $\mathbf{h}_{K,\mathrm{w}} = \mathbf{H}^{-1}\mathbf{w}$. In summary, there is a unique solution $\mathbf{u}$ and it is evaluated only by operations on vectors of length n. Matrix $\mathbf{H}$ involves neither R nor $\mathbf{w}$. The vector $\mathbf{s}$ is a linear function of both R and $\mathbf{w}$, and therefore so is the solution $\mathbf{u}$. For $R = 0$, $\mathbf{u}$ does not depend on $\mathbf{w}$.

2.2 Relation to ridge calibration

In this section we show that the priorities $\mathbf{p}$ have a role similar to the ridging costs in Chambers (1996), equation (10), although function F incorporates all constraints and aims of calibration, and therefore involves additional parameters. Just like ridging can be motivated as a compromise between applying no ridging and infinite ridging, our proposal is a compromise of $\mathbf{p} = \mathbf{0}$ and infinitely large $\mathbf{p}$. Both approaches yield estimators that can be interpreted as shrinkage estimators.

In equation (2.1), $\mathbf{w}_R$ is a convex combination of $\mathbf{w}$ and t_0/n, interpreted as shrinking $\mathbf{w}$ towards its mean or expectation. Write $\mathbf{w}_R = \mathbf{X}\mathbf{v}_R + \boldsymbol{\varepsilon}_R$, where $\mathbf{v}_R = (\mathbf{X}^\top\mathbf{X})^- \mathbf{X}^\top\mathbf{w}_R$ is a projecting vector and $\boldsymbol{\varepsilon}_R$ is such that $\mathbf{X}^\top\boldsymbol{\varepsilon}_R = \mathbf{0}$. Here $(\mathbf{X}^\top\mathbf{X})^-$ is a generalised inverse of $\mathbf{X}^\top\mathbf{X}$. Its non-uniqueness is resolved below. We have the analogous decomposition $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\omega}$, where $\boldsymbol{\gamma} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{y}$, so that $\mathbf{X}^\top\boldsymbol{\omega} = \mathbf{0}$. The vectors $\mathbf{v}_R$ and $\boldsymbol{\gamma}$ are respective least-squares fits for $\mathbf{w}_R$ and $\mathbf{y}$ in terms of $\mathbf{X}$. We do not assume any aspects of validity of the implied ordinary regressions $(\mathbf{w}_R \mid \mathbf{X})$ and $(\mathbf{y} \mid \mathbf{X})$. Note that the residuals $\boldsymbol{\varepsilon}_R$ and $\boldsymbol{\omega}$ are unique, and $\boldsymbol{\varepsilon}_R^\top\mathbf{y} = \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$.

Let $\mathbf{e}_0 = (1, 0, \ldots, 0)^\top$ be the vector comprising one unity and K zeros. For a matrix $\mathbf{A}$ with $K+1$ columns, $\mathbf{A}\mathbf{e}_0$ is equal to its first column. Owing to orthogonality, $\mathbf{X}^{\circ\top}\mathbf{x}_0 = \mathbf{0}$, arranged by standardisation, we have the identity $(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{1}_n = n(\mathbf{X}^\top\mathbf{X})^-\mathbf{e}_0 = \mathbf{e}_0$. Hence

$$\mathbf{v}_R \;=\; R\left(\mathbf{X}^\top\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{w} \;+\; \frac{(1-R)\,t_0}{n}\;\mathbf{e}_0, \tag{2.2}$$

and by its substitution in (2.1),

$$\mathbf{u} \;=\; \mathbf{H}^{-1}\mathbf{s} \;=\; \left(\mathbf{I} \;+\; \mathbf{X}\mathbf{P}\mathbf{X}^\top\right)^{-1}\left\{\mathbf{X}\left(\mathbf{P}\mathbf{t} \;+\; \mathbf{v}_R\right) \;+\; \boldsymbol{\varepsilon}_R\right\}.$$

The identity $(\mathbf{I} + \mathbf{X}\mathbf{P}\mathbf{X}^\top)^{-1} = \mathbf{I} - \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ implies that

$$\begin{aligned}\mathbf{u} \;&=\; \mathbf{X}\left\{\mathbf{I} - \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\right\}\left(\mathbf{P}\mathbf{t} + \mathbf{v}_R\right) \;+\; \left\{\mathbf{I} - \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}\boldsymbol{\varepsilon}_R \\ &=\; \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t} + \mathbf{P}^{-1}\mathbf{v}_R\right) \;+\; \boldsymbol{\varepsilon}_R.\end{aligned}$$

Therefore the calibration estimator is $\hat{\theta}(\mathbf{u}) = \hat{\theta}_1 + \hat{\theta}_2(\mathbf{w}) + \hat{\theta}_3(\mathbf{w})$, where

$$\begin{aligned}\hat{\theta}_1 \;&=\; \mathbf{t}^\top\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \hat{\theta}_2(\mathbf{w}) \;&=\; \mathbf{v}_R^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \hat{\theta}_3(\mathbf{w}) \;&=\; \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}.\end{aligned} \tag{2.3}$$

Here $\hat{\theta}_1$ is an estimator of the population total θ based on the ridge regression prediction by the linear model $(\mathbf{y} \mid \mathbf{X})$. Not involving the weights $\mathbf{w}$ and based on ridging, $\hat{\theta}_1$ is unbiased only when the biases due to no weighting and ridging happen to cancel out. When $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega} = 0$, $\hat{\theta}(\mathbf{u})$ has the form of ridge calibration (Chambers, 1996; Goga and Shehzad, 2010), with $\mathbf{P}$ in the role of the (diagonal) cost matrix and $\mathbf{t}$ adjusted by $\mathbf{P}^{-1}\mathbf{v}_R$. However, the crossproducts $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top\mathbf{y}$ are evaluated without the weights $\mathbf{w}$.

As $\mathbf{P}^{-1}$ converges to the zero matrix, which corresponds to diminishing interest in the deviation of $\mathbf{u}$ from $\mathbf{w}$ as well as in the dispersion of $\mathbf{u}$, $\hat{\theta}(\mathbf{u})$ approaches $\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$ when $\mathbf{X}$ is nonsingular. This reduces to the least-squares predictor when $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega} = 0$. If $\mathbf{X}$ has deficient rank and all priorities in $\mathbf{p}$ are large, then $\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}$ has one or more small eigenvalues, and so $\hat{\theta}(\mathbf{u})$ is unstable.

As each priority in $\mathbf{p}$ converges to zero, $\hat{\theta}(\mathbf{u})$ converges to the shrinkage estimator $\mathbf{w}_R^\top \mathbf{y} = R\mathbf{w}^\top \mathbf{y} + (1-R)\, t_0\, \bar{y},$ where $\bar{y}$ is the sample (unweighted) mean of y. For $R = 0$ this is the trivial (unweighted) estimator $t_0 \bar{y}$ and for $R = 1$ it is the weighted mean $\mathbf{w}^\top \mathbf{y}$. Note that the elements of $\mathbf{p}$ (the diagonal of $\mathbf{P}$) are not set by the usual considerations of ridge regression to reduce the sampling variance in exchange for a small bias. Introducing these considerations is not straightforward because of the contribution of $\hat{\theta}_2 + \hat{\theta}_3$ to $\hat{\theta}$.

Denote by $\hat{\mathbf{y}}$ the projection vector $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{y}$ and by $\hat{\mathbf{y}}_{\mathrm{P}}$ its "ridged" version $\mathbf{X}(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Then $\hat{\theta}_3(\mathbf{w}) = R\boldsymbol{\varepsilon}_1^\top \boldsymbol{\omega} = R\mathbf{w}^\top(\mathbf{y} - \hat{\mathbf{y}})$ and, owing to (2.2),

$$\begin{aligned}
\hat{\theta}_2(\mathbf{w}) &= R\mathbf{w}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-}\mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&\quad + \frac{(1-R)\, t_0}{n}\, \mathbf{e}_0^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= R\mathbf{w}^\top\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-} - \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}\mathbf{X}^\top\mathbf{y} + \frac{1-R}{1+np_0}\,\frac{t_0}{n}\,\mathbf{y}^\top\mathbf{1}_n \\
&= R\mathbf{w}^\top\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + (1-R)\,\frac{t_0\bar{y}}{1+np_0}.
\end{aligned}$$

Hence the total $\hat{\theta}_2(\mathbf{w}) + \hat{\theta}_3(\mathbf{w})$ is

$$\hat{\theta}_{23} = R\mathbf{w}^\top\left(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + (1-R)\,\frac{t_0\bar{y}}{1+np_0}.$$

It is a linear function of both $\mathbf{w}$ and R, shrinking the weighted total of the prediction errors, $\mathbf{w}^\top(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}})$, toward a quantity that does not involve $\mathbf{X}$. Column 0 of $\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}$ is equal to $\left(1/p_0 + n\right)\mathbf{e}_0$, and so, recalling the notation $\mathbf{X}^\circ$, $\mathbf{P}^\circ$ and $\mathbf{t}^\circ$, $\hat{\theta}(\mathbf{u})$ can be expressed as

$$\mathbf{t}^{\circ\top}\left(\mathbf{P}^{\circ -1} + \mathbf{X}^{\circ\top}\mathbf{X}^\circ\right)^{-1}\mathbf{X}^{\circ\top}\mathbf{y} + R\mathbf{w}^\top\left(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + \left(1 - \frac{R}{1+np_0}\right) t_0\bar{y}.$$

It is useful to present $\hat{\theta}(\mathbf{u}; R)$ as the linear interpolation of

$$\begin{aligned}
\hat{\theta}(\mathbf{u}; R=0) &= \frac{t_0\,\bar{y}}{1+np_0} + \mathbf{t}^\top\hat{\boldsymbol{\beta}} \\
&= t_0\,\bar{y} + \mathbf{t}^{\circ\top}\hat{\boldsymbol{\beta}}^\circ \\
\hat{\theta}(\mathbf{u}; R=1) &= \mathbf{w}^\top\mathbf{y} - \boldsymbol{\delta}_{\mathrm{w}}^\top\hat{\boldsymbol{\beta}},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, $\hat{\boldsymbol{\beta}}^\circ = (\mathbf{P}^{\circ -1} + \mathbf{X}^{\circ\top}\mathbf{X}^\circ)^{-1}\mathbf{X}^{\circ\top}\mathbf{y}$, and $\boldsymbol{\delta}_{\mathrm{w}} = \mathbf{X}^\top\mathbf{w} - \mathbf{t}$ is the discrepancy vector evaluated with weights $\mathbf{w}$. The $K \times 1$ vectors $\mathbf{t}^\circ$ and $\hat{\boldsymbol{\beta}}^\circ$ are formed from $\mathbf{t}$ and $\hat{\boldsymbol{\beta}}$ by dropping their respective first elements. Thus, $\hat{\theta}(\mathbf{u}; R=0)$ can be described as the trivial estimator $t_0\,\bar{y}$ adjusted by ridge prediction at $\mathbf{t}^\circ$. For $R \geq 0$,

$$\hat{\theta}(\mathbf{u}; R=r) = \hat{\theta}(\mathbf{u}; R=0) + r\left\{\mathbf{w}^\top\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) - \frac{t_0\,\bar{y}}{1+np_0}\right\}; \tag{2.4}$$

that is, $\hat{\theta}(\mathbf{u}; R=0)$ is adjusted by R-multiple of the estimator of the population-total error in the prediction of $\mathbf{y}$ and a fraction of $t_0\,\bar{y}$. Setting the value of R matters more when the prediction errors $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}$ are large and are correlated with the base weights $\mathbf{w}$. Note that $\hat{\theta}$ is evaluated without using $(\mathbf{X}^\top\mathbf{X})^-$.

The bias of $\hat{\theta}$ is a linear function of R, so it is minimised either for $R=0$ or 1. Even though $\hat{\theta}$ does not involve $\mathbf{w}$ when $R=0$, it would have the smallest bias when the effect of the weights is estimated well by the prediction model. This is the case when the weights are constructed using only variables in $\mathbf{X}$, for which the population totals are available. We confirm this by simulations in Section 5.

In Appendix A we derive and discuss an expression for $\hat{\theta}$ in (2.4) when calibration is on a single categorical variable (and the population size). The estimator has a decomposition $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$, where $\hat{\theta}^{(1)}$ depends on $\boldsymbol{\mu}$, $\mathbf{p}$ and $\mathbf{t}$ only through p_0 and t_0, and $\hat{\theta}^{(Z)}$ does not depend on p_0. Also, $\hat{\theta}^{(Z)} = 0$ when $p_1 = \ldots = p_K = 0$. In this case, calibrating only on the population size,

$$\hat{\theta} = \hat{\theta}^{(1)} = R\left(\mathbf{w}^\top\mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,\mathbf{w}^\top\mathbf{1}_n\,\bar{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right)t_0\,\bar{y}.$$

For $R=0$, we have $\hat{\theta} = t_o\,\bar{y}$ for any p_0. For $R=1$, $\hat{\theta} = \mathbf{w}^\top\mathbf{y} + np_0/(1+np_0) \times (t_0 - \mathbf{w}^\top\mathbf{1}_n)\,\bar{y}$, which converges to the "obvious" adjustment $\mathbf{w}^\top\mathbf{y} + (t_0 - \mathbf{w}^\top\mathbf{1}_n)\,\bar{y}$ as $p_0 \to +\infty$.

3. Setting the priorities

In some applications, the constraints are equally important for all the variables, but only after taking into account the dispersions of these variables. As stated earlier, we assume that every variable in $\mathbf{X}^\circ$ is standardised. Setting the priorities $p_1, \ldots, p_K$ to a common constant is a reasonable starting point or default. For large K we may define a small number of groups of variables and assign a common priority within each group. For example, these groups of variables may be associated with distinct categorical variables. The intercept $\mathbf{x}_0 = \mathbf{1}$ has a special status among the variables in $\mathbf{X}$. High priority p_0 corresponds to the desire for small $|\delta_0|$. When $t_0 = N$ and p_0 is set sufficiently large, $\mathbf{u}^\top\mathbf{1} \doteq N$.

By construction, $0 < R < 1$, so $R = 0.5$ might be a default. A more profound approach would weigh the relative importance of small change (R) and small variance (S) of the calibration weights. With greater n, concerns about bias become dominant, so smaller S and greater $R = 1 - S$ are appropriate. Setting R to a very small value is usually unwise because then $\hat{\theta}$ depends only weakly on $\mathbf{w}$. Reducing the difference $\mathbf{u} - \mathbf{w}$ and the dispersion of the weights are both devices to avoid extreme (very large and negative) weights. These goals are given greater prominence by reducing the priorities, e.g., from $\mathbf{p}$ to $c\,\mathbf{p}$ for $0 < c < 1$.

These guidelines are admittedly rather vague and incomplete. However, the computational simplicity of the solution allows us to explore a range of plausible settings of $\mathbf{p}$ and R, especially when $\mathbf{p}$ involves only

a few distinct values. For example, when $\mathbf{X}^\circ$ is based on a single discrete variable with H categories, then $p_1 = \ldots = p_H$, and so we have only three tuning parameters, p_0, p_1 and R. When there are two discrete variables four parameters are involved. When the entire two-way table of the population margins for these two variables is available, matching the univariate margins is usually more important than matching the two-way subtotals.

Dependence of the fitted discrepancy $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$ on p_k can be explored analytically. By matrix differentiation we obtain the identity

$$\begin{aligned} \frac{\partial \delta_l^2}{\partial p_k} &= -2\delta_l \left(\mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial p_k} \mathbf{H}^{-1} \mathbf{s} - \mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{s}}{\partial p_k} \right) \\ &= -2\delta_k \delta_l \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_l, \end{aligned} \tag{3.1}$$

and its special case,

$$\frac{\partial \delta_k^2}{\partial p_k} = -2\delta_k^2 \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k < 0,$$

so that

$$\frac{\partial \log\left(\delta_k^2\right)}{\partial p_k} = -2\, \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k.$$

Hence, $|\delta_k|$ is a decreasing function of p_k, in accord with the motivation of p_k as the relative importance of reducing $|\delta_k|$. However, the decrease is slower when $|\delta_k|$ is smaller;

$$\frac{\partial^2 \log\left(\delta_k^2\right)}{\partial p_k^2} = 2\left(\mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k\right)^2,$$

so δ_k^2 is a log-convex function. Interpreted loosely, trying to wipe out a small discrepancy $|\delta_k|$ requires a sizeable change of p_k, and that could inflate some other discrepancies $|\delta_l|$.

The solution $\mathbf{u}$ is a linear function of R, which is involved only in $\mathbf{s}$;

$$\frac{\partial \mathbf{u}}{\partial R} = \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\mathbf{1}_n\right),$$

and

$$\frac{\partial \delta_k}{\partial R} = \mathbf{x}_k^\top \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\,\mathbf{1}_n\right). \tag{3.2}$$

The expression for $\mathbf{s}$ in equation (2.1), as well as the original intent, imply that R acts as a shrinkage factor, reducing the deviation of the weights $\mathbf{u}$ from $\mathbf{w}$ and $1-R$ as reducing their dispersion. Indeed, when no auxiliary information is available and $\mathbf{u}^\top \mathbf{1}_n$ is matched to $\mathbf{w}^\top \mathbf{1}_n$, $\mathbf{H} = \mathbf{I}_n$ and the solution is the convex

combination $\mathbf{w}_R$ introduced in equation (2.1). In this case, we have a simple expression, $\mathrm{E}(\mathbf{w}_R^\top \mathbf{y}) - \theta = -(1-R)\,\mathrm{cov}(w, y)$, for the bias of the calibrated estimator when the weights $\mathbf{w}$ are correct and $\mathbf{w}^\top \mathbf{y}$ is unbiased.

The bias of $\hat{\theta}(\mathbf{u})$ is likely to be largest when R is set to zero, when $\mathbf{u}$ does not depend on $\mathbf{w}$. At the same time, $\delta_0 = 0$ only when $R = 0$, unless $\mathbf{w}^\top \mathbf{1} = t_0$; see Appendix B for proof. So, focus on one goal of calibration to the detriment of others may be ill-advised.

4. Examples

In this section we demonstrate the control of the discrepancies $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_K)^\top$ and dispersion of the calibrated weights $\mathbf{u}$ by the priorities $\mathbf{p}$ and parameter R. We use a synthetic population of size $N =$ 120,000 from which we draw a sample of expected size $\mathrm{E}(n) \doteq 1{,}000$ according to a design with unequal probabilities and independent inclusions in the sample, without replacement. In the population, we have a discrete variable Z with six categories and a continuous variable X.

The six subpopulations defined by Z have sizes between 13,000 and 26,000. The values of the sole continuous background variable X are generated as random samples, one within each category k, from gamma distributions with shapes ξ_k and common rate 5.0, where $\boldsymbol{\xi} =$ (6.4, 6.7, 6.1, 6.6, 6.9, 6.4), so that the within-category means of X are in the range 1.22 − 1.28 and their standard deviations in the range 0.244 − 0.276. The outcome Y is generated as $\frac{1}{2}X + \exp(\zeta)$, where ζ comprises independent random samples from the normal distributions with means ξ_k in category k and standard deviations (0.05, 0.07, 0.04, 0.06, 0.08, 0.05). The entire setting has been selected arbitrarily, to generate a dataset in which the distributions of X and Y differ across the categories of Z, are skewed to the right, and X and Y are moderately correlated both overall (their correlation is 0.30), and within the six categories. The relevant summaries of X, Y and Z are given in Table 4.1.

Table 4.1
Population summaries of the variables X, Y and Z.

	Category (k) of Z						
	1	2	3	4	5	6	All
$t_1 \ldots t_6$	25,000	15,000	22,000	26,000	13,000	19,000	120,000
$\sum X$	31,949	20,053	26,710	34,329	17,890	24,421	155,352
$\bar{X}$ (sd)	1.28 (0.51)	1.34 (0.52)	1.21 (0.49)	1.32 (0.51)	1.38 (0.52)	1.29 (0.51)	1.29 (0.51)
$\sum Y$	117,449	92,376	79,480	146,142	96,201	89,336	620,984
$\bar{Y}$ (sd)	4.70 (0.32)	6.16 (0.47)	3.61 (0.27)	5.62 (0.39)	7.40 (0.60)	4.70 (0.33)	5.17 (1.18)

4.1 Sampling design

For drawing a sample from this population we consider a design that would be planned, and from which the base weights are calculated. In this design, the probabilities of inclusion are constant within the six categories, set to 0.0075, 0.0093, 0.0070, 0.0090, 0.0110 and 0.0075. The inclusions are mutually independent.

We inject realism into the simulated survey data collection by generating the "real" inclusion probabilities that reflect the imperfect implementation of the survey. The design probabilities are multiplied by a random sample from the log-normal distribution based on $\mathcal{N}(0, \sigma^2)$. No imperfection corresponds to $\sigma = 0$ and larger σ, with greater deviations from the base weights, to greater impact of the imperfections in the conduct of the survey. That is, $\hat{\theta}(\mathbf{w})$ is biased because the base weights $\mathbf{w}$ have been distorted in the implementation of the design, which has led to an altered (perturbed) set of weights $\mathbf{w}^\dagger$. By applying $\hat{\theta}\{C(\mathbf{w})\}$ we aim to remedy this problem. In practice, the altered weights $\mathbf{w}^\dagger$ cannot be recovered, nor estimated, but there may be some information (or well-founded opinion) about the extent of such perturbation. In our model, this perturbation is characterised by the variance σ^2 of the changes of the log-weights, or equivalently, of the log-probabilities.

In this section we use a sample drawn with $\sigma = 0.1$. This amounts to a substantial perturbation; the standard deviations of the perturbed weights in this sample are 13.3, 10.4, 14.4, 11.5, 9.7 and 12.3 within the respective categories 1, …, 6, whereas the base weights are constant within the categories. The average base weight is $\bar{w} = 122.44$. The average of the perturbed weights is $\bar{w}^\dagger = 123.22$.

4.2 Calibration

We calibrate a single sample and we are concerned solely with the discrepancies $\boldsymbol{\delta}$. We use the tuning parameter values

$$p_0 = 1.5,\ p_1 = \ldots = p_6 = 0.15,\ p_7 = 0.5,\ R = 0.9, \tag{4.1}$$

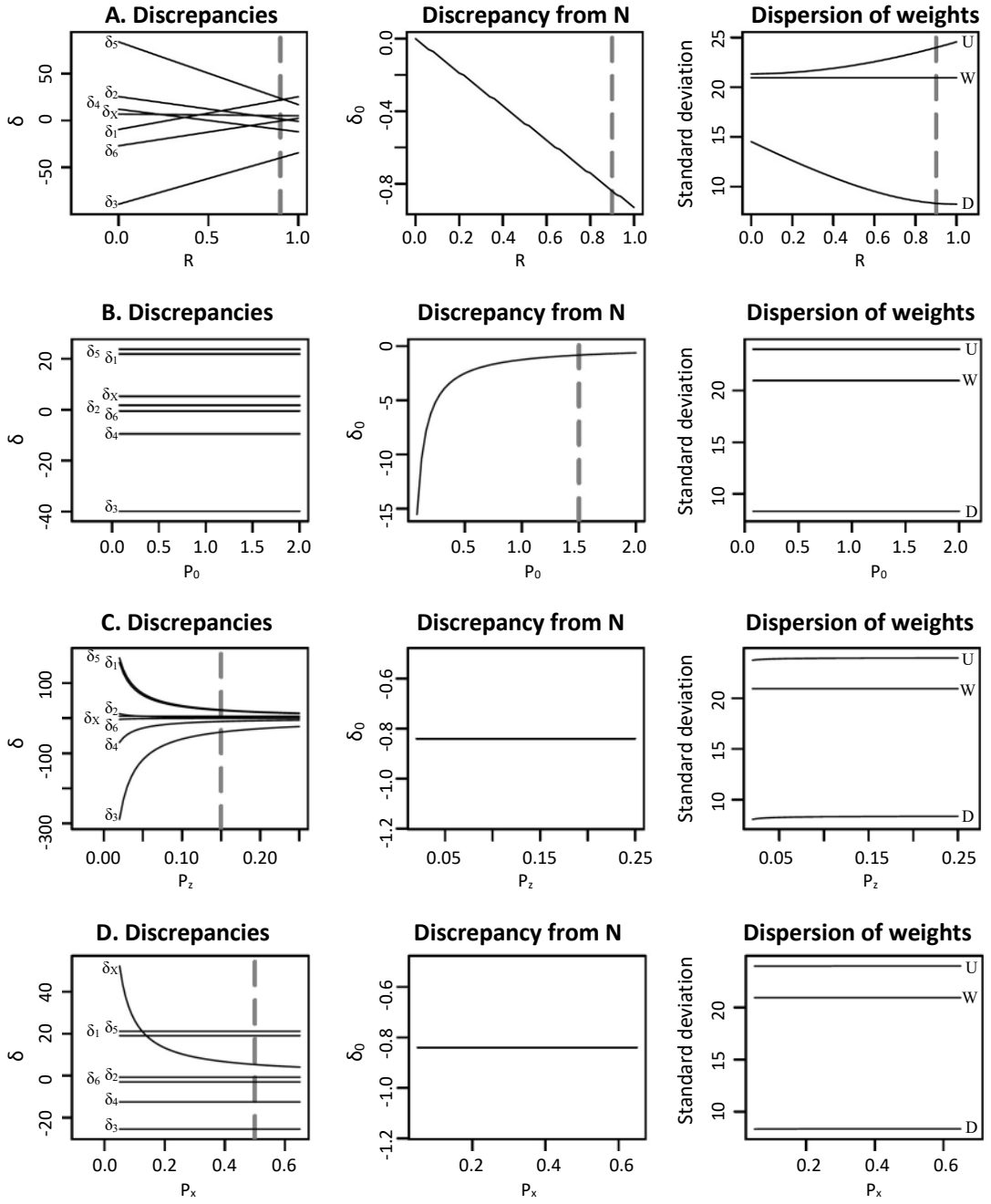
and call them the reference setting. We use the notation p_X for p_7 and δ_X for δ_7. When $p_1 = \ldots = p_6$, we denote the common value by p_Z.

We apply first calibration with values of R in the range $(0,1)$ and the reference setting for $\mathbf{p}$. The discrepancies, as functions of R, are plotted in row A at the top of Figure 4.1. In the left-hand panel, the discrepancies are plotted for X and the six categories of Z. The plot confirms that these functions $\delta_k(R)$ are linear; see equation (2.4). The discrepancies are dispersed more for $R = 0$ than for $R = 1$, although some of the functions $\delta_k(R)$ cross zero at $R \in (0,1)$.

Function $\delta_0(R)$ is plotted in a separate panel because a much narrower scale is required for the vertical axis. The match is perfect, $\delta_0 = 0$, for $R = 0$, in accord with the proof in Appendix B. The standard deviations of $\mathbf{u}$ and $\mathbf{u} - \mathbf{w}$, marked by the respective symbols U and D, are drawn in the right-hand panel, together with the standard deviation of $\mathbf{w}$ (W), which, of course, is constant. The standard deviation of $\mathbf{u}$

increases and the standard deviation of $\mathbf{u} - \mathbf{w}$ decreases with R. Based on these three plots we would choose a large value of R, to reduce the values of δ_k overall, even at a small sacrifice of δ_0. We settle on $R = 0.9$, indicated by vertical dashes.

Figure 4.1 Discrepancies δ and standard deviations of the weights as functions of the parameter $R \in (0,1)$ in row A (top), of $p_0 \in (0,2)$ in row B, of $p_Z \in (0.02, 0.25)$ in row C and of $p_X \in (0.05, 0.65)$ in row D, all adapted from the reference setting given by (4.1), indicated in the plots by long vertical dashes.



In row B, the discrepancies are plotted as functions of p_0 in the range $(0, 2)$ with the other parameters given by the reference setting. Apart from $\delta_0(p_0)$, each discrepancy and standard deviation is very close to a constant. Function $|\delta_0(p_0)|$ decreases rapidly for small values of p_0 and converges to zero very slowly for large values of p_0, as discussed following equation (3.1). Based on this plot, we set p_0 to 1.5.

In row C, the value of $p_Z = p_1 = \ldots = p_6$ is varied in the range (0.02, 0.25), and the other parameters are held at their reference values. The discrepancies $\delta_1, \ldots, \delta_6$ approach zero rapidly for small values of p_Z and converge to zero slowly for large values of p_Z. The other functions of p_Z, δ_0 and δ_X, as well as the two standard deviations in the right-hand panel, are each in a very narrow range, and their curvature is appreciable only for very small values of p_Z. We set $p_Z = 0.15$. Since $|\delta_3|$, equal to 39.9 at $p_Z = 0.15$, is larger than the discrepancies for the other five categories, we increase p_3 to 0.25. Now $\delta_3 = -25.4$, much smaller in absolute value. The second largest discrepancy is for category 5, $\delta_5 = 21.2$, reduced from 23.7 with the original setting.

The panels in row D (bottom) present the setting with p_X being varied while the other parameters are held at their reference values, except for $p_3 = 0.25$. The diagram confirms that $|\delta_X(p_X)|$ decreases rapidly for small values of p_X and more slowly as it approaches zero for large values of p_X. The other discrepancies and the two standard deviations depend on p_X very weakly. We set $p_X = 0.5$.

In summary, Figure 4.1 illustrates that we can reduce any one of the absolute discrepancies $|\delta_k|$ by increasing the corresponding priority p_k. Also, by altering R, we can trade off small dispersion of $\mathbf{u}$, essential for efficiency, for small alteration of the weights (small dispersion of $\mathbf{u} - \mathbf{w}$), indirectly related to bias. Figure 4.1 confirms that eradicating a small discrepancy requires substantial increase of the corresponding priority. It suggests that such an increase would affect the other discrepancies only slightly. Equation (3.1) and its discussion imply that this is not the case in general, especially with many variables in $\mathbf{X}$, some of them highly correlated, when several discrepancies are nontrivial.

5. Simulations

In this section we study the empirical bias and root-MSE (rMSE) of the calibration estimator $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}, \mathbf{t})\} = \mathbf{u}^\top \mathbf{y}$ for several settings ($\mathbf{p}$ and R) and levels of imperfect implementation of the sampling design, characterised by the perturbation parameter σ. We also assess the contribution a variable used in calibration makes to the reduction of rMSE. The simulations we describe involve sets of 1,000 replications. We checked that this number is sufficient by comparing the results with their re-runs using 2,000 replications for a selection of cases. We use the same population ($N = 120{,}000$), planned design with $\mathrm{E}(n) \doteq 1{,}000$ and the process of perturbing the base weights to represent the imperfect implementation of the design, as described in Section 4. The bias and rMSE are presented as functions of R, using the interpolation based on (2.4). These functions are evaluated for $\sigma = 0, 0.02, \ldots, 0.10$ and a selection of vectors $\mathbf{p}$.

Figure 5.1 presents by solid black lines the bias and the root-MSE of $\hat{\theta}$, as functions of $R \in (0, 1)$ for the reference setting of $\mathbf{p}$ and $\sigma = 0, 0.02, \ldots, 0.1$, as indicated at the right-hand margin. For each value

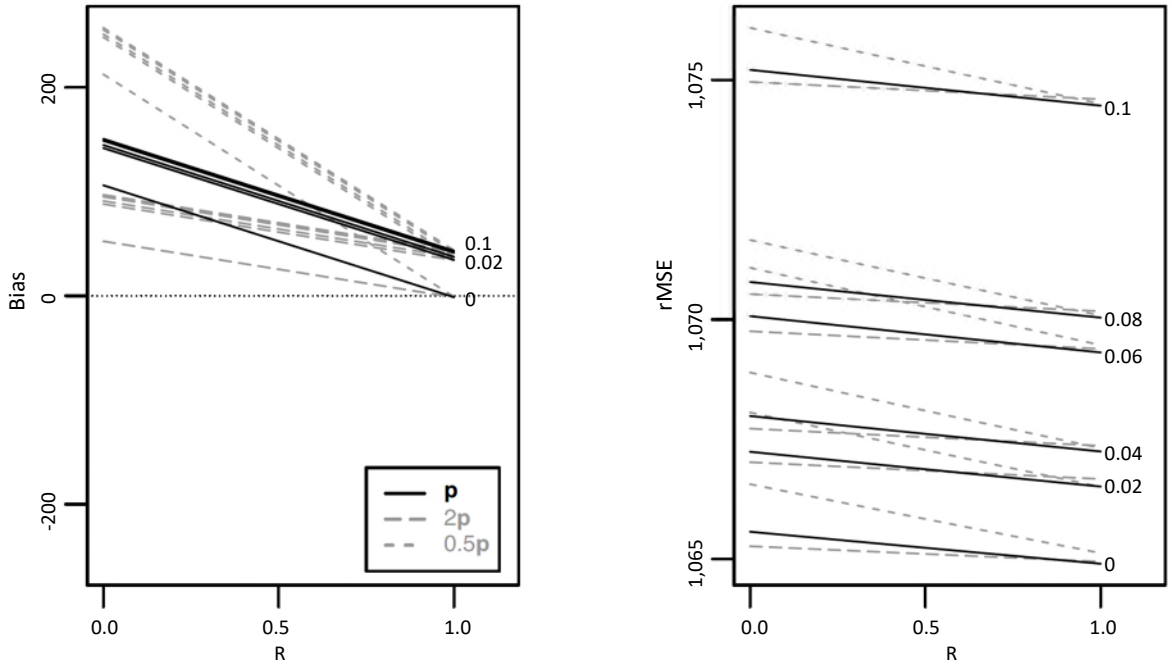
of σ, the bias is smallest for $R = 1$. The bias functions are parallel and for $\sigma \geq 0.02$ differ only slightly. The rMSE functions have very little curvature, are nearly parallel and also attain their smallest values for $R = 1$. The rMSE increases with σ for all R, although not evenly. The bias and rMSE functions for priorities $2\mathbf{p}$ and $\frac{1}{2}\mathbf{p}$ are drawn by grey lines with long and short dashes, respectively. All the functions have steeper gradients for $\frac{1}{2}\mathbf{p}$ and lower gradients for $2\mathbf{p}$, and differ from their counterparts for the reference value of $\mathbf{p}$ at $R = 1$ only slightly. Table 5.1 presents the values of rMSE at $R = 0$ and 1 in a tabular form, together with some results discussed below.

Table 5.1
Root-MSEs of the calibrated estimators of the population total of Y. The first column indicates the variables on which the design probabilities are based and, in brackets underneath, the variables used in calibration.

Design [calibration]		σ					
		0	**0.02**	**0.04**	**0.06**	**0.08**	**0.10**
		$\mathbf{p}$ (reference)					
	$R = 0$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
	$R = 1$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
		$2\mathbf{p}$					
	$R = 0$	1,065.3	1,067.0	1,067.7	1,069.8	1,070.5	1,075.0
	$R = 1$	1,064.9	1,066.7	1,067.4	1,069.4	1,070.2	1,074.6
		$\frac{1}{2}\mathbf{p}$					
	$R = 0$	1,066.6	1,068.1	1,068.9	1,071.1	1,071.7	1,076.1
	$R = 1$	1,065.1	1,066.5	1,067.3	1,069.5	1,070.1	1,074.5
Z, X		$p_3 = 0.25$					
$[Z, X]$	$R = 0$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
	$R = 1$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
		$p_0 = 15$					
	$R = 0$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
	$R = 1$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
		Ridge estimators					
	$\hat{\theta}_1$	1,082.5	1,100.4	1,103.7	1,103.1	1,103.1	1,105.4
	$\tilde{\theta}_W$	1,082.3	1,100.6	1,103.8	1,103.1	1,103.3	1,105.5
		$\mathbf{p}$ (reference)					
Z, X	$R = 0$	1,431.8	1,440.9	1,443.0	1,445.7	1,449.0	1,449.1
$[Z]$	$R = 1$	1,431.3	1,440.3	1,442.3	1,445.1	1,448.4	1,448.4
		$\mathbf{p}$ (reference)					
	$R = 0$	1,397.2	1,403.1	1,403.9	1,405.5	1,408.2	1,408.1
	$R = 1$	1,378.6	1,384.9	1,385.5	1,386.9	1,389.4	1,389.3
		$2\mathbf{p}$					
	$R = 0$	1,396.7	1,402.6	1,403.3	1,404.9	1,407.6	1,407.4
	$R = 1$	1,378.5	1,384.8	1,385.4	1,386.7	1,389.2	1,389.1
Z, X, U		$\frac{1}{2}\mathbf{p}$					
$[Z, X]$	$R = 0$	1,398.5	1,404.7	1,405.5	1,407.2	1,409.9	1,409.9
	$R = 1$	1,379.1	1,385.6	1,386.2	1,387.6	1,390.1	1,390.1
		Ridge estimators					
	$\hat{\theta}_1$	1,445.4	1,461.8	1,463.9	1,468.6	1,473.5	1,479.5
	$\tilde{\theta}_W$	1,426.6	1,446.8	1,448.7	1,452.7	1,457.8	1,463.6

Figure 5.1 shows that rMSE depends on R and the factor c in the priorities $c\mathbf{p}$ only weakly, much less than it depends on σ. For R much smaller than unity, the bias is a nontrivial fraction of the rMSE but the gradient of the bias is to a large extent ameliorated in rMSE. For example, if the sampling variance of $\hat{\theta}$ were equal to $1{,}065^2$ for all $R \in (0,1)$ and the bias decreased from 100 at $R = 0$ to zero at $R = 1$, then the rMSE would attain values 1,069.7 and 1,065.0 at $R = 0$ and 1, respectively, that is, 4.7 apart. The corresponding empirical values, 1,065.6 and 1,064.9, are only 0.7 apart.

Figure 5.1 Bias and root-MSE as functions of the parameter R for the reference setting and perturbation σ = 0, 0.02, …, 0.1, indicated at the right-hand margin; settings with the reference values of $\mathbf{p}$ (solid black lines), $2\mathbf{p}$ (long grey dashes) and $\frac{1}{2}\mathbf{p}$ (short grey dashes). Calibration on $\boldsymbol{Z}$ and X.



We compare our calibration estimators with two alternatives based on ridge regression,

$$\hat{\theta}_1 = \mathbf{t}^\top \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$\tilde{\theta}_W = \mathbf{t}^\top \left(\lambda \mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

where $\mathbf{W}$ is the diagonal matrix with $\mathbf{w}$ on its diagonal and $\lambda = n/N$. We selected $\hat{\theta}_1$ because it is a component in the decomposition of $\hat{\theta}$ in (2.3); $\tilde{\theta}_W$ is motivated by Chambers (1996).

These ridge estimators have much greater biases, in the range 300 – 500, but their rMSE's are greater than the rMSE's of our calibration estimators by much smaller margins; the rMSE's of $\hat{\theta}_1$ and $\tilde{\theta}_W$ are in the range 1,082.3 – 1,105.5 for $\sigma \in$ (0, 0.1); details are given at the bottom of the first block of Table 5.1.

For any given setting, the rMSE's of $\hat{\theta}_1$ and $\tilde{\theta}_W$ differ by less than 0.1. The weighted estimator $\mathbf{w}^\top\mathbf{y}$ and the Hájek estimator $N\mathbf{w}^\top\mathbf{y}\,/\mathbf{w}^\top\mathbf{1}$ have small biases but their rMSE's are far greater, exceeding 2,000.

By increasing p_3 from 0.15 to 0.25, as suggested in the discussion of Figure 4.1, the bias and rMSE functions are altered only slightly. The bias is reduced, by up to 0.05 but, contrary to expectation, the rMSE is increased for all values of σ, although by no more than 0.1. When p_0 is increased from 1.5 to 15, with p_3 set to 0.15, the bias and rMSE are altered imperceptibly, by less than 0.01.

In conclusion, Figure 5.1 suggests that calibration increases the efficiency of $\hat{\theta}$, and the estimator is not very sensitive to the choice of the tuning parameters, even though the discrepancies $\boldsymbol{\delta}$ are sensitive to the choice. In all settings we used, $R = 1$ is the optimal choice; that is, the criterion of small change $\|\mathbf{u} - \mathbf{w}\|$ should be ignored. It also results in small bias, much smaller than for $R = 0$. This does not apply in general. Positive S, that is, $R < 1$, is useful when the weights are much more dispersed and not highly correlated with the outcome. In small samples, variance reduction is relatively more important than bias reduction, and then S should be set to a larger positive value.

The value of a margin t_k can be quantified by calibration with the corresponding variable X_k omitted, or assigned very small priority. By dropping the margin t_X from calibration the rMSE functions are inflated substantially, to the range (1,431.8, 1,449.1) for $\sigma \in (0, 0.1)$ at $R = 0$ and are smaller by between 0.5 and 0.6 at $R = 1$. Details are given in the middle block of Table 5.1.

The sampling design and calibration setting explored so far are unrealistically congenial in that the base weights $\mathbf{w}$ depend only on the variables for which the population totals in $\mathbf{t}$ are available and therefore $\hat{\theta}_3 = 0$ in equation (2.3). We generate a more realistic setting by including a variable U correlated with the outcome variable Y, and define design probabilities that depend, apart from Z, also on U. In particular, $\mathbf{U}$ is generated as a random sample from the log-normal distribution based on $\mathcal{N}(1, 0.4)$, so that its mean is 2.94 and standard deviation is 1.23. Further, the design probabilities are set to

$$\boldsymbol{\pi}_U = \boldsymbol{\pi}_0 + \frac{0.004\,(\mathbf{U} - \bar{U})}{\max(\mathbf{U}) - \min(\mathbf{U})},$$

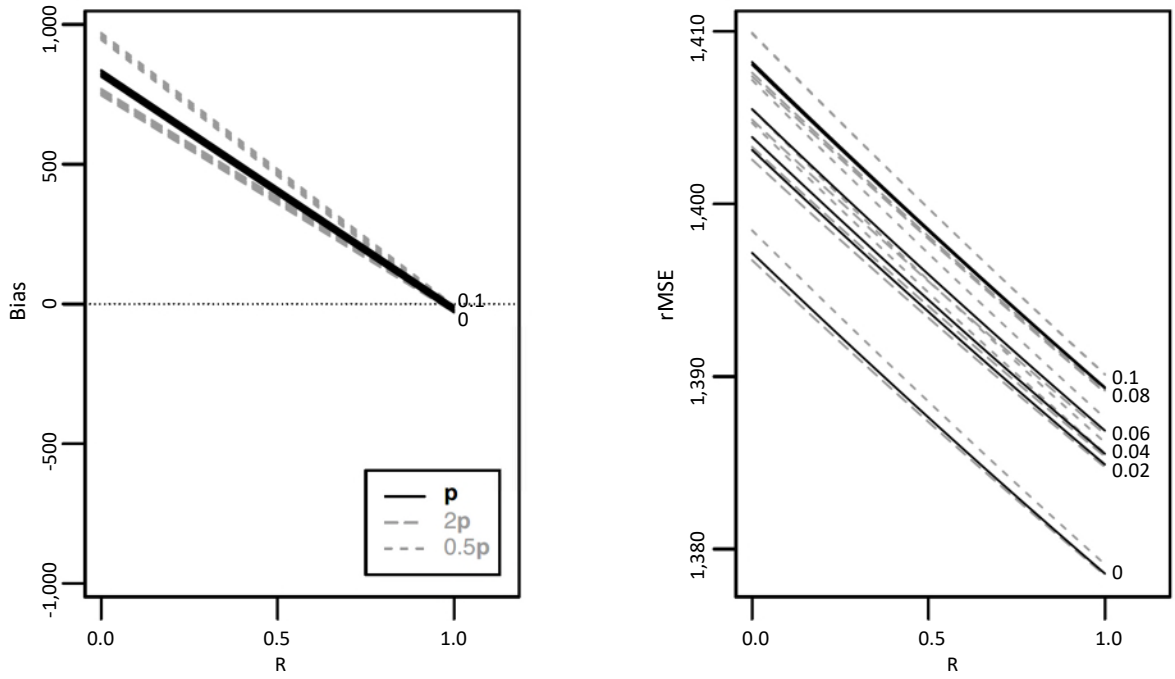
where $\boldsymbol{\pi}_0$ are the design probabilities used thus far. The probabilities $\boldsymbol{\pi}_U$ are standardised to have population total equal to 1,000. The "new" outcome variable is set to $Y + \frac{1}{5}U$. The corresponding target is $(\mathbf{Y} + \frac{1}{5}\mathbf{U})^\top\mathbf{1}_N$ = 691,516.8. The design probabilities $\boldsymbol{\pi}_U$ scaled to yield a sample of size $n \doteq$ 1,000, have standard deviations around 2.0×10^{-4} within the six categories of Z. (The average probability is $n/N = 8.33 \times 10^{-4}$).

Figure 5.2 presents the results using the same layout as Figure 5.1. It shows that the rMSE functions attain far greater values than in the setting without variable U, the rMSE functions attain their minima for $R = 1$ but the slopes of these functions are much steeper than with the original setting.

The ridge estimators $\hat{\theta}_1$ and $\tilde{\theta}_W$ are less efficient than $\hat{\theta}$, although $\tilde{\theta}_W$ is now discernibly more efficient than $\hat{\theta}_1$; their rMSE's differ by between 14 (for $\sigma = 0.1$) and 19 ($\sigma = 0$); see the bottom of the last block in Table 5.1. The bias of $\hat{\theta}_1$ is in the range (330, 370), comparable to the bias of $\hat{\theta}$ at $R = 0.6$. The bias of

$\tilde{\theta}_{\mathrm{W}}$ is much smaller, in the range $(-15, 24)$. Clearly, not involving the weights $\mathbf{w}$ is a handicap of $\hat{\theta}_{1}$ when the regression model $(y \mid X)$ is deficient.

Figure 5.2 Bias and root-MSE as functions of the parameter R for the reference setting and perturbation $\sigma = 0, 0.02, \ldots, 0.1$, indicated at the right-hand margin. Settings with the reference values of p (solid black lines), 2p (long grey dashes) and $\frac{1}{2}$p (short grey dashes). Calibration on Z, X and U.



The simulations provide evidence that our calibration estimator $\hat{\theta}$ is more efficient than the two ridge estimators $\hat{\theta}_{1}$ and $\tilde{\theta}_{\mathrm{W}}$, but the gains in efficiency are rather modest. Although the priorities $\mathbf{p}$ can be set with a view to minimising the discrepancies in $\boldsymbol{\delta}$, a careful choice is not always rewarded by greater efficiency. However, bias and efficiency are fairly insensitive to the setting of $\mathbf{p}$. We obtained very similar results using designs with fixed sample size and stratification on Z.

We have identified two factors that have a strong impact on the efficiency of $\hat{\theta}$ – the imperfection in how the sampling design is implemented, governed in the simulations by the parameter σ, and the completeness of vector $\mathbf{t}$. That is, efficiency is enhanced by good implementation of the sampling design and by calibrating on all the variables on which the sampling design is based, or by constructing the base weights using only variables with known population totals. Conclusions based on our simulations do not warrant a generalisation to settings with much more extensive calibration vector $\mathbf{t}$ and complex sampling designs.

6. Discussion and conclusion

We have introduced a method of calibration that matches the prescribed margins (population and subpopulation sizes and totals) subject to small discrepancies, while attending to the concern of efficiency. The imperative of an exact match and the dichotomy of matching or ignoring each available margin is replaced by a set of priority coefficients that quantify the importance or urgency of matching each margin, together with limiting the adjustment of the base weights and preference for less dispersed weights. The algorithm for calibrating the weights entails no iterations, nor handling any large matrices, and so it can be applied multiple times, searching for the best compromise of the competing constraints.

The calibration estimator $\hat{\theta}$ is related to a ridge regression predictor. We derived two decompositions of this estimator, one to the sum of a ridge regression predictor that does not involve the sampling weights, a bias adjustment that also has the form of ridge prediction, and a residual term that vanishes in some congenial settings. The other decomposition leads to a linear interpolation between the values of $\hat{\theta}$ for $R = 0$ and 1, both of which involve simple expressions.

Our exploration, analytically and by simulations, indicates that reduction of an absolute discrepancy $|\delta_k|$ is achieved by a small alteration of the priority p_k while the discrepancy is large, but p_k has to be increased substantially when δ_k is close to zero. This suggests that relaxed calibration applied after a careful exploration of the settings for the priority coefficients may be much more constructive than the binary choice of matching exactly or ignoring the available margin. In our experiments with relatively simple settings, we showed that the discrepancies $\boldsymbol{\delta}$ are easy to control but refined control is not necessarily rewarded by greater efficiency of the calibration estimator. However, the efficiency is quite insensitive to the detailed setting of the vector of priorities $\mathbf{p}$ and R.

We identified three factors that influence efficiency of $\hat{\theta}$ strongly. The first, the level or extent of the imperfection in how the sampling design is implemented comes as no surprise, even though calibration is meant to address this problem. It may do so to some extent, but does not compensate for it fully. The second is the availability of the population totals for the variables on which the design weights $\mathbf{w}$ are based. The third is the residual variance of the linear model relating the outcome y to the variables in $\mathbf{X}$. Of course, these conclusions have to be confirmed in more complex settings, sampling designs and information about population totals.

Our method of calibration is entirely model-free but it does not preclude adaptations that involve models. For example, the base weights $\mathbf{w}$ may be first adjusted by a model-based or model-assisted method, and the resulting weights subjected to relaxed calibration. The priorities $\mathbf{p}$ (and maybe also the coefficient R) may be set by considerations related to models, in addition to the expert judgement that we presumed originally.

The decompositions we derived involve ridge regression coefficients and predictors based on implied models. An avenue to explore in the future is whether and when it is advantageous to replace these terms using model-based considerations. Another challenge is to devise ways of incorporating known or

conjectured properties of the outcome variable(s) in the calibration. In regularly conducted surveys, they may be based on insights gained from their previous rounds. In some simple settings we have identified the correlation of the weights and outcomes as an important factor. We conjecture that these correlations are important also in more complex settings.

Adaptation of the presented method to settings with imprecise auxiliary information (West and Little, 2012; Opsomer and Erciulescu, 2021) is another outstanding challenge. The chi-squared distance for the weights used in optimisation may be replaced by alternatives, possibly at the price of a more complex (iterative) algorithm. One exception is the chi-squared distance for the log-weights (or another transformation of the weights), which may be motivated by reference to multiplicative adjustments of the weights.

We set out with no intention to posit a model for the outcome variable but derived an estimator that is related to a model. This suggests that we should reduce the emphasis on model validity in a model-based approach and focus on exploiting all the available information. The offset in the ridge regression to which the calibration estimator is related can be considered similarly, and set by the considerations of the trade-off of bias and variance. Maybe the committment to an approach or paradigm is inferior to their well-devised combination or compromise, exploiting the strengths of each and ameliorating their weaknesses.

All the computing described in this article was accomplished in R. The code developed can be obtained from the author on request.

Appendix

A. Calibration on a single categorical variable

This appendix derives an expression for the estimator $\hat{\theta}$ when calibrating only on the population size t_0 and the subpopulation sizes t_k of the categories $k = 1, \ldots, K$ of a discrete variable Z. We assume that the probability that a category does not appear in the sample is negligible. The starting point is equation (2.4), according to which we have to evaluate $\mathbf{w}^\top \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{t}^\top\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}$ and $\mathbf{X}$ is matrix of auxiliary variables after its columns $1, \ldots, K$ are standardised.

The original $n \times (K+1)$ matrix of auxiliary data, denoted by $\mathbf{X}_{\text{or}}$, comprises $\mathbf{1}_n$ in column 0 and the indicator of category k in column $k = 1, \ldots, K$. Denote by $\boldsymbol{\mu} = (0, \mu_1, \ldots, \mu_K)^\top$ the vector of sample proportions of the categories of Z, preceded by a zero, $s_0^2 = 1$ and $s_k^2 = \mu_k(1-\mu_k)$ for $k = 1, \ldots, K$, and $\mathbf{S} = \text{diag}(1, s_1, \ldots, s_K)$. Then standardisation is the transformation $\mathbf{X} = (\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\,\mathbf{S}^{-1}$; it leaves column 0 intact. Since the other columns of $\mathbf{X}_{\text{or}}$ are pairwise orthogonal indicators, $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is expressed in terms of $\mathbf{X}_{\text{or}}$, $\boldsymbol{\mu}$ and $\mathbf{S}$ as

$$\hat{\boldsymbol{\beta}} = \mathbf{S}\left\{\mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\left(\boldsymbol{\mu}^{(1)}\right) - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right\}^{-1}\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\mathbf{y},$$

where $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu} + \mathbf{e}_0$; element 0 of $\boldsymbol{\mu}^{(1)}$ is equal to 1 and its other elements are μ_k. Denote $\boldsymbol{\Pi} = \mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\,(\boldsymbol{\mu}^{(1)})$; its diagonal elements are $1/p_0 + n$ and $s_k^2/p_k + n\mu_k$, $k = 1, \ldots, K$. The inverse involved in $\hat{\boldsymbol{\beta}}$ is

$$\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1} = \boldsymbol{\Pi}^{-1} + \frac{n}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}$$

where, since $\boldsymbol{\mu}^\top\mathbf{1} = 1$,

$$\begin{aligned} G &= 1 - n\,\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \sum_{k=1}^{K}\left(\mu_k - \frac{n\,p_k\mu_k^2}{\mu_k(1-\mu_k) + n\,p_k\mu_k}\right) = \sum_{k=1}^{K} g_k s_k^2, \end{aligned}$$

and $g_k = 1/(1 - \mu_k + n\,p_k)$, $k = 1, \ldots, K$.

Denote $\bar{\mathbf{Y}} = (\bar{y}, \bar{y}_1\,\mu_1, \ldots, \bar{y}_K\mu_K)^\top$, where $\bar{y}$ is the sample mean of y and $\bar{y}_k$ the subsample mean of y in category k of Z. Define $\bar{\mathbf{W}} = (\bar{w}, \bar{w}_1\,\mu_1, \ldots, \bar{w}_K\mu_K)^\top$ similarly. We have the identities $(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu} = \frac{1}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}$ and $\mathbf{a}^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} = \sum_{k=1}^{K} a_k p_k g_k$ for any vector $\mathbf{a} = (a_0, a_1, \ldots, a_K)^\top$. Therefore

$$\begin{aligned} \mathbf{w}^\top\mathbf{X}\hat{\boldsymbol{\beta}} &= n\,\mathbf{w}^\top\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &= n^2\left(\bar{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &\quad + \frac{n^3}{G}\left(\bar{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \times \left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \frac{n^2 p_0}{1 + n\,p_0}\bar{w}\,\bar{y} + n^2\sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{w}_k - \bar{w}\right)\left(\bar{y}_k - \bar{y}\right) \\ &\quad + \frac{n^3}{G}\sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{w}_k - \bar{w}\right) \times \sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{y}_k - \bar{y}\right). \end{aligned}$$

(The sign $\times$ is added to emphasise that the multiplication is applied to two scalars.) The same sequence of operations yields the identity

$$\begin{aligned} \mathbf{t}^\top\hat{\boldsymbol{\beta}} &= n\left(\mathbf{t}_{\text{or}} - t_0\boldsymbol{\mu}\right)^\top\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &= \frac{n\,p_0}{1 + n\,p_0}t_0\,\bar{y} + n t_0\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right)\left(\bar{y}_k - \bar{y}\right) \\ &\quad + \frac{n^2 t_0}{G}\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right) \times \sum_{k=1}^{K} g_k p_k\left(\bar{y}_k - \bar{y}\right). \end{aligned}$$

Hence, after substituting to (2.4), $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$, where

$$\begin{aligned}\hat{\theta}^{(1)} &= R\left(\mathbf{w}^\top\mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,n\,\overline{w}\,\overline{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right) t_0\,\overline{y} \\ \hat{\theta}^{(Z)} &= n\sum_{k=1}^{K} g_k p_k \Lambda_k \left(\overline{y}_k - \overline{y}\right) + \frac{n^2}{G}\sum_{k=1}^{K} g_k p_k \Lambda_k \times \sum_{k=1}^{K} g_k p_k \left(\overline{y}_k - \overline{y}\right)\end{aligned} \tag{6.1}$$

and $\Lambda_k = t_k - n\,R\,\overline{w}_k - \mu_k(t_0 - n\,R\,\overline{w}) = t_k - t_0\mu_k - nR(\overline{w}_k - \overline{w}\mu_k)$. Note that $\hat{\theta}^{(1)}$ depends on $\boldsymbol{\mu}$, $\mathbf{p}$ and $\mathbf{t}$ only through p_0 and t_0, whereas $\hat{\theta}^{(Z)}$ does not depend on p_0. In fact, $\hat{\theta}^{(Z)} = 0$ when calibrating only on the population size, when $p_1 = \ldots = p_K = 0$. We have $\hat{\theta}^{(Z)} = 0$ also when $\Lambda_1 = \ldots = \Lambda_K = 0$. An example of such a reduction arises when the weights are constant, $\mathbf{w} = \overline{w}\,\mathbf{1}_n$ and the subsample sizes within the K categories are fixed, $t_k = t_0\,\mu_k$. In general, $\hat{\theta}^{(Z)}$ can be regarded as an adjustment of $\hat{\theta}^{(1)}$ for the priorities associated with Z.

Estimator $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$ depends on $\mathbf{w}$ and $\mathbf{y}$ only through $\mathbf{w}^\top\mathbf{y}$ and the K pairs of within-category means $\overline{w}_k$ and $\overline{y}_k$. The product $g_k p_k = p_k/(1-\mu_k + np_k)$ is an increasing function of p_k with its respective limits of 0 at $p_k = 0$ and $1/n$ as $p_k \to +\infty$. The denominator G is a decreasing function of each p_k. It attains its maximum of 1 for $p_1 = \ldots = p_K = 0$ and its limit as all K priorities p_k diverge to $+\infty$ is zero. In the latter case $\hat{\theta}$ becomes unstable. However, in practice p_0 is set higher than the other priorities, so this case is of little practical relevance.

As p_0 diverges to $+\infty$, $\hat{\theta}^{(1)}$ converges to

$$R\,(\mathbf{w}^\top\mathbf{y} - n\,\overline{w}\,\overline{y}) + t_0\,\overline{y} = (n-1)\,R\,\mathrm{cov}\,(w, y) + t_0\,\overline{y}\,.$$

This confirms that the weights are especially important when they are highly correlated with the outcome variable.

B. $R = 0$ implies $\delta_0 = 0$

In the decomposition $\mathbf{w}_R = \mathbf{X}\mathbf{v}_R + \boldsymbol{\varepsilon}_R$ for $\mathbf{w}_0$, we have $\mathbf{v}_0 = \mathbf{e}_0$ and $\boldsymbol{\varepsilon}_R = \mathbf{0}$. Hence

$$\mathbf{u} = \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t} + \frac{t_0}{n}\,\mathbf{P}^{-1}\mathbf{e}_0\right).$$

Owing to standardisation, $\mathbf{X}^\top\mathbf{1}_n = n\,\mathbf{e}_0$, and column 0 of $\mathbf{X}^\top\mathbf{X}$ is also equal to $n\,\mathbf{e}_0$. Further, $\mathbf{P}^{-1}\mathbf{e}_0 = p_0^{-1}\mathbf{e}_0$, and so

$$\begin{aligned}\mathbf{u}^\top\mathbf{1} &= \left(\mathbf{t} + \frac{t_0}{np_0}\,\mathbf{e}_0\right)^\top\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{1} \\ &= \frac{n}{p_0^{-1} + n}\left(\mathbf{t} + \frac{t_0}{np_0}\,\mathbf{e}_0\right)^\top\mathbf{e}_0 = \mathbf{t}_0.\end{aligned}$$

Therefore $\delta_0 = \mathbf{u}^\top \mathbf{1} - t_0 = 0$. We explore when the slope of δ_0, a linear function of R, vanishes. According to equation (3.2), this slope is

$$\begin{aligned}
\frac{\partial \delta_0}{\partial R} &= \mathbf{e}_0^\top \mathbf{X}^\top \left\{ \mathbf{I}_n - \mathbf{X} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \right\} \left(\mathbf{w} - \frac{t_0}{n} \mathbf{1}_n \right) \\
&= \mathbf{e}_0^\top \left\{ \mathbf{I} - \mathbf{X}^\top \mathbf{X} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \right\} \mathbf{X}^\top \left(\mathbf{w} - \frac{t_0}{n} \mathbf{1}_n \right) \\
&= \mathbf{e}_0^\top \mathbf{P}^{-1} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \left(\mathbf{X}^\top \mathbf{w} - \frac{t_0}{n} \mathbf{X}^\top \mathbf{1}_n \right) \\
&= \frac{1}{p_0} \frac{1}{p_0^{-1} + n} \mathbf{e}_0^\top \left(\mathbf{X}^\top \mathbf{w} - t_0 \mathbf{e}_0 \right) \\
&= \frac{1}{1 + n p_0} \left(\mathbf{w}^\top \mathbf{1} - t_0 \right).
\end{aligned}$$

Hence $\delta_0 = 0$ when $R = 0$ or $\mathbf{w}^\top \mathbf{1} = t_0$, as stated at the end of Section 3. Note that in most surveys with complex sampling design and nontrivial (random) sample size the probability that $\mathbf{w}^\top \mathbf{1} = t_0$ is very small or zero.

References

Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 5-20.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 323-353.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Dagdoug, M., Goga, C. and Haziza, D. (2023). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*, 50, 761-785.

Davies, G. (2018). *Examination of Approaches to Calibration in Survey Sampling.* PhD. thesis, Cardiff University, UK.

Devaud, D., and Tillé, Y. (2019). Deville and Särndal's calibration: Revisiting a 25-years-old successful optimization problem. *Test*, 28, 1033-1065.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association*, 87, 1013-1020.

Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.

Goga, C., and Shehzad, M.A. (2010). Overview of ridge estimators in survey sampling. Université de Bourgogne, Dijon, France.

Guggemos, F., and Tillé, Y. (2010). Penalized calibration in survey sampleing: Design based estimation assisted by mixed model. *Journal of Statistical Planning and Inference*, 140, 3199-3212.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

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

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

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

Longford, N.T. (2024). Statistical balancing as an unconstrained optimisation problem. Submitted.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Opsomer, J.D., and Erciulescu, A.L. (2021). Replication variance estimation after sample-based calibration. *Survey Methodology*, 47, 2, 265-277. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00006-eng.pdf.

Rao, J.N.K., and Singh, A.C. (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics*, 25, 371-383.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 2, 99-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488-eng.pdf.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: John Wiley & Sons, Inc.

Vera, J.F., Sánchez Zuleta, C.C. and Rueda, M. (2023). A unified approach based on multidimensional scaling for calibration estimation in surveys with qualitative auxiliary information. *Statistical Methods in Medical Research*, 32, 760-772.

West, B.T., and Little, R.J.A. (2012). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society*, Series A, 176, 211-225.

Wu, C., and Lu, W.W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, 84, 21-39.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 287-301
Statistics Canada, Catalogue No. 12-001-X

A hierarchical gamma prior for modeling random effects in small area estimation

Xueying Tang and Liangliang Zhang[1]

Abstract

Small area estimation (SAE) is becoming increasingly popular among survey statisticians. Since the direct estimates of small areas usually have large standard errors, model-based approaches are often adopted to borrow strength across areas. SAE models often use covariates to link different areas and random effects to account for the additional variation. Recent studies showed that random effects are not necessary for all areas, so global-local (GL) shrinkage priors have been introduced to effectively model the sparsity in random effects. The GL priors vary in tail behavior, and their performance differs under different sparsity levels of random effects. As a result, one needs to fit the model with different choices of priors and then select the most appropriate one based on the deviance information criterion or other evaluation metrics. In this paper, we propose a flexible prior for modeling random effects in SAE. The hyperparameters of the prior determine the tail behavior and can be estimated in a fully Bayesian framework. Therefore, the resulting model is adaptive to the sparsity level of random effects without repetitive fitting. We demonstrate the performance of the proposed prior via simulations and real applications.

Key Words: Adaptive shrinkage; Fay-Herriot model; Global-local priors; Normal-gamma priors.

1. Introduction

Small area estimation (SAE) aims at producing reliable estimates of crucial statistics at a finer geographic level or for a small subpopulation. The results often provide important information for public policy design and resource allocation. An example of SAE is the Small Area Income and Poverty Estimation (SAIPE) program conducted by the United States Census Bureau. The goal of the program is to provide estimates related to income and poverty at various levels including counties and school districts based on data collected from the American Community Survey. Small areas and small subgroups are often associated with small sample sizes in a survey, making direct estimates possess large standard errors and coefficients of variation. Therefore, model-based approaches are often used to produce estimates with desirable precision by borrowing strength across small areas. The models for SAE are often classified into unit-level models and area-level models. The focus of this article is the latter class and we refer readers to Pfeffermann (2013), Rao and Molina (2015), and Sugasawa and Kubokawa (2020) for detailed reviews of small area estimation models.

In area-level models, the direct estimate of each small area is often written as the sum of the small area mean and sampling error. The sampling errors are often assumed to be independent normal random variables with mean zero and known variances. The small area means are further decomposed into fixed effects and area-specific random effects. The fixed effect part uses auxiliary information from administrative records and population census as covariates to link different small areas while the random effects characterize the

1. Xueying Tang, Department of Mathematics, University of Arizona, 617 N. Santa Rita Avenue, Tucson, AZ 85721, USA. E-mail: xytang@arizona.edu; Liangliang Zhang, Department of Population and Quantitative Health Sciences, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA. E-mail: lxz716@case.edu.

variation of small area means that is not captured by the fixed effects. The most famous area level model in the SAE literature is the Fay-Herriot (FH) model (Fay and Herriot, 1979) where the random effects are assumed to be independent normals with mean zero and a common unknown variance. These assumptions are convenient for theoretical analysis and application in practice, making the FH model one of the most popular models for SAE.

Despite the convenience, the assumptions of the FH model have been questioned as they are often violated in practice. Various extensions have been made to relax the assumptions and further improve the model performance. Examples of these efforts include Datta and Lahiri (1995), Li and Lahiri (2007), Ybarra and Lohr (2008), Fabrizi and Trivisano (2010), and Porter, Wikle, and Holan (2015). A question that has been raised recently is whether the inclusion of random effects for all areas is necessary. The exploration starts with Datta, Hall and Mandal (2011) where a hypothesis testing procedure was designed to determine whether eliminating the random effects is appropriate. The null hypothesis is that the variance of the random effects equals zero. The test is based on the goodness-of-fit of the fixed effect model and works fine if there exists a small or moderate number of small areas. However, when the number of small areas is large, the null hypothesis is often rejected due to the large discrepancy between the fixed effect and the direct estimate in a few areas. Based on these observations, Datta and Mandal (2015) proposed to use spike-and-slab priors for modeling random effects. Under this model, the distribution of the random effects is assumed to be a mixture of a point mass at zero (the spike part) and a zero-mean normal distribution (the slab part). The spike part enables the removal of the random effects in areas where it is appropriate, and the slab part characterizes the non-zero random effects. This idea is further extended by Chakraborty, Datta and Mandal (2016) where a mixture of two normal distributions with different variances is used for modeling random effects.

More recently, Tang, Ghosh, Ha and Sedransk (2018) proposed to use global-local (GL) shrinkage priors for describing the random effects with various sparsity structures. The model still assumes the random effects follow independent zero-mean normal distributions, but the variances are area-specific. Each variance is expressed as a product of an area-specific local parameter and a global parameter shared across small areas. A small global parameter tends to shrink all direct estimates to the synthetic estimator to account for the small or close-to-zero random effects while a large local parameter compensates for the shrinkage for areas that need a large random effect. Possible choices of the priors for the local parameters include a wide range of heavy-tailed distributions such as Laplace priors (Park and Casella, 2008), horseshoe priors (Carvalho, Polson and Scott, 2009), and three-parameter-beta priors (Armagan, Clyde and Dunson, 2011). The flexibility in the local priors enables the GL model to characterize random effects in various settings. Tang et al. (2018) showed that the performance of the GL model is often better than that of the spike-and-slab model.

The outstanding performance of the GL model relies on choosing the priors for the local parameters appropriately. The priors are often classified into exponential-tailed priors and polynomial-tailed priors in theoretical analysis. It has been shown that polynomial-tailed priors are more appropriate when random

effects are minimal in the majority of the areas, and exponential-tailed priors are more suitable otherwise. Using an inappropriate prior may lead to undesirable performance such as low estimation accuracy or low coverage rate of the credible intervals. Since the underlying structure of random effects is unknown for a given dataset, data-driven methods for prior selection are crucial for applying the GL model in practice. Tang et al. (2018) used the deviance information criterion (DIC; Spiegelhalter, Best, Carlin and Van Der Linde, 2002) for this purpose. This method can often select a reasonable prior. However, calculating the DICs and other model selection criteria requires fitting models multiple times, each time with a different choice of local priors. This requires substantial computing resources especially when the number of small areas is large. In addition, uncertainty quantification based solely on the selected model may underestimate the variation of the estimates since it does not take into account the variation brought by different models.

In this paper, we propose a new model for the random effects of the area-level model. The model is adaptive to various sparsity levels and structures of random effects as the GL model while not requiring repeated fitting for prior selection. Similar to the GL model, we assume the normally distributed random effects with area-specific variances. A gamma prior is then placed on the variances. With different choices of the shape and rate parameters, the prior can have an exponential tail or (almost) polynomial tail, accommodating various sparsity levels and structures of the random effects in a way similar to the GL model. Since the tail behavior of the prior is indexed by the hyperparameters, the problem of selecting the most appropriate prior becomes the problem of estimating the hyperparameters. We further place hyper-priors on the shape and rate parameters of the gamma distribution to estimate the hyperparameters in a fully Bayesian framework. In this way, fitting models with different priors can be avoided and the variation brought by different models can also be taken into account.

The rest of the paper is organized as follows. In Section 2, we describe the hierarchical gamma model for achieving adaptive shrinkage in small area estimation and its connection with some existing models. An MCMC algorithm for drawing posterior samples is described in Section 3. The performance of the proposed model is demonstrated through simulation studies in Section 4 and applications to two real datasets in Section 5. We conclude with final remarks in Section 6.

2. Model

Let $y_1,\ldots,y_n$ denote the direct estimates of small area means $\theta_1,\ldots,\theta_m$ of m small areas. We assume

$$y_i = \theta_i + \varepsilon_i, \text{ and } \theta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \text{ for } i = 1,\ldots,m, \tag{2.1}$$

where $\varepsilon_1,\ldots,\varepsilon_m$ are independent sampling errors with $\varepsilon_i \sim N(0, D_i)$ and D_i being a known error variance, $\mathbf{x}_i$ is a p-dimensional vector of auxiliary variables, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ is the corresponding coefficient vector, and $u_1,\ldots,u_m$ are independent area-specific random effects. The random effects characterize the variation in θ_i that cannot be explained by the auxiliary variables. Throughout the paper, we use the

following compact notation to denote the model components: $\mathbf{y} = (y_1,\ldots,y_m)^\top$, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)^\top$, $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_m)^\top$, $\mathbf{u} = (u_1,\ldots,u_m)^\top$, and $\mathbf{D} = \text{diag}\{D_1,\ldots,D_m\}$. The notation $N(\mu,\sigma^2)$ represents the normal distribution with mean μ and variance σ^2.

We assume the random effects follow

$$\begin{aligned} u_i \mid \sigma_i^2 &\sim \text{N}(0,\sigma_i^2) \\ \sigma_i^2 \mid a,b &\sim \text{Gamma}(a,b) \end{aligned} \tag{2.2}$$

where $0 < a < 1$ and $b > 0$ are two hyperparameters and $\text{Gamma}(a,b)$ denote the gamma distribution with shape parameter a and rate parameter b. The probability density function of $\text{Gamma}(a,b)$ is $\pi(x \mid a,b) = (b^a / \Gamma(a))\, x^{a-1} \exp(-bx)$. Unlike the classic Fay-Herriot (FH) model where the random effects follow independent normal distributions with a common variance σ^2 as

$$u_i \mid \sigma^2 \sim N(0,\sigma^2), \tag{2.3}$$

we assign distinct variances for random effects of different areas and further place a gamma distribution on the variance parameters. Marginally, the scale mixture structure in (2.2) assumes a heavy tail distribution on u_i. Also, we constrain hyperparameter a to the interval $(0,1)$ so that the marginal distribution of u_i has a significant amount of probability mass around zero. These features of u_i allow the model to capture the high variation in the random effects which typically occurs when the number of small areas is large.

The scale mixture structure is also used in the Global-Local (GL) model proposed by Tang et al. (2018). There

$$\begin{aligned} u_i \mid \lambda_i^2, \tau^2 &\sim N(0, \lambda_i^2 \tau^2), \\ \lambda_i^2 \sim \pi_L(\lambda_i^2), \quad \tau^2 &\sim \pi_G(\tau^2), \end{aligned} \tag{2.4}$$

where λ_i^2 and τ^2 are called the local and the global parameters, respectively, and π_L and π_G denote their respective priors. Our model is similar to the GL model in the sense that both models assume area-specific variances of random effects and place hyperpriors on the variance parameters. Although we do not explicitly include a global parameter in (2.4), $1/b$ is a scale parameter of λ_i and thus u_i, playing the role of the global parameter. Our model can be rewritten as a GL model with $\lambda_i^2 \sim \text{Gamma}(a,1)$, which is one of the choices of π_L considered in Tang et al. (2018). However, a is treated as a hyperparameter to be estimated in our model while a known constant in the GL model. In our model, the tail heaviness of $\pi(\sigma_i^2)$ varies with the values of a and b. If a is close to one, the exponential component $\exp(-b\sigma_i^2)$ in the gamma density dominates. If b is close to zero, then the polynomial term $(\sigma_i^2)^{a-1}$ dominates. In the GL model, the authors of Tang et al. (2018) divided the choices of π_L into polynomial-tailed priors and exponential-tailed priors. They showed the two groups of priors have their own best-performing scenarios in terms of estimating small area means. The polynomial-tailed priors perform better when only a few areas need random effects while the exponential-tailed priors are more suitable when more areas need random effects. By considering a gamma prior on σ_i^2 with varying a and b, our model unifies the polynomial-tailed priors

and the exponential-tailed priors for λ_i^2 in the GL model. The problem of choosing π_L in the GL model becomes the problem of estimating a and b in the proposed model.

To estimate a and b, we consider a fully Bayesian framework and further place hyperpriors on a and b. More specifically, we consider

$$a \sim \text{Uniform}(0,1), \quad b \sim \text{Gamma}(s_0, t_0), \tag{2.5}$$

where s_0 and t_0 are set to small positive values. Since $1/b$ is analogous to the global parameter, the gamma prior on b is similar to a weakly informative inverse gamma prior on τ^2 in the GL model. Although other choices for the hyperpriors are possible, we choose the gamma prior for convenience as it is conditional conjugate. Following Tang et al. (2018), we set $s_0 = t_0 = 10^{-10}$. Because of the hierarchical structure (2.2) and (2.5), we call our model hierarchical gamma (HG) model. This prior is closely related to the normal-gamma prior in the context of variable selection (Griffin and Brown, 2005, 2010).

The model specification completes with a prior on $\boldsymbol{\beta}$. Following the literature in Bayesian small area estimation, we consider a flat prior

$$\pi(\boldsymbol{\beta}) \propto 1. \tag{2.6}$$

Although this prior is improper, one can easily show that the resulting posterior distribution is proper under minor regularity conditions.

Theorem 1. *The posterior distribution of the model specified by (2.1), (2.2), (2.5), and (2.6) is proper if* $\text{rank}(\mathbf{X}) = p < m$.

The proof of the theorem is included in the appendix.

3. Computation

Under the proposed HG model (2.1) (2.2), (2.5), and (2.6), the posterior density of $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)^\top$, a, and b is

$$\begin{aligned} \pi(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2, a, b \mid \mathbf{y}) \quad &\propto \exp\left[-\frac{1}{2} (\mathbf{y} - \mathbf{u} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{D}^{-1} (\mathbf{y} - \mathbf{u} - \mathbf{X}\boldsymbol{\beta}) \right] \\ &\times \prod_{i=1}^{m} \left\{ (\sigma_i^2)^{-1/2} \exp\left(-\frac{u_i^2}{2\sigma_i^2} \right) \frac{b^a}{\Gamma(a)} (\sigma_i^2)^{a-1} \exp(-b\sigma_i^2) \right\} \\ &\times b^{s_0 - 1} \exp(-t_0 b). \end{aligned} \tag{3.1}$$

We use Gibbs sampling (Gelfand and Smith, 1990) to draw samples from the posterior distribution. To this end, it is easy to find the full conditionals as

- $\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}, \boldsymbol{\sigma}^2, a, b \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, where $\boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta \mathbf{X}^\top \mathbf{D}^{-1} \mathbf{w}$, $\boldsymbol{\Sigma}_\beta = (\mathbf{X}^\top \mathbf{D}^{-1} \mathbf{X})^{-1}$, and $\mathbf{w} = \mathbf{y} - \mathbf{u}$;

- $\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, a, b \sim N(\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\mathbf{u}})$, where $\boldsymbol{\mu}_{\mathbf{u}} = (\mathbf{I} - \mathbf{B})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $\boldsymbol{\Sigma}_{\mathbf{u}} = (\mathbf{I} - \mathbf{B})\mathbf{D}$, $\mathbf{I}_m$ is the m -dimensional identity matrix, and $\mathbf{B} = \text{diag}(B_1, \ldots, B_m)$ with $B_i = D_i / (D_i + \sigma_i^2)$;
- $\pi(\boldsymbol{\sigma}^2 \mid \mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, a, b) \propto \prod_{i=1}^{m} (\sigma_i^2)^{a-3/2} \exp\left[-(u_i^2 / 2\sigma_i^2) - b\sigma_i^2\right]$;
- $b \mid \mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2, a \sim \text{Gamma}\,(am + s_0, \sum_{i=1}^{m} \sigma_i^2 + t_0)$;
- $\pi(a \mid \mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2, a) \propto \frac{b^{ma}}{\Gamma(a)^m} \left(\prod_{i=1}^{m} \sigma_i^2 \right)^{a-1}$.

Hence, sampling $\boldsymbol{\beta}$, $\mathbf{u}$, and b from their respective full conditional distribution is straightforward. For $\boldsymbol{\sigma}^2$, its elements can be sampled independently. Noticing that the probability density function of a generalized inverse Gaussian distribution (denoted by $\text{GIG}(\eta, \chi, \psi)$) is $f(x \mid \eta, \chi, \psi) \propto x^{\eta-1} \exp[-\chi / (2x) - \psi x / 2]$, we can sample σ_i^2 from $\text{GIG}(a - 1/2, u_i^2, 2b)$. The full conditional of a is not a common distribution. We use a slice sampling (Neal, 2003) step to draw samples of a within the Gibbs sampler.

4. Simulations

4.1 Settings

In this section, we investigate the performance of the proposed model on simulated datasets. The datasets are generated from model (2.1). We consider three choices of the number of small areas $m =$ 100, 500, 1,000. For each area, the covariate vector consists of one and an element randomly generated from $N(10, 2)$. The covariate coefficient vector is fixed at $\boldsymbol{\beta} = (20, 1)^\top$. The candidate values of the error variance D_i are 0.5, 1, 1.5, …, 5. Each value is assigned to the same number of areas in each dataset. We consider five settings for generating the random effects u_i's:

(i) Normal: $u_i \sim N(0, 4)$,

(ii) Mixture 0.2: $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0.2)$,

(iii) Mixture 0.5: $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0.5)$,

(iv) Mixture 0.8: $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0.8)$,

(v) Student's T: $u_i \sim t_3$,

where $\text{Ber}(p)$ denotes the Bernoulli distribution with mean p. For ease of comparison, random effects generated from settings (ii)-(v) are rescaled to have the same standard deviation as those generated from setting (i).

We generate 100 datasets for each combination of m and the setting for u_i. The proposed HG model is fitted to each dataset. Posterior samples are obtained using the Gibbs sampler described in Section 3. The sampler is run for 20,000 iterations with the first half discarded as burn-in. The small area means θ_i's are estimated by the corresponding posterior sample means. For comparison, we also use the FH model and the GL model to estimate the small area means. For the GL model, two choices of the priors of the local

parameters λ_i^2's are considered: the horseshoe (HS) prior $\pi_L(\lambda_i^2) \propto (\lambda_i^2)^{-1/2}(1+\lambda_i^2)^{-1}$ (Carvalho, Polson and Scott, 2010) and the Laplace (LA) prior $\pi_L(\lambda_i^2) = \exp(-\lambda_i^2)$. They are chosen as the representatives of the polynomial-tailed and exponential-tailed priors, respectively.

We use Average Absolute Deviance (AAD) and Average Squared Deviance (ASD) to quantify the difference between the estimated small area means and the true values. The two criteria are defined as

$$\text{AAD} = \sum_{i=1}^{m} |\hat{\theta}_i - \theta_i|, \text{ and ASD} = \sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)^2. \tag{4.1}$$

We also construct the 95% empirical credible intervals for θ_i and compute the coverage rates over 100 datasets to examine the uncertainty quantification.

4.2 Results

The main results from our simulation study are presented in Figures 4.1-4.3. Among them, Figures 4.1 and 4.2 provide the AAD and ASD of estimated small area means from different models and under different settings of generating random effects. Figure 4.3 presents the average coverage of empirical 95% credible intervals over all small areas. These figures show that, among the models we consider, the HG model has the most robust performance in terms of both estimation accuracy and uncertainty quantification. Note that the model producing the smallest deviation measurements varies across different settings. Although the HG model is not always the best model in terms of the two deviation measures, its performance is often close to that of the best model regardless of the number of small areas and the settings of generating random effects. For the FH, HS, and LA models, although each of them produces the smallest AAD and ASD under certain settings, their performance could be significantly worse than that of the HG model in other settings. For example, FH performs best under the Normal setting and LA performs best under the Mixture 0.8 setting. However, under the Mixture 0.2 setting, they produce higher AAD and ASD than the HS model and the HG model. The HS model performs best under the Mixture 0.2 setting, but it produces the highest deviation measurements under the Normal setting and the Mixture 0.8 setting. Also, the coverage of the credible intervals from the HS model is significantly lower than that from other models and the nominal coverage rate under the two settings.

The robust performance of the HG model is realized through the adaptive choice of the hyperparameters a and b. Figure 4.4 provides the posterior mean estimates $\hat{a}$ and $\hat{b}$ of the hyperparameters. Under the three Mixture settings, both $\hat{a}$ and $\hat{b}$ increase as the proportion of nonzero random effects increases. In the sparsest setting (Mixture 0.2), $\hat{b}$ is close to zero and $\hat{a}$ is significantly smaller than one, indicating that the polynomial component in the gamma density of σ_i^2 plays a critical role in describing the random effects. In the densest setting (Mixture 0.8), $\hat{a}$ is close to one, meaning the exponential component plays a critical role. Tang et al. (2018) showed that in the GL model, polynomial-tailed local priors are better at characterizing small random effects while the exponential tail priors are better at characterizing large random effects. Our results are consonant with theirs.

Figure 4.1 AAD of estimated small area means from different models.

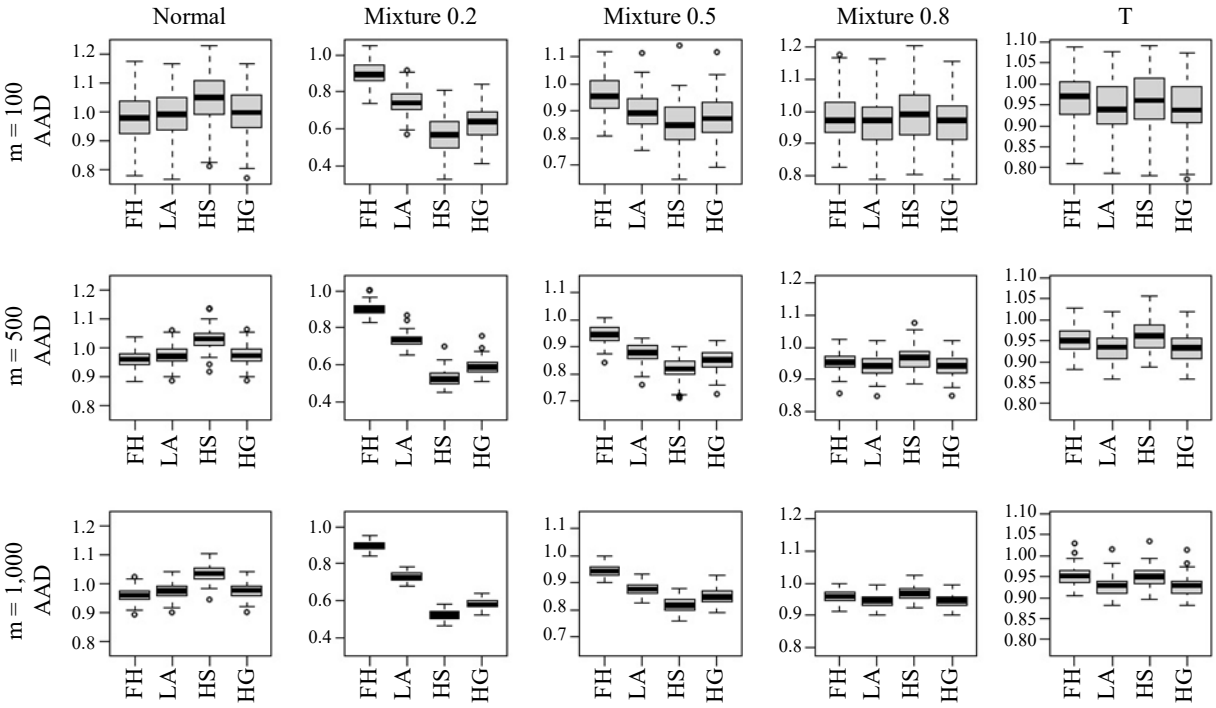


Figure 4.2 ASD of estimated small area means from different models.

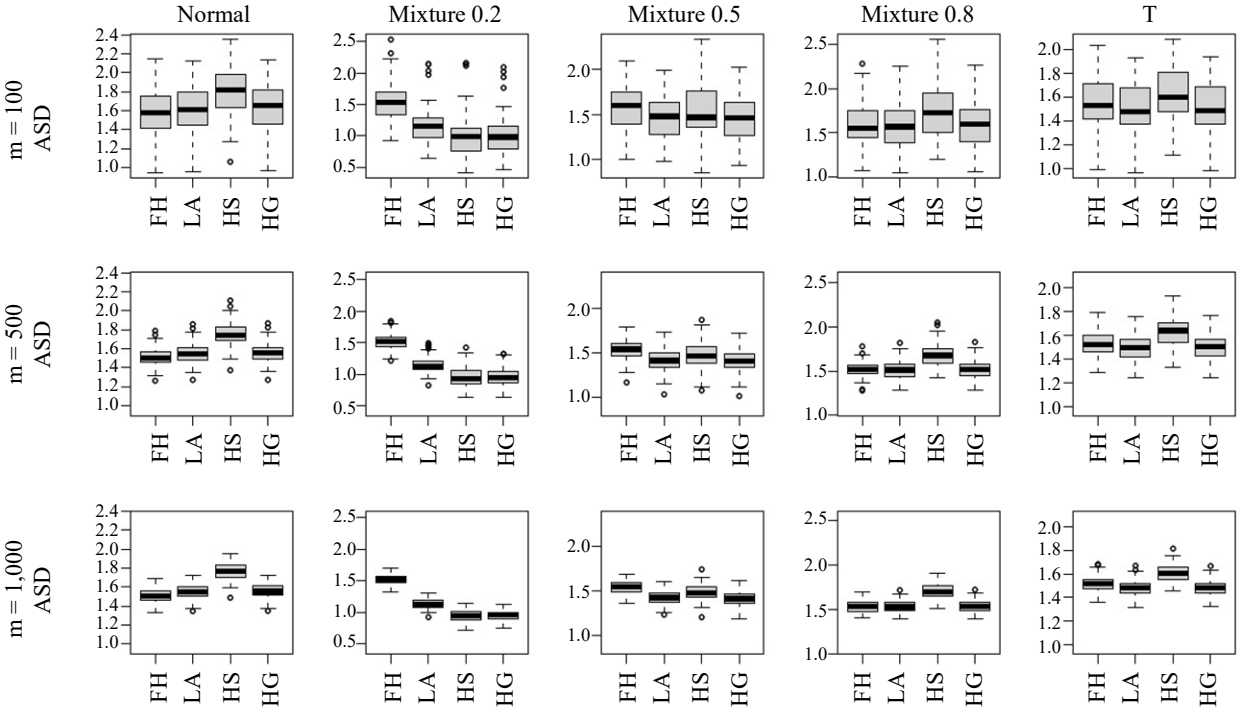


Figure 4.3 Average coverage of 95% empirical credible interval of small area means from different models.

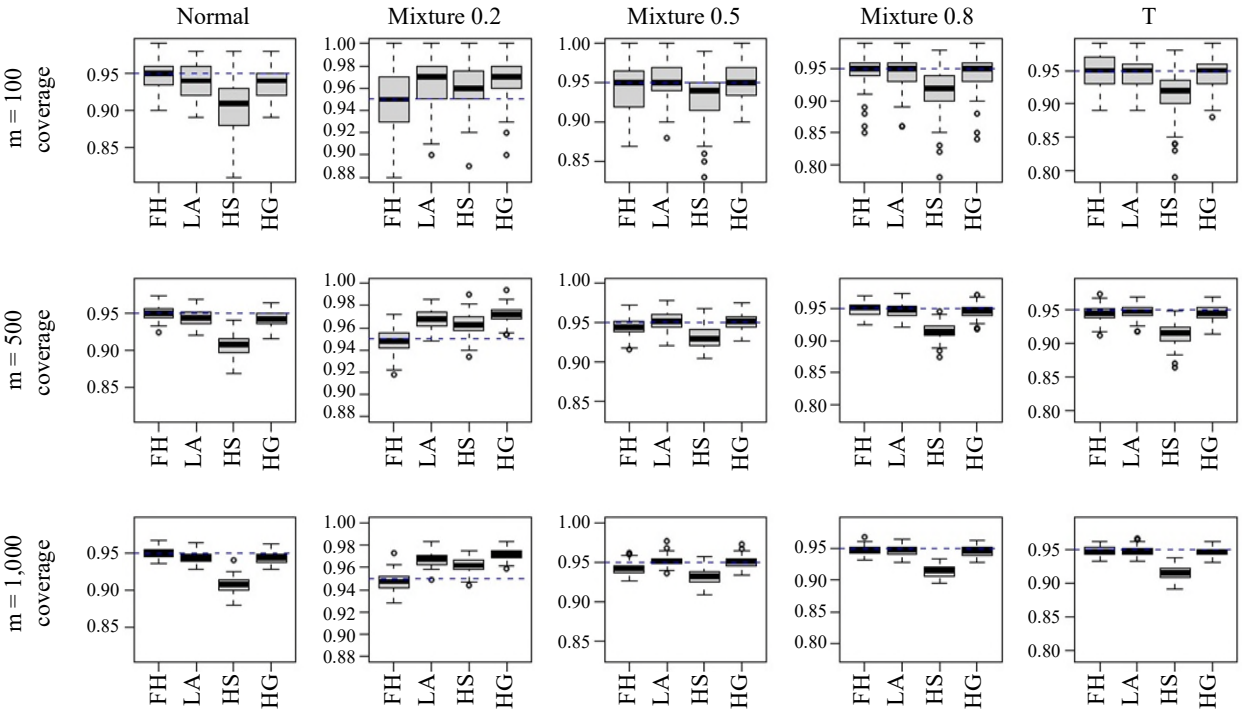


Figure 4.4 Estimated hyperparameters a and b in the HG model.

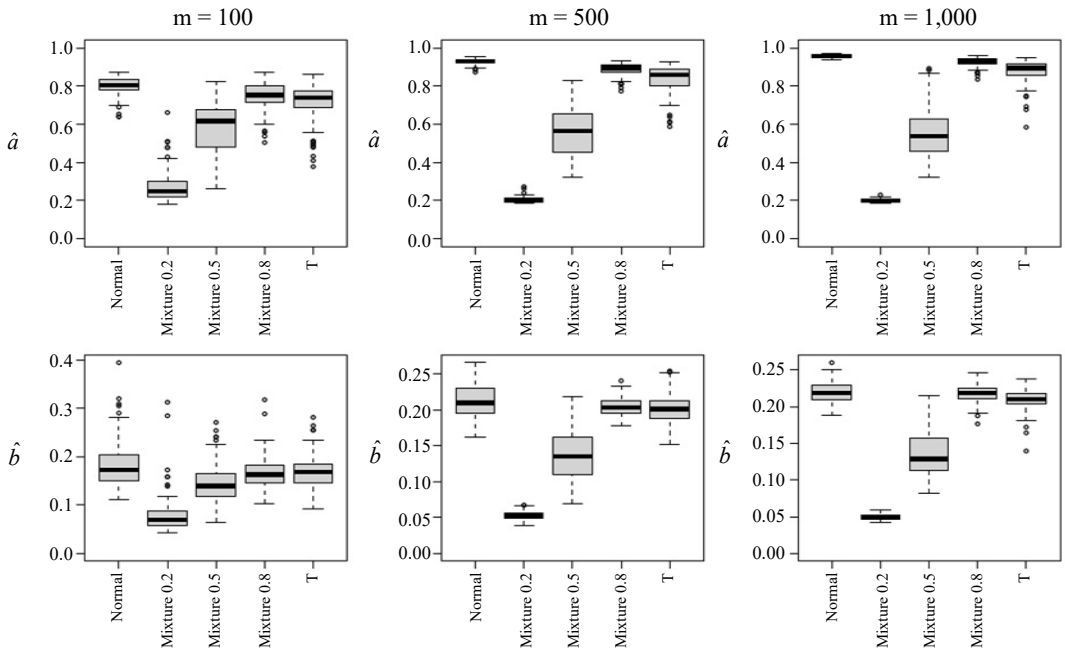
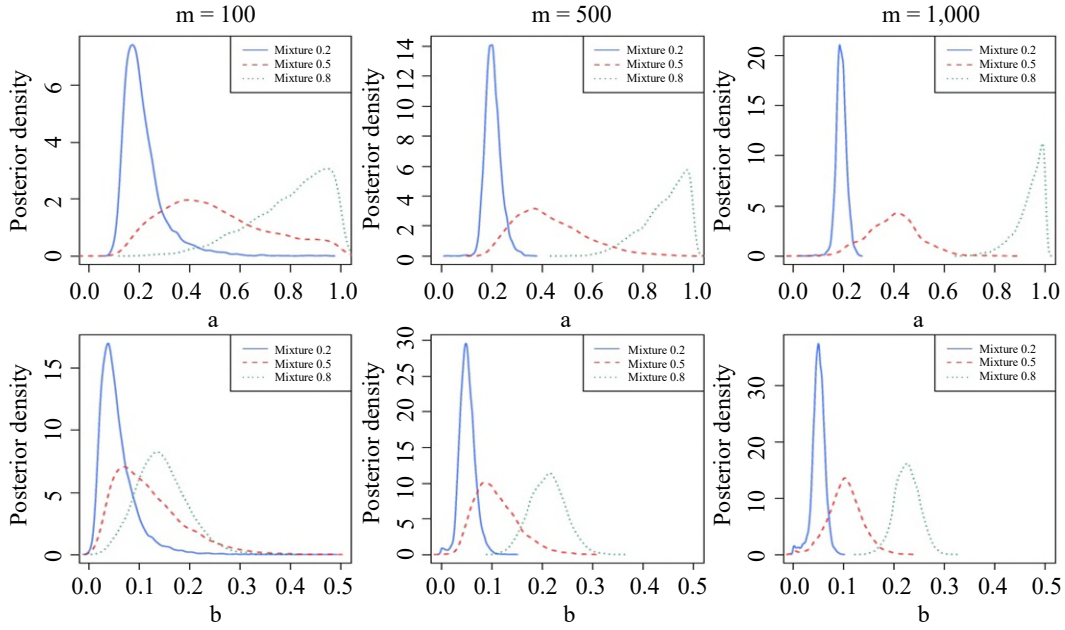


Figure 4.5 presents the posterior density of a and b under different settings for selected datasets. For both a and b, the posterior distribution concentrates on larger values for settings with a higher proportion of nonzero random effects.

Figure 4.5 Posterior density of a and b in the HG model for a typical dataset.



Figures 4.4 and 4.5 also show that the variation in the two hyperparameters decreases as the number of small areas m increases. More small areas provide more random effects and thus more information for characterizing the distribution of the random effects.

The HG model achieves the robust performance without sacrificing much in computational cost. Table 4.1 presents the time used for obtaining small area mean estimates from different models. Because of the extra effort in estimating the hyperparameters a and b, fitting the HG model often takes more time than the GL models. However, the increase is often no more than 15%. Also, for the GL models, as suggested in Tang et al. (2018), one can use the deviance information criterion (Spiegelhalter et al., 2002) to select the most appropriate prior for the local parameter. This requires fitting the GL models under several different local parameter priors, multiplying the computational cost while obtaining similar results as the HG model.

Table 4.1
The average (standard deviation) of computing time in seconds over 100 datasets under various settings.

m	Model	Random Effect Setting				
		Normal	Mixture 0.2	Mixture 0.5	Mixture 0.8	Student's T
100	FH	2.78 (0.11)	2.76 (0.16)	2.74 (0.11)	2.75 (0.12)	2.77 (0.13)
	LA	16.37 (0.80)	16.32 (0.82)	16.34 (0.87)	16.28 (0.75)	16.24 (0.79)
	HS	17.00 (0.79)	17.01 (0.97)	17.00 (0.82)	16.92 (0.76)	16.94 (0.84)
	HG	17.43 (0.87)	17.72 (0.80)	17.61 (0.94)	17.44 (0.77)	17.50 (0.82)
500	FH	3.96 (0.17)	3.92 (0.18)	3.95 (0.22)	3.90 (0.17)	3.93 (0.17)
	LA	70.54 (3.84)	70.52 (3.76)	70.30 (3.61)	70.27 (3.66)	70.29 (3.62)
	HS	73.15 (3.76)	73.17 (3.76)	72.97 (3.56)	72.92 (3.56)	72.95 (3.39)
	HG	77.41 (3.87)	79.69 (4.12)	78.43 (4.00)	77.25 (3.57)	77.52 (3.79)
1,000	FH	5.33 (0.20)	5.32 (0.21)	5.29 (0.20)	5.27 (0.20)	5.22 (0.24)
	LA	138.12 (7.22)	138.06 (7.29)	137.38 (7.09)	137.03 (7.00)	134.73 (7.06)
	HS	144.01 (7.65)	143.69 (7.84)	142.83 (7.52)	142.66 (7.51)	138.57 (6.83)
	HG	152.88 (7.42)	157.93 (7.68)	154.94 (7.67)	152.45 (7.34)	151.80 (8.23)

5. Real data analysis

In this section, we estimate the state- and county-level poverty rates in the United States using the proposed model. The two datasets we analyze come from Datta and Mandal (2015) and Tang et al. (2018). The first dataset concerns the state-level child poverty ratio for age group 5-17. Besides the direct estimates obtained from the 1999 Current Population Survey, the dataset also includes the number of child exemptions, Internal Revenue Service non-filer rate, and the residuals of regressing 1989 census poverty rates on the two previous variables. We include the three variables as covariates and an intercept in our model. In the second dataset, we have direct estimates of five-year (2007-2011) pooled county-level poverty rates for 3,141 counties from the American Community Survey. Food stamp participation rate is used as a covariate in addition to intercept.

Besides the proposed model, we also fit the FH model and the GL model with HS and LA priors to each dataset for comparison. Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) is used for comparing the model fit and the values are presented in Table 5.1. In the state-level analysis, following Datta and Mandal (2015) and Tang et al. (2018), we treat the ratio benchmarked state-level poverty ratios obtained from the 2000 census as the "true" small area means and measure the errors of the estimated values using AAD and ASD defined in (4.1). The results are also presented in Table 5.1.

For state-level estimation, the HG model is the most appropriate model in terms of DIC. The posterior median estimates of parameters a and b in the HG model are 0.53 and 0.22, respectively. The relatively small estimate of b suggests that the polynomial component in the prior of random effects is important. This agrees with the DIC results that the HS prior is preferred over the LA prior in GL models. The values of the two deviation measures also suggest that the HG and HS models have similar performance with the HG model incurring slightly larger errors. It is worth noting that the LA model produces the smallest errors, especially in terms of ASD, although DIC is not in favor of it.

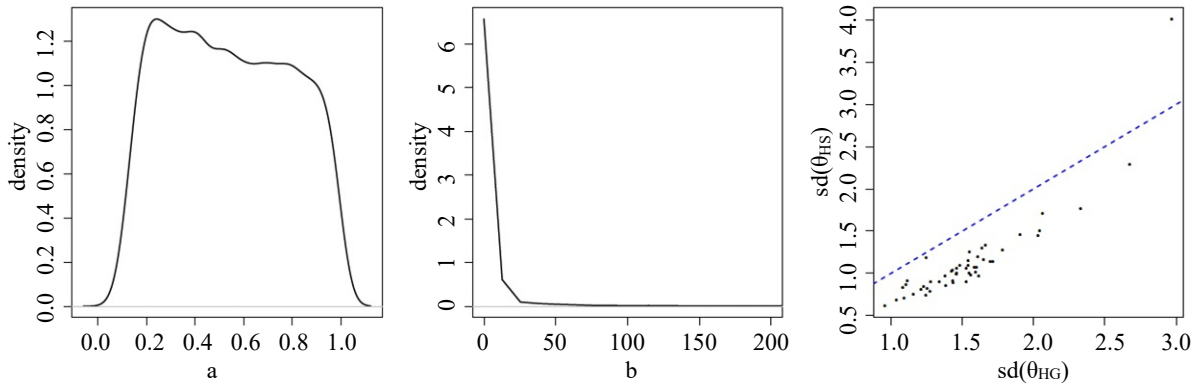
Table 5.1
Performance of various models for state-level and county-level data.

	Measure	HG	FH	HS	LA
State Level	DIC	271.52	273.29	273.09	275.92
	AAD	1.05	1.19	1.01	0.99
	ASD	2.19	2.55	2.04	1.68
County Level	DIC	-15,946.23	-15,883.34	-15,751.12	-15,946.96

The left and middle panels of Figure 5.1 present the posterior densities of a and b. Both distributions exhibit high variations. The density of a does not vary much between 0.1 and 0.9. The density of b has a very long tail although a significant portion of probability is distributed around zero. The high variation is mainly a result of the small number of small areas $(m = 51)$, suggesting a high degree of uncertainty in determining a suitable model. This is also reflected in the close DIC values for different models. The uncertainty in the hyperparameters also has an impact on the variability of small area means. The right panel of Figure 5.1 shows a plot of the posterior standard deviations of small area means from the HS model

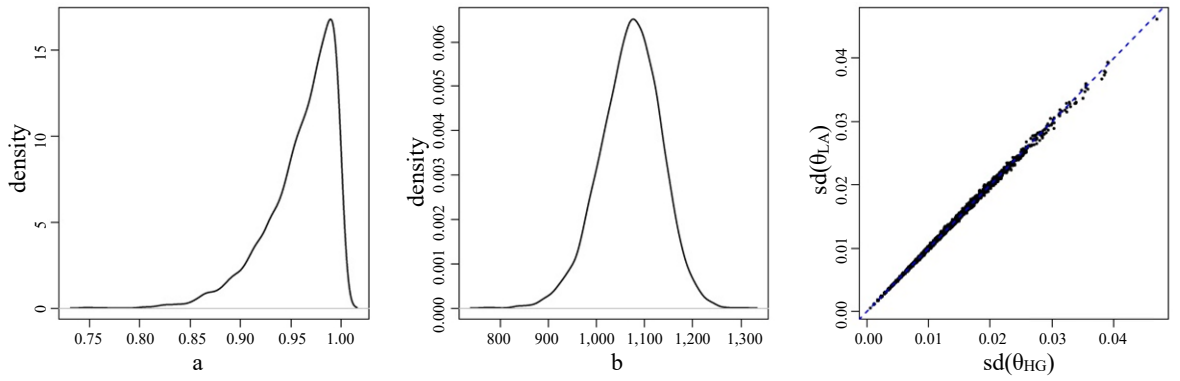
against those from the HG model. Because of the uncertainty in a and b, the HG model produces a higher posterior standard deviation in most of the small areas.

Figure 5.1 Results of state-level estimations: left: posterior densities of parameter a; middle: posterior densities of parameter b; right: posterior standard deviations from HG and HS models.



For county-level estimation, Tang et al. (2018) showed that the LA prior performs the best among other choices of priors in the GL model. The DIC values presented in Table 5.1 indicate that the proposed HG model achieves a similar fit as the LA model. In fact, under the HG model, the posterior median of a is 0.97, which is very close to 1, and the posterior median of b is 1073.25, indicating that the HG prior resembles the exponential tail LA prior in this case. Moreover, as a result of a much larger number of small areas in the county-level data, the posterior densities of a and b shown in the left and middle panels of Figure 5.2 have a smaller variation than those obtained in the state-level estimation. Correspondingly, as shown in the right panel of Figure 5.2, the posterior standard deviations of small area means produced by the HG and LA models are also close.

Figure 5.2 Results of county-level estimations: left: posterior densities of parameter a; middle: posterior densities of parameter b; right: posterior standard deviations from HG and HS models.



The datasets and R code for producing the results in this section are available at https://github.com/xytangtang/HGSAE.

6. Conclusion

In this paper, we proposed a hierarchical gamma (HG) model for the random effects in small area estimation. It assumes the random effects follow a scale mixture of normal distributions with the mixing distribution being gamma. Hyperpriors are further placed on the shape and rate parameter of the gamma distribution. We showed through simulations and real data analyses that the proposed model is capable of characterizing heterogeneous random effects across small areas as the global-local (GL) model while not requiring fitting the model multiple times to choose the most appropriate priors for the local parameters.

The HG model can be seen as a mixture of different global-local models. Because of such formulation, the posterior variation of small area means from the HG model takes into account model uncertainty (the variation in the estimation results from different models) to some degree. When the number of small areas is small, there is often not much information for the hyperparameters a and b, leading to higher model uncertainty and thus higher variation of small area means compared to using the GL model with a specific choice of the priors for local parameters. For this reason, we recommend to use the proposed model when the number of small areas is large to avoid large estimation variance.

In this article, we focus on the area-level models for small area estimation. Developing a similar method for unit-level models is a possible future direction.

Appendix

A. Proof of Theorem 1

Under the proposed model, the posterior density can be written as

$$\begin{aligned}\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y}) = \; & K\prod_{i=1}^{m}\exp\left[-\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-u_i)^2}{2D_i}\right]\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\\ & \times\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]b^{s_0-1}\exp(-bt_0),\end{aligned} \tag{A.1}$$

where K is a generic constant that does not depends on $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\sigma}^2$, a, and b. It is sufficient to show the integral of $\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y})$ with respect to $\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,$ and b is finite.

First, consider integration with respect to $\boldsymbol{\beta}$. Let $\mathbf{z}=(z_1,\ldots,z_m)^\top$ with $z_i=y_i-u_i$ for $i=1,\ldots,m$. Since $\mathbf{X}$ is of full column rank, we can define $\hat{\boldsymbol{\beta}}=(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{z}}$. Note that

$$\prod_{i=1}^{m}\exp\left[-\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-u_i)^2}{2D_i}\right]=\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top(\mathbf{X}^\top\mathbf{D}\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-\frac{1}{2}\tilde{\mathbf{z}}^\top(\mathbf{I}-\mathbf{P})\tilde{\mathbf{z}}\right],$$

where $\mathbf{P}=\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top$. Integrating $\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y})$ with respect to $\boldsymbol{\beta}$ leads to

$$\begin{aligned}\pi(\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y}) &= K\exp\left[-\frac{1}{2}\tilde{\mathbf{z}}^{\top}(\mathbf{I}-\mathbf{P})\tilde{\mathbf{z}}\right]\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\\ &\quad\times\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]b^{s_0-1}\exp(-bt_0)\\ &\le K\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]\\ &\quad\times b^{s_0-1}\exp(-bt_0).\end{aligned} \tag{A.2}$$

Note that the expression in the last two lines in (A.2) gives the joint prior of $\mathbf{u}$, $\boldsymbol{\sigma}^2$, a, and b. Since the prior distribution is proper, the proof now completes.

References

Armagan, A., Clyde, M. and Dunson, D.B. (2011). Generalized beta mixtures of Gaussians. *Advances in Neural Information Processing Systems*, 523-531.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2009). Handling sparsity via the horseshoe. *International Conference on Artificial Intelligence and Statistics*, 73-80.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480.

Chakraborty, A., Datta, G.S. and Mandal, A. (2016). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition New Series*, 17(1), 67-90.

Datta, G.S., and Lahiri, P. (1995). Robust hierarchical bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, 54(2), 310-328.

Datta, G.S., and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110(512), 1735-1744.

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(493), 362-374.

Fabrizi, E., and Trivisano, C. (2010). Robust linear mixed models for small area estimation. *Journal of Statistical Planning and Inference*, 140(2), 433-443.

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(366a), 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(410), 398-409.

Griffin, J.E., and Brown, P.J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick.

Griffin, J.E., and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1), 171-188.

Li, Y., and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102(478), 664-673.

Neal, R.M. (2003). Slice sampling. *The Annals of Statistics*, 31(3), 705-767.

Park, T., and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

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(1), 15-29.

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

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: Series B*, 64(4), 583-639.

Sugasawa, S., and Kubokawa, T. (2020). Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*, 3, 693-720.

Tang, X., Ghosh, M., Ha, N.S. and Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113(524), 1476-1489.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 303-321
Statistics Canada, Catalogue No. 12-001-X

Design-based estimation of small and empty domains in survey data analysis using order constraints

Xiyue Liao, Mary C. Meyer and Xiaoming Xu[1]

Abstract

Recent work in survey domain estimation has shown that incorporating *a priori* assumptions about orderings of population domain means reduces the variance of the estimators and provides smaller confidence intervals with good coverage. Here we show how partial ordering assumptions allow design-based estimation of sample means in domains for which the sample size is zero, with conservative variance estimates and confidence intervals. Order restrictions can also substantially improve estimation and inference in small-size domains. Examples with well-known survey data sets demonstrate the utility of the methods. Code to implement the examples using the `R` package `csurvey` is given in the appendix.

Key Words: Domain means; Isotonic; Order restrictions; R; Small area estimation; Survey.

1. Background and Introduction

Consider a finite population with labels $U = \{1,\ldots,N\}$ and let U_d, $d = 1,\ldots,D$ denote a partition of the population into domains where U_d has N_d elements. For a study variable y, suppose interest is in estimating the population domain means

$$\bar{y}_{U_d} = \frac{\sum_{k\in U_d} y_k}{N_d}$$

for each d, and providing inference such as confidence intervals for each $\bar{y}_{U_d}$. Given a survey design, a sample $s \subset U$ is chosen; let $s_d = s \cap U_d$ for $d = 1,\ldots,D$. The standard Hájek estimator $\tilde{\mathbf{y}}_s = (\tilde{y}_{s_1},\ldots,\tilde{y}_{s_D})^\top$ of the population domain means is a weighted average of the sample observations in each domain d. Specifically,

$$\tilde{\mathbf{y}}_{s_d} = \frac{\sum_{i\in s_d} y_i \big/ \pi_i}{\sum_{i\in s_d} 1 \big/ \pi_i},$$

where π_i is the sampling probability for the ith population element calculated from the sampling design (see Särndal, Swensson and Wretman (1992), page 185).

Because the estimate for each domain uses only the observations within a domain, a small domain sample size results in unreliable estimators within that domain. Traditional small area estimation methods use observations in other domains to provide more information for the small sample size domains. The Fay-Herriot estimator accomplishes this by imposing a parametric model on the domain means, with a domain random effect to account for the departure of domain means from the assumed overall model. See Rao and Molina (2015) and Pfeffermann (2013) for a comprehensive treatment of small area estimation methods.

1. Xiyue Liao, Mary C. Meyer and Xiaoming Xu, San Diego State University, Colorado State University and Duke University. E-mail: xliao@sdsu.edu.

We will consider *a priori* assumptions on the population domain means that can be expressed as a partial ordering of domains. For example, in a workplace survey we could assume that average salary increases with job rank, within job type and location. In an environmental survey, it might be reasonable to assume that amount of pollution decreases in the distance from the source. Orderings imposed in the examples of Section 4 include assuming that test scores decrease as poverty increases, and that average cholesterol level increases with the age and waist size of the subject. The orderings allow information to be shared across domains without parametric modeling.

Let $S = \{1,\ldots,D\}$ enumerate the domains; a partial ordering of S is specified by a binary relation $\preceq$, so that for d_i and $d_j \in S$, the expression $d_i \preceq d_j$ means that we assume $\overline{y}_{U_i} \le \overline{y}_{U_j}$. The partial ordering must have the following properties: reflexive ($d \preceq d$ for all $d \in S$), anti-symmetric (if $d_i, d_j \in S$, $d_i \preceq d_j$, and $d_j \preceq d_i$, then $\overline{y}_{U_i} = \overline{y}_{U_j}$), and transitive (for $d_i, d_j, d_k \in S$, if $d_i \preceq d_j$, and $d_j \preceq d_k$, then $d_i \preceq d_k$). A complete ordering has the additional property that all pairs of points in S are comparable (given $d_i, d_j \in S$, then either $d_i \preceq d_j$ or $d_j \preceq d_i$ or both). Orderings of interest in survey domain mean estimation include complete orderings, orderings in grids of domains, and block orderings.

Wu, Meyer and Opsomer (2016) considered a complete ordering on the sequence of domain means, applying the pooled adjacent violators algorithm (Brunk, 1958) for domain mean estimation. They derived confidence intervals that have smaller width without sacrificing coverage, compared to the estimators that do not consider the ordering. Oliva-Aviles, Meyer and Opsomer (2020) developed methodology for partial orderings as well as more general linear constraints on domains.

A partial ordering can be imposed on the domain mean estimators using linear inequality constraints in the form of an $m \times D$ constraint matrix $\mathbf{A}$, and the constrained estimator $\tilde{\boldsymbol{\theta}}$ of the vector of domain means is found by minimizing

$$\min_{\theta}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta})^\top \mathbf{W}_s(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}) \text{ such that } \mathbf{A}\boldsymbol{\theta} \ge \mathbf{0}. \tag{1.1}$$

The weight matrix $\mathbf{W}_s$ is diagonal with ith element $\hat{N}_i / \hat{N}$, where $\hat{N}_i = \sum_{i \in s_d} 1/\pi_i$, and $\hat{N} = \sum_{i=1}^{D} \hat{N}_i$.

For a simple example of a constraint matrix, consider five domains with a complete ordering, where we assume $\overline{y}_{U_1} \le \overline{y}_{U_2} \le \overline{y}_{U_3} \le \overline{y}_{U_4} \le \overline{y}_{U_5}$. Perhaps these are average cholesterol levels over five age groups, or average wages over employee ranks. The constraint matrix is

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

For complete orderings on D domains, the constraint matrix is $(D-1) \times D$. For an example of a partial ordering, suppose we consider five age groups for women and for men, and we still want to order cholesterol level by age within the sexes, but not have any ordering between the sexes. If the first five domains represent the five age groups for women, and the domains 6-10 represent the age groups for men, then the 8×10 constraint matrix is

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

where the first four rows set the constraints for the women's age groups, and the last four rows are the constraints for the men's age groups. The ordering is not complete, because, for example, domain 2 is not comparable with domain 8.

The *a priori* assumptions about the ordering may be verified using the cone information criterion (CIC) developed by Oliva-Aviles, Meyer and Opsomer (2019). The CIC is similar to the familiar Akaike information criterion in that it is a measure of goodness of fit, with a penalty for effective degrees of freedom, and is provided by the `csurvey` package (see the code in Appendices A and B). The CIC is reported for both the constrained and unconstrained Hajék estimators, and if the CIC is smaller for the constrained estimator, this is evidence that the order assumptions are correct.

The solution to (1.1) is the weighted projection of $\tilde{\mathbf{y}}_s$ onto the subset C of $\mathbb{R}^D$ defined by $\mathbf{A}$:

$$C = \left\{ \boldsymbol{\theta} \in \mathbb{R}^D : \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0} \right\}. \tag{1.2}$$

This subset C of $\mathbb{R}^D$ is a *cone* because for any $\boldsymbol{\theta} \in C$ and any $a \geq 0$, $a\boldsymbol{\theta} \in C$. Oliva-Aviles, Meyer and Opsomer (2020) explained how such a cone projection leads to pooling of information across domains, which provides more precise estimators. The order-constrained estimator of domain means is constructed by an optimal pooling of the domains over which the unconstrained estimators violate the ordering, and the pooling reduces the estimated variance because the averages are over larger numbers of observations. They constructed a covariance estimator based on the observed pooling and showed how this produced smaller confidence intervals with good coverage. Xu, Meyer and Opsomer (2021) developed a variance estimator based on a mixture of covariance matrices, which we will call the mixture covariance estimator. Instead of constructing the covariance estimator using the observed pooling, the mixture covariance estimator recognizes that another data set might result in a different pooling, hence a weighted average of covariance matrices for all possible poolings is used. They provided some large sample theory and showed that the mixture covariance estimator improves coverage of confidence intervals while retaining smaller interval lengths.

In this paper we extend the previous works, providing estimation and inference for empty cells that are not on the "boundary" of the assumed ordering. We generalize these ideas with a method for imposing order constraints on the upper and lower confidence bounds, which allows for smaller confidence intervals in small sample size domains. We show that the coverage of the adjusted confidence intervals is at least as good as that of the original confidence intervals provided by the mixture covariance matrix.

In Section 2 we propose a simple method for estimating domain means (and providing confidence intervals) where the sample size in the domain is zero, and a partial ordering is assumed. This is a design-based estimator that does not incorporate a parametric model to assist estimation. In Section 3 we impose the assumed ordering on the upper and lower confidence interval bounds, leading to smaller confidence interval lengths for domains with small sample sizes. We also show that the method provides valid confidence intervals, and we provide some simulations comparing the proposed confidence intervals to those for the unconstrained Hájek estimator and the Fay-Herriot estimator. Examples of small area estimation using well-known data sets are given in Section 4, and some discussion is provided in Section 5. The methods are available in the `csurvey` package; the appendix contains the code to reproduce results from the examples in Section 4.

2. Estimation for empty domains in a partial ordering

One of the major advantages to implementing valid inequality constraints is that information is pooled across domains to construct the estimates, while for the unconstrained design-based estimator, each domain estimator uses only the observations in that domain. When the sample size within the domain is small, the unconstrained design-based variance estimates are unreliable, and for cells with only one observation, or for empty cells, the variance cannot be estimated at all without using additional assumptions such as in some small area estimation methods. See Rao and Molina (2015), preface, for a nice discussion of why and when small area estimation methods are needed. The order constraints, if appropriate, allow for a completely design-based approach to estimation of means in small sample size domains.

If a domain d has no observations it cannot be included in the estimation based on (1.1), but if we can assume some inequality constraints that involve $\overline{y}_{U_d}$, then an estimate for $\overline{y}_{U_d}$ can be provided along with a conservative confidence interval. For an illustrating example, suppose there are 20 domains with a simple non-decreasing ordering, and only domain 16 has no observations. We can get estimates and confidence intervals for the 19 non-empty domains, and argue that the 16th population domain mean must not exceed the 17th but must be at least as big as the 15th. The bottom of the confidence interval for the 15th domain mean, combined with the top of the interval for the 17th domain mean, provides a conservative confidence interval for the 16th domain mean. The 16th domain mean estimator can be taken to be the center of its confidence interval, or as close to the center as possible while satisfying the constraints. If there are consecutive empty domains, the non-empty boundary domains provide the confidence intervals.

For more complicated constraints, we can apply the same idea of getting upper and lower bounds for empty domains, using the estimates and confidence intervals calculated from observations in the other domains. Let D' be the number of domains and m' be the total number of constraints imposed on the domain means, using an $m' \times D'$ constraint matrix $\mathbf{A}'$. Let D be the number of non-empty domains and let m be the number of constraints imposed on the non-empty domain means; then an $m \times D$ constraint matrix $\mathbf{A}$ can be obtained by modifying $\mathbf{A}'$. Using these, the constrained estimator $\tilde{\boldsymbol{\theta}}$ as well as the $D \times D$ mixture covariance matrix can be obtained for the D non-empty domains. For an empty domain d, we look at the

dth column of $\mathbf{A}'$ to find non-empty domains d_1 and d_2 such that $\overline{y}_{U,d_1} \leq \overline{y}_{U,d} \leq \overline{y}_{U,d_2}$. If two such domains do not exist, as for a "corner" domain, then the domain mean cannot be estimated, and the confidence interval will be unbounded at one end. If there is at least one d_1 such that $\overline{y}_{U,d_1} \leq \overline{y}_{U,d}$, then we look at the confidence intervals for all such d_1, and choose the largest lower bound as the lower bound for the confidence interval of the empty domain. Similarly, if there is at least one d_2 for which $\overline{y}_{U,d} \leq \overline{y}_{U,d_2}$, then we look at the confidence intervals for all such d_2, and choose the smallest upper bound as the upper bound for the confidence interval of the empty domain.

3. Ordering the confidence interval bounds

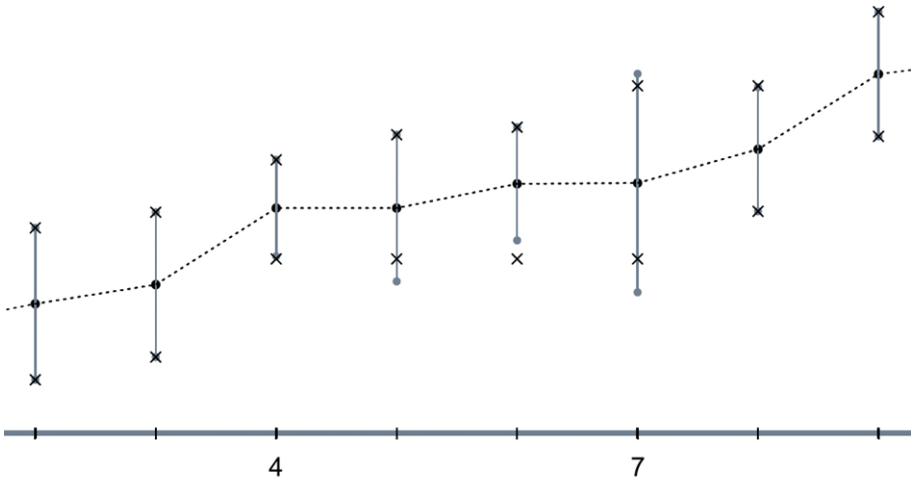
For the constrained methods, the problem of small sample sizes in domains is mitigated because both the domain mean estimator and the mixture variance estimator use information from other domains. The adjustments of the confidence intervals proposed here, imposing the order constraints on the upper and lower confidence bounds as well as on the domain mean estimates, allow for more sharing of information across domains.

The order constraints should also hold in the upper confidence interval bounds, and in the lower bounds. This can be seen by considering a case where the mean of domain 6, for example, is assumed to be not greater than the mean for domain 7. The upper confidence bound for domain 7 indicates a level of confidence that the population domain mean is smaller than this, so this level of confidence should apply also to the (smaller) population mean for domain 6.

The confidence intervals provided by the mixture covariance matrix do not necessarily satisfy the constraints. Recall that this estimated covariance matrix is a weighted average of linear covariance matrices, and the upper and lower bounds calculated using this mixture covariance estimate do not necessarily follow the partial ordering assumed for the domains. This is illustrated in Figure 3.1, where domain 7 has a small sample size compared to the surrounding domains. The estimated domain means are marked with black dots, and the grey dots are the bounds for the unadjusted confidence intervals computed from the mixture covariance matrix. For the example in the figure, the estimated variance for domain 7 is larger, causing the lower bound to be less than the lower bound for domain 6. However, we are assuming that the 7th population domain mean is at least as large as the 6th, so we adjust the confidence bounds so that they satisfy the order constraints.

More generally, if $\tilde{\mathbf{u}}$ are the upper bounds for the confidence intervals computed from the variances given by the mixture covariance estimator, we project $\tilde{\mathbf{u}}$ onto the cone (1.2), using the sample sizes (or effective sample sizes) as weights. The projection $\hat{\mathbf{u}}$ is the new set of upper bounds, satisfying the constraints. Similarly $\hat{\ell}$ is the new set of lower bounds obtained by the weighted projection of the original lower bounds $\tilde{\ell}$ onto the constraint cone. The $\times$ marks in Figure 3.1 indicate the new, order-constrained upper and lower bounds. Confidence intervals for domains with small sample sizes are adjusted more than those for larger domains, resulting in more precise small-domain estimation.

Figure 3.1 Example of confidence bound adjustment to ensure that the bounds follow the order constraints.



Notes: The domain mean estimates are marked with black dots; the confidence bounds provided by the mixture covariance matrix are given as grey dots; the adjusted confidence bounds marked with ×.

The new upper bounds $\hat{\mathbf{u}}$ are the weighted projection of $\tilde{\mathbf{u}}$ onto C. That is, $\hat{\mathbf{u}}$ minimizes $\sum_{d=1}^{D} n_d(\tilde{u}_d - u_d)^2$ over $\mathbf{u} \in C$, where n_d is the number of observations in domain $d \in S$. Let $A_c \subseteq S$ be defined as $\{d \in S: \hat{u}_d = c\}$. The following Lemma is similar to Theorem 1.3.5 of Robertson, Wright and Dykstra (1988) and tells us that any $\hat{u}_d$ is a weighted average of some subset of the $\tilde{u}_1,\ldots,\tilde{u}_D$.

Lemma 1. *If A_c is not empty, then*

$$c = \frac{\sum_{d\in A_c} \tilde{u}_d n_d}{\sum_{d\in A_c} n_d}.$$

Proof: Let $L_c = \{d : \hat{u}_d < c\}$ and $U_c = \{d : \hat{u}_d > c\}$. Let c_1 be the maximum of $\hat{u}_j$ for $j \in L_c$, and let c_2 be the minimum of $\hat{u}_j$ for $j \in U_c$. Then $c_1 < c < c_2$. Write

$$\sum_{d=1}^{D} [\hat{u}_d - \tilde{u}_d]^2 n_d = \sum_{d\in L_c} [\hat{u}_d - \tilde{u}_d]^2 n_d + \sum_{d\in A_c} [c - \tilde{u}_d]^2 n_d + \sum_{d\in U_c} [\hat{u}_d - \tilde{u}_d]^2 n_d,$$

and note that the value of c that minimizes the middle term is given in the Lemma. If the result were not true, then the middle term could be made smaller by moving c up or down, toward the given weighted average of the $\tilde{u}_d$, and if we stay within (c_1, c_2), the function would still satisfy the constraints.

The confidence bounds $(\tilde{\ell}_d, \tilde{u}_d)$ obtained from the mixture covariance matrix were shown by Xu, Meyer and Opsomer (2021) to have correct asymptotic coverage if the constraints hold strictly. If for any d, $\hat{u}_d \geq \tilde{u}_d$, then the upper coverage of $\bar{y}_{U_d}$ is at least as good as for the unadjusted interval. Similarly, if $\hat{\ell}_d \leq \tilde{\ell}_d$ then $\hat{\ell}_d \leq \bar{y}_{U_\ell}$ if $\tilde{\ell}_d \leq \bar{y}_{U_\ell}$. The coverage for the adjusted confidence intervals when $\hat{u}_d < \tilde{u}_d$ or $\hat{\ell}_d > \tilde{\ell}_d$ is addressed in the following.

Theorem 1. *For $d \in S$ such that $\hat{u}_d < \tilde{u}_d$, let $c = \hat{u}_d$ and define $A_c^+ = \{j \in S: \hat{u}_j = c$ and $d \preceq j\}$. Then A_c^+ has at least two elements, and if for any $j \in A_c^+$ we have $\tilde{u}_j \geq \bar{y}_{U_j}$, this guarantees $\hat{u}_d \geq \bar{y}_{U_d}$. Similarly, for*

$d \in S$ such that $\hat{\ell}_d > \tilde{\ell}_d$, let $c = \hat{\ell}_d$ and define $A_c^- = \{j \in S: \hat{\ell}_j = c \text{ and } j \preceq d\}$ Then A_c^- has at least two elements, and if for any $j \in A_c^-$ we have $\tilde{\ell}_j \le \bar{y}_{U_j}$, this guarantees $\hat{\ell}_d \le \bar{y}_{U_d}$.

Proof: Suppose for some domain d, $\hat{u}_d \neq \tilde{u}_d$. If $\hat{u}_d > \tilde{u}_d$, then certainly $\tilde{u}_d \ge \bar{y}_{U_d} \Rightarrow \hat{u}_d \ge \bar{y}_{U_d}$. If $\hat{u}_d < \tilde{u}_d$, let $c = \hat{u}_d$. We know by Lemma 1 that A_c has more than one element. There is at least one $j \in A_c^+$ such that $\tilde{u}_j < \tilde{u}_d$; if not, then $\hat{u}_d$ could be larger and closer to $\tilde{u}_d$ without changing other $\hat{u}_j$. Because $\bar{y}_{U_d} \le \bar{y}_{U_j}$ and $\tilde{u}_j < \hat{u}_j = \hat{u}_d$,

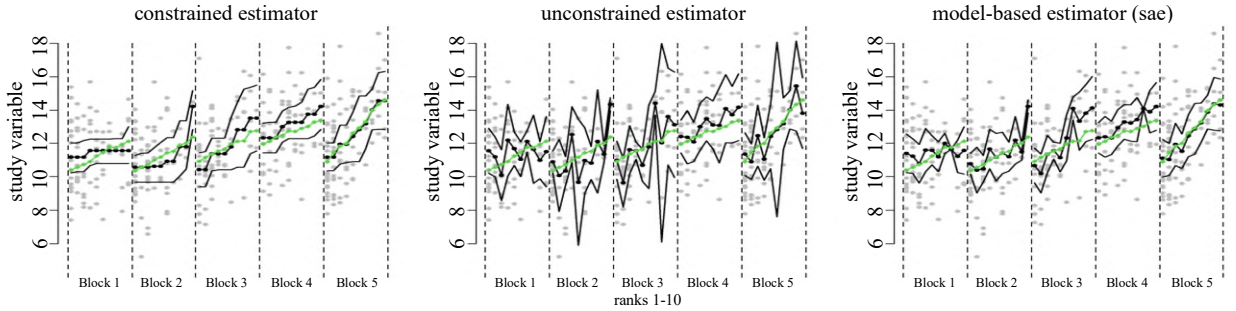
$$\tilde{u}_j \ge \bar{y}_{U_j} \Rightarrow \hat{u}_j \ge \bar{y}_{U_j} \Rightarrow \hat{u}_d \ge \bar{y}_{U_j} \Rightarrow \hat{u}_d \ge \bar{y}_{U_d}.$$

The proof for the lower bound is similar.

To demonstrate the effectiveness of the empty-and-small domain methods, we generated a population of size 40,000 using a vector $\boldsymbol{\mu}$ of 50 domain means, so that each domain has 800 population values from a normal distribution with mean μ_i and variance 4. Suppose the units of the simulated population are workers in a certain field, and the values represent the log of the salaries. There are 10 job ranks and 5 locations comprising the 50 domains, and the domains with lower ranks have higher numbers of workers. We assume that within each location, the salaries are increasing in rank, and we impose a block ordering on the locations: For each rank, the mean salary in location 4 is higher than the mean salary of the corresponding rank in any of locations 1, 2, and 3, and the mean salaries for the ranks in location 5 are also higher than those in the first three locations. However, there is no ordering imposed within locations 1, 2, and 3; similarly we do not impose an order on the mean salaries of locations 4 and 5. A stratified design is used to sample from the population.

A sample of size $n = 400$ from this population is shown as the gray dots in Figure 3.2. The true population means are shown as the diamonds, which are almost linear in the finite population. The estimates are shown as black dots, and the 95% confidence intervals are shown as well. The constrained estimator is compared to the unconstrained Hajék estimator, and to the Fay-Herriot estimate given by the `sae` package using rank and block as ordinal and nominal predictors, respectively. The average domain sample size is 8, but some domains corresponding to higher ranks are more likely to have smaller sample sizes. The confidence intervals for the constrained estimator appears to have better coverage and small length.

Figure 3.2 A simulated data set from a population of workers where the study variable is log(salary).

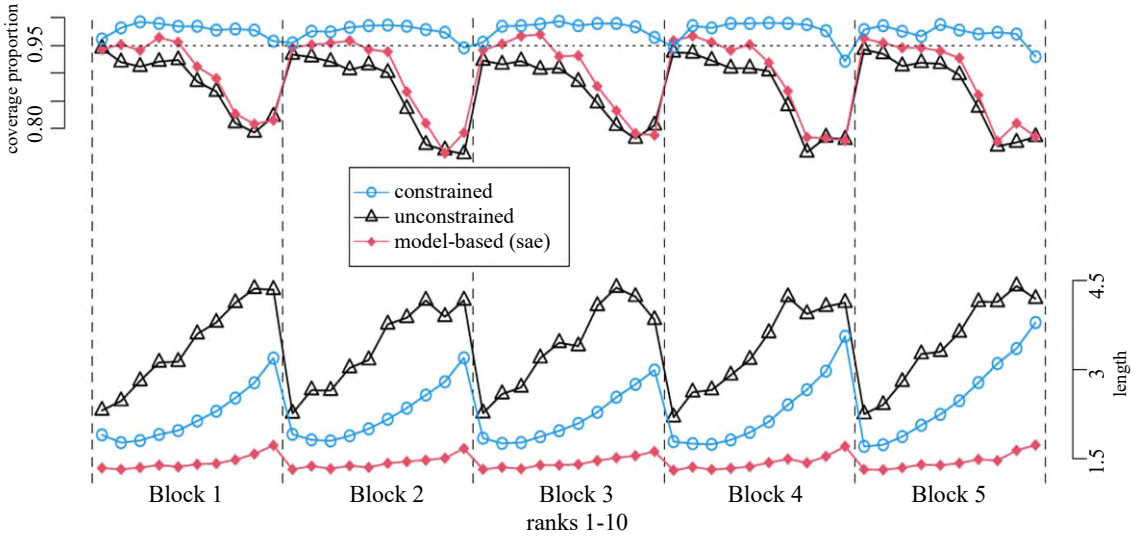


Notes: For the constrained estimator, salary is assumed to be increasing in ranks 1-10 within each location, and across locations salaries follow a block order, with locations 4 and 5 having higher mean salaries compared to locations 1, 2, and 3. The population means are shown as the lighter-colored diamonds, while the estimates are shown by the black dots. The approximate 95% confidence intervals are shows as the lines. The unconstrained estimator given by `survey` and the Fay-Herriot estimator given by `sae` are also shown.

To demonstrate the coverage and precision of the order-restricted estimates with small sample size, we sampled 1,000 data sets of size 400, using the stratified design. For each data set, we computed the constrained estimates, the Hajék estimates, the Fay-Herriot estimates, and the 95% confidence intervals for each. For each of the 50 domains, we determined the proportion of data sets for which the confidence interval captured the population mean, and we also determined the lengths of the confidence intervals. The coverage rates and interval lengths are summarized for both estimators in Figure 3.3. Because the Hajék estimate cannot be computed in empty domains, the reported percent coverage is limited to the data sets with non-empty domains. Similarly, the percent coverage for the model-based estimator is over the simulated data sets for which the estimator can be computed.

For the Hajék estimator, the intervals for the higher ranks have poor coverage within each block, due to the smaller sample sizes. The model-based `sae` estimator is an improvement in that the coverage is higher than for the unconstrained estimator, and the lengths are considerably smaller. However, the `sae` coverage is near the target only for ranks 1-5, which have larger population and sample sizes. The coverage for the order-constrained estimator is still good for these domains, because information from the domains with larger sample sizes is used. The lengths of the constrained intervals are consistently smaller than for the Hajék estimator, while the coverage is consistently good. Complete code to produce these simulation results is available in the Supplemental Materials.

Figure 3.3 Coverage probabilities (top) and interval lengths (bottom) for 95% confidence intervals for 50 domain means with $n = 400$, with assumption that the domain mean is increasing in rank, and that blocks 4 and 5 have higher means than blocks 1, 2, and 3.



Notes: The higher ranks have smaller sample sizes, so the coverage is poor without the constraints.

4. Applications

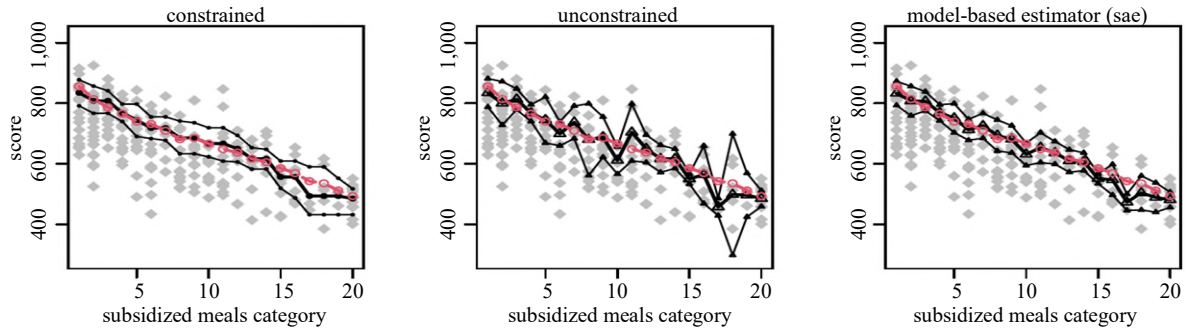
The first example uses a complete ordering. The dataset `apipop` in the `R` package `survey` contains standardized testing data from $N = 6{,}194$ California schools. We will use this data set as a population from

which we sample, to compare the performance of the constrained estimator with the standard unconstrained estimator. Because we know the population domain means, we can compare the errors of fit and also determine coverage proportions for the confidence intervals.

Suppose interest is in the average standardized test score for the school, called `api00` in the data set, and how this might be related to a measure of affluence. The variable `meals` represents the proportion of students at the school who are eligible for subsidized meals; we categorize this measure into 20 levels of five percentage points each, and assume that the average test score is decreasing as the proportion of eligible students increases.

Observations for a single sample of size $n = 240$ are shown as grey diamonds in Figure 4.1. The sample is stratified by type of school, with 60 each from elementary and middle schools, and 120 high schools. The population domain means shown as circles connected with dashed lines. On the left, the domain mean estimates constrained to be decreasing are shown as the dots connected by solid lines, and the 95% confidence interval bounds are shown as well. In the center, the unconstrained (Hajék) domain mean estimators are shown with their confidence interval bounds. The confidence interval lengths for the constrained estimators can be seen to be smaller, and for this sample, these confidence intervals all capture the population domain means. On the right are the estimates and confidence intervals for the model-based Fay Herriot estimator as provided by the `sae` package. The lengths of the confidence intervals are smallest for this estimator, but they do not capture all of the population means.

Figure 4.1 Domain means and confidence intervals for a stratified sample in the R dataset `api`, $n = 240$.

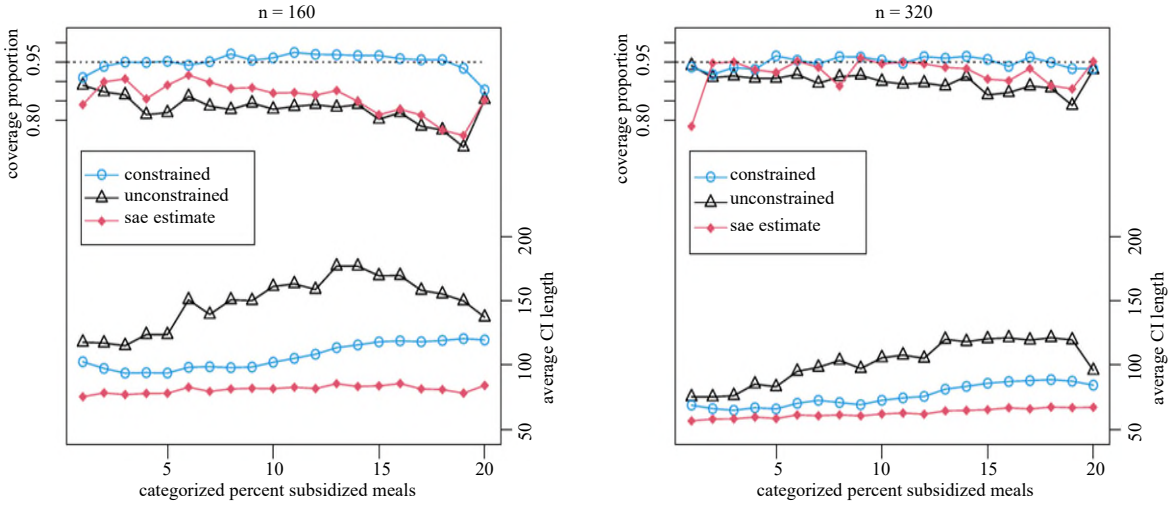


Notes: The population values are shown as circles, and the sample values as grey diamonds. The 95% confidence intervals are indicated.

The actual coverage rates and interval lengths for various sample sizes can be established through repeated sampling. We use the same sampling scheme, and for each of 1,000 samples, determine the population domain estimates and their 95% confidence intervals. The average lengths of the confidence intervals and the coverage proportions are seen in Figure 4.2, where it is seen that the coverage rates for the unconstrained estimates are lower than the target, with reasonable coverage only for larger samples. We chose $n = 160$ for a "moderate" sized sample with an average of eight observations per domain, and $n = 320$ for a "large" sample. For sample size $n = 160$, the unconstrained estimators have poor coverage, which

is improved for $n = 320$. The constrained estimator, however, has coverage proportions that are often higher than the target. In addition, the lengths for the constrained intervals are consistently smaller than those for the Hajék estimator.

Figure 4.2 Coverage probabilities and interval lengths for stratified samples of sizes 160 and 320 for 20 domains.



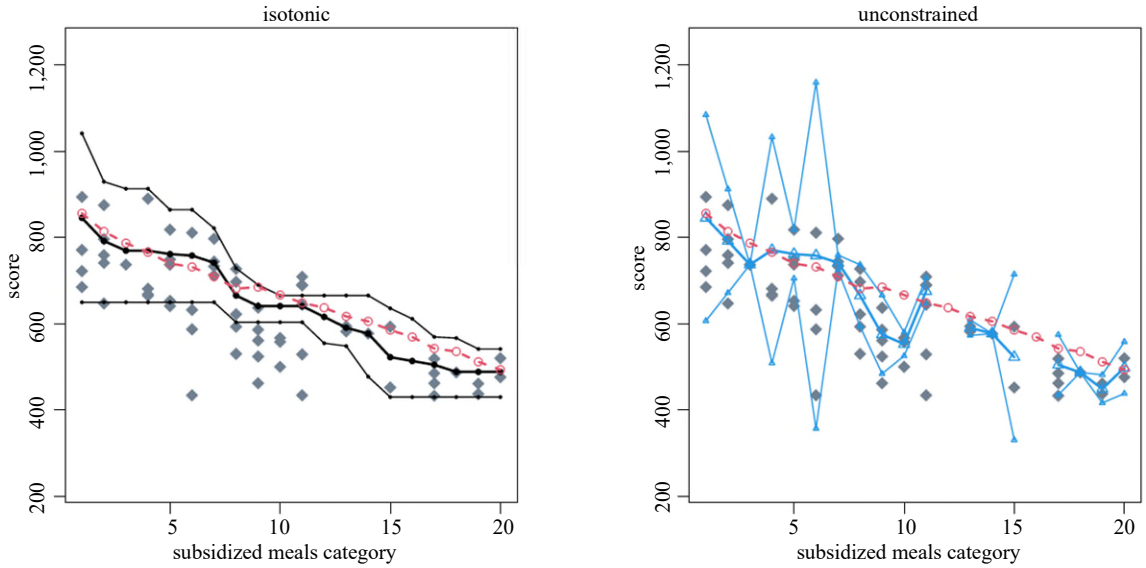
Notes: The constrained estimators have consistently better coverage and smaller lengths, compared to the unconstrained estimators.

We next choose a single stratified sample of size $n = 60$, with 15 each from elementary and middle schools, and 30 high schools. This sample size is small enough to result in empty and small-sample-size domains for each of the samples. The sample shown in Figure 4.3 has three empty domains and two domains with only one observation. For the unconstrained estimators, it is not possible to get design-based estimates for the empty domains, and the confidence intervals are unreliable for domains with small sample sizes. For the constrained estimator, information from other domains is used for the estimators of the domain means that are empty or have a small number of observations, resulting in valid confidence intervals with reasonable length. Results for repeated simulations at $n = 60$ are shown in Figure 4.4. For this sample size, there is an average of three observations per domain, but the coverage proportion is still good.

For an example of a binary study variable, we use data from the NHANES study, which provides health data for a sample of the U.S. population, and is available to the public at http://www.cdc.gov/nchs/nhanes.htm. There are $n = 1{,}680$ observations with complete records for cholesterol level, age, height, and waist size for adults ages 21-40; we will use these to demonstrate a partial ordering by estimating proportion of the population with cholesterol level above 200, by age, waist size, and gender. The waist size is divided by height so that it is a measure of relative girth, then divided into four levels, with level one the smallest and level four the largest. There are 160 domains that represent age/waist/gender combinations for a $20 \times 4 \times 2$ grid of domains. The number of domains is large for this sample size, resulting in many domains with fewer than five observations. In the absence of order assumptions, domains would have to be pooled to obtain

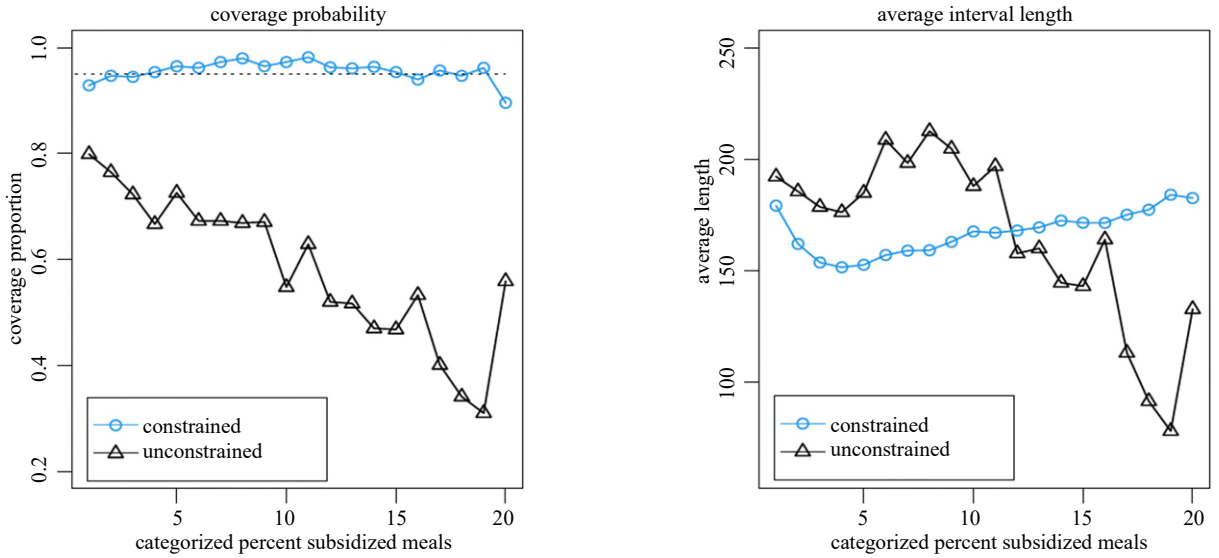
reliable estimation and inference. It is reasonable, however, to assume that the probability of having higher cholesterol will increase in both age and waist size. The subset of the NHANES data used here is included in the object `nhdat` in the `csurvey` package.

Figure 4.3 Domain means and confidence intervals for a stratified sample in the `R` dataset `api`, $n = 60$.



Notes: The population domain means are shown as circles, and the sample values are gray diamonds. The sample size is too small for design-based estimation without constraints.

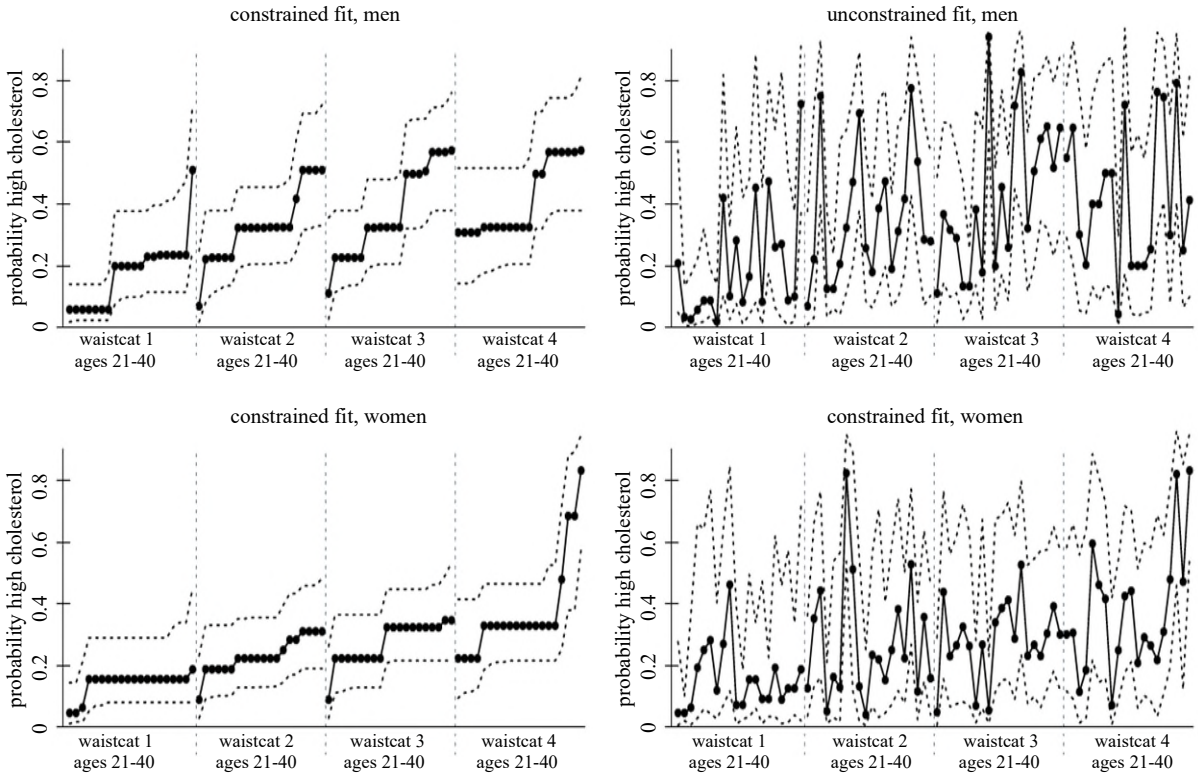
Figure 4.4 Coverage probabilities and interval lengths, for 1,000 simulations from the `apipop` data set with sample size $n = 60$.



Notes: For the unconstrained estimator, the average length and coverage probability are computed over only the non-empty domains.

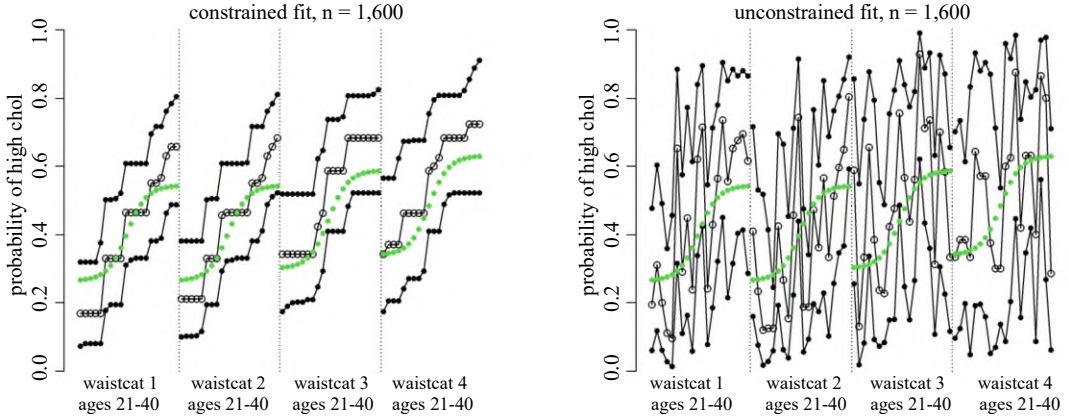
The estimates and 95% confidence intervals for the 160 domains are shown in Figure 4.5, where we see that the constrained estimates are more stable and tend to have smaller confidence intervals compared to the unconstrained Hájek estimators. (We did not include the Fay-Herriot estimators as a comparison because a binary response is not as straight-forward to implement with small sample sizes.) Although in this case we do not know the population domain proportions, it is unlikely that they "jump" up and down as age increases, within a waist category. The jumps in the unconstrained estimator are likely the result of random chance, due to small sample sizes within the 160 domains.

Figure 4.5 Estimates of the probability of high cholesterol for $D = 160$ domains, with 95% confidence intervals by age, waist size, and gender, using the NHANES data with a sample size of $n = 1{,}680$ observations.



To verify the method we carried out simulations to compare coverage rates and interval lengths for the constrained and unconstrained estimates of the probability estimates. We have 160 domains as in the NHANES example, and sample sizes of $n = 1{,}600$ and $n = 4{,}800$. We specify true probabilities of high cholesterol as shown (for one of the genders) in Figure 4.6 (the other gender has slightly higher probabilities). The sample size $n = 1{,}600$ is too small to get reasonable estimates for the traditional unconstrained estimator; researchers would aggregate the domains to get larger domain sample sizes and smaller variances.

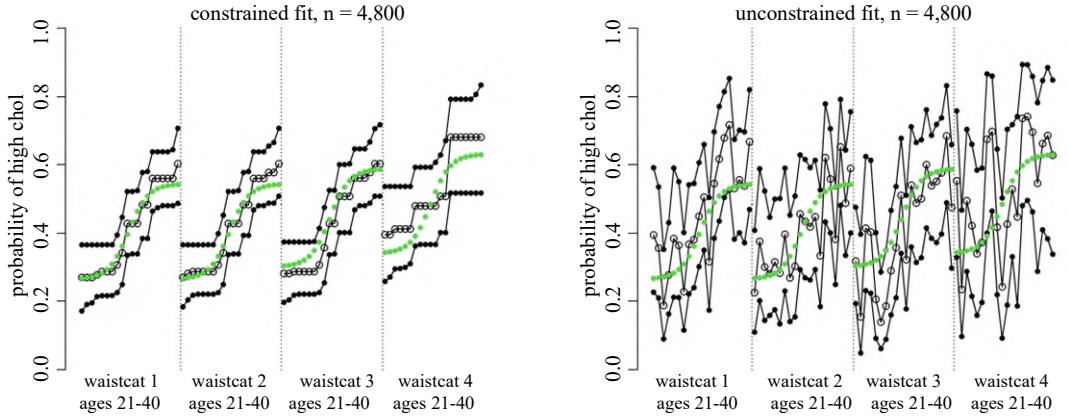
Figure 4.6 Simulated NHANES data with true probabilities of high cholesterol shown as the sigmoidal shapes.



Notes: The estimates on the left for $D = 80$ domains are constrained to be increasing in waist size and in age, while the estimates on the right are unconstrained.

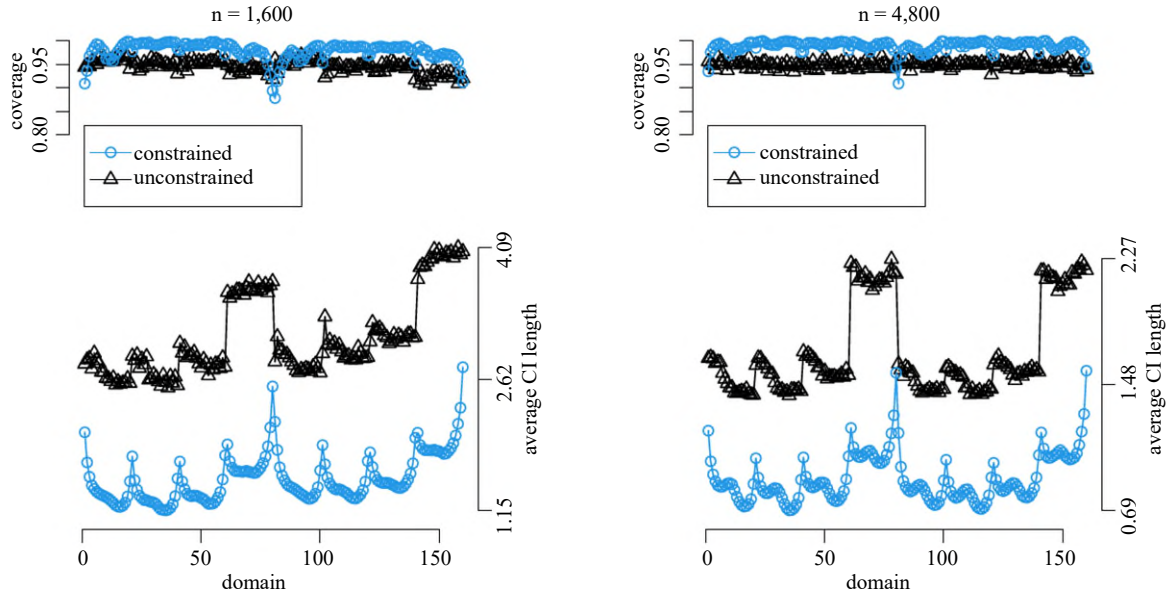
The simulated data set shown in Figure 4.7 has a larger sample size. The jumps are not as drastic as for the smaller sample size, but the constrained estimator still produces more reasonable estimates and smaller confidence intervals. For the smaller sample size, many of the domain sample sizes are small; too small for the unconstrained fit to give much information. Imposing the constraints results in more precise estimates and tighter confidence intervals. The simulations results in Figure 4.8 show the coverage proportions and the lengths of the confidence intervals for the log odds of high cholesterol in each of the 160 domains; imposing the constraints gives a dramatic reduction in confidence interval length. The constrained estimator performs best for the domains that are in the "middle" of the partial ordering; that is, if the domain mean is constrained to be both lower than some domain main means and higher than others. The "edge" domains give coverage somewhat below the target, and larger confidence intervals.

Figure 4.7 Simulated NHANES data with true probabilities of high cholesterol shown as the sigmoidal shapes.



Notes: The estimates on the left are constrained to be increasing in waist size and in age, while the estimates on the right are unconstrained. This larger data set shows more precise estimates.

Figure 4.8 Coverage proportions and confidence interval lengths for the log-odds of high cholesterol, for $D = 160$ domains, over 1,000 simulated data sets.



Notes: The constrained estimates have consistently smaller lengths.

5. Discussion

We have introduced a novel method for design-based estimation and inference of domain means with small or zero sample size, given *a priori* inequality constraints. Estimation and inference for domain means with survey data can be substantially improved if estimators based on natural orderings and the mixture variance estimator are used, and the improvement is larger for smaller sample sizes. These estimators use information from other domains in a design-based approach. The constrained methods were introduced by Wu, Meyer and Opsomer (2016) and Oliva-Aviles, Meyer and Opsomer (2020), who emphasized that the order assumptions were imposed on an imaginary "super-population" or mechanism that generates the finite population, so that the finite population itself might not exactly satisfy the ordering. They showed through simulations that if the population contains small deviations from the ordering (as in the California school example), the inference is still improved over the unconstrained estimator. We have extended these methods, providing reliable estimates and confidence intervals for small sample size or empty cells. The simulations show that the confidence intervals computed with the proposed methods give consistently good coverage compared to the standard Hajék estimator and the Fay-Herriot estimator, and the length of the confidence intervals are smaller than for the Hajék estimator. The `csurvey` package implements these methods, allowing users to specify orderings on grids of domains, and obtain estimates of and confidence intervals for population domain means. The utility of the methods has been demonstrated with well-known survey data sets.

Acknowledgements

This work was partially funded by NSF-MMS 1533804.

Appendix

A. Code for California school data example

The data set `api` in the package `survey` contains information about elementary, middle, and high schools in California. The unit is school and for this example we are concerned with the average standardized test score for the year 2000, `api00`. We expect the average scores to be decreasing over 20 levels of the `meals` variable, the proportion of students qualifying for free or reduced-price lunch.

```
mcat = apipop$meals
M = 20
for(i in 1:M){mcat[trunc(apipop$meals / 5) + 1 == i] = i}
mcat[mcat == 100] = M
mcat = as.factor(mcat)
```

For purposes of comparison, we compute the true population domain means:

```
tsc = 1:M
for(i in 1:M){tsc[i] = mean(apipop$api00[mcat == i])}
```

The `stype` variable indicates the type of school; we will choose a stratified sample based on this variable. The `snum` variable is the school ID number; the following code chooses a simple random sample from each school type, with 60 each from elementary and middle schools, and 120 from high schools.

```
nsp = c(60, 60, 120)
es = sample(apipop$snum[apipop$stype == "E" & !is.na(apipop$avg.ed) &
!is.na(apipop$api00)], nsp[1])
ms = sample(apipop$snum[apipop$stype == "M" & !is.na(apipop$avg.ed) &
!is.na(apipop$api00)], nsp[2])
hs = sample(apipop$snum[apipop$stype == "H" & !is.na(apipop$avg.ed) &
!is.na(apipop$api00)], nsp[3]) sid = c(es, ms, hs)
```

The probability weights and the finite population correction are computed next: 6,194 is the total number of schools in the data frame, of which there are 4,421 elementary schools, 1,018 middle schools, and 755 high schools.

```
pw = 1:6194 * 0 + 4421 / nsp[1]
pw[apipop$stype == "M"] = 1018 / nsp[2]
pw[apipop$stype == "H"] = 755 / nsp[3]
fpc = 1:6194 * 0 + 4421
fpc[apipop$stype == "M"] = 1018
fpc[apipop$stype == "H"] = 755
```

The design is specified using the functions svydesign and as.svrepdesign in the survey package.

```
strsamp = cbind(apipop, mcat, pw, fpc)[sid, ]
dstrat = svydesign(ids = ~snum, strata = ~stype, fpc = ~fpc, data = strsamp, weight
= ~pw)
rds = as.svrepdesign(dstrat, type = "JKn")
```

For more information about the design specification, see Lumley (2004), Lumley (2010), and Lumley (2023).

To get the proposed constrained domain mean estimate, we use the csvy function in the csurvey package. In this example, the decr function is used to constrain the domain means of api00 to be decreasing for larger values of mcat. Arguments in the csvy function are similar to those required by the svyglm function in the R package survey. An additional argument is nD, which specifies total number of domains in a data set. The user must provide this argument such that the csvy function can do estimation and inference for empty domains.

```
ans = csvy(api00 ~ decr(mcat), design = rds, nD = M)
```

CIC value for the constrained and unconstrained estimator can be extracted as

```
ans$CIC
ans$CIC.un
```

A smaller CIC shows a better fit.

The confint function can be used to extract the confidence intervals for domain mean estimates from the object ans. The svyby and svymean functions in the survey package are used to get the unconstrained domain mean estimate together with the standard error.

```
cstr = confint(ans, level = 0.95, type = "link")
unc = svyby(formula = ~api00, by = ~mcat, design = rds, FUN = svymean, covmat =
TRUE)
```

The mseFH function in the sae package is used to get the Fay-Herriot estimate together with the standard error. We need to provide the unconstrained estimator and standard error from the svyby function in the survey package as the input values of the mseFH function.

```
mhatu = unc$y
seu = unc$se
ysae = mhatu
doms = expand.grid(1:10,1:5)
x1sae = doms[,1]
x2sae = doms[,2]
anss = mseFH(ysae ~ x1sae*factor(x2sae), vardir = seu^2)
```

The ebp function in the emdi package is used to get the Empirical Best Prediction estimate by mol10 together with parametric bootstrapping standard error. We need to provide a population data set and a sample

data set. We also need to specify the name of a variable that indicates domains in the population data and the sample data.

```
emdi_model = ebp(fixed = y ~ x1*factor(x2), pop_data = pop,
      pop_domains = "domain", smp_data = sample.stsi, smp_domains = "domain",
      MSE = TRUE, seed = NULL, na.rm = TRUE)
```

B. Code for NHANES data example

The data set `nhdat` in the package `csurvey` is a subset of the data collected in the National Health and Nutrition Examination Survey (NHANES), which combines in-person interviews and physical examinations to produce a comprehensive data set from a probability sample of residents of the U.S. Included in `nhdat` are observations from 1,680 subjects. We use this data set to estimate the probability of an individual having high cholesterol level with the assumption that the average cholesterol level will increase when age and waist size increase, but we have no ordering of gender. The response variable `chol` is coded as 1 if someone's cholesterol level is larger than 200 mg/dl and 0 otherwise. Age is categorized and it takes integer values in between 21 and 40. Another variable `wcat` categorizes the ratio of an individual's waist size and height. It has 4 categories and the 3 cut-off threshold values are .48, .55 and .66. Another covariate `gender` is coded as 1 and 2, where 1 represents male and 2 represents female.

After we import the data set from `csurvey`, we use the `svydesign` function to specify a stratified sampling design with `str` being the stratas:

```
library(csurvey)
data(nhdat)
dstrat = svydesign(ids = ~ id, strata = ~ str, data = nhdat, weight = ~ wt)
```

To get the constrained estimator, we use the symbolic function `incr` twice as in `incr*incr` to specify that the domain mean of cholesterol level, i.e., the probability of getting high cholesterol, is increasing in both `age` and `wcat`, and the effects are not expected to be additive. When the response is binary, we specify `family = quasibinomial(link = "logit")` in `csvy`. Here, we use `family = quasibinomial(link = "logit")` for the same reason to avoid a warning about non-integer numbers of successes, which is recommended by author for the `svyglm` function in `survey` package. Finally, the total number of domains will be $M = 160$ and we need to provide it for estimating empty domains.

```
M = 160
ans = csvy(chol ~ incr(age) * incr(wcat) * gender, design = dstrat, nD = M,
family = quasibinomial(link = "logit"))
```

CIC value for the constrained and unconstrained estimator can be extracted as

```
ans$CIC
ans$CIC.un
```

A smaller CIC shows a better fit.

To predict the probability of a person with a set of characteristics falling in the high cholesterol group, we call the `predict` function. The arguments are similar to those of the `predict.glm` function. For example, if we want to predict the probability of a person whose `age` = 40, `wcat` = 4 and `gender` = 2, we create a new data frame containing these characteristics and provide it to the `predict` function as:

```
pred.muhat = predict(ans, newdata = data.frame(age = 40, wcat = 4, gender = 2),
      type = "response", se.fit = FALSE)
```

References

Brunk, H.D. (1958). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 29(2), 437-454.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(8), 1-19.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. New York: John Wiley & Sons, Inc.

Lumley, T. (2023). survey: Analysis of complex survey samples. *R package*.

Oliva-Aviles, C., Meyer, M.C. and Opsomer, J.D. (2019). Checking validity of monotone domain mean estimators. *Canadian Journal of Statistics*, 47(2), 315-331.

Oliva-Aviles, C., Meyer, M.C. and Opsomer, J.D. (2020). Estimation and inference of domain means subject to qualitative constraints. *Survey Methology*, 46, 2, 145-180. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020002/article/00002-eng.pdf.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

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

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley & Sons, Inc.

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

Wu, J., Meyer, M.C. and Opsomer, J.D. (2016). Survey estimation of domain means that respect natural orderings. *Canadian Journal of Statistics*, 44(4), 431-444.

Xu, X., Meyer, M.C. and Opsomer, J.D. (2021). Improved variance estimation for inequality-constrained domain mean estimators using survey data. *Journal of Statistical Planning and Inference*, 215, 47-71.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 323-345
Statistics Canada, Catalogue No. 12-001-X

A small area estimation approach for reconciling differences in two surveys of recreational fishing effort

Teng Liu, F. Jay Breidt and Jean D. Opsomer[1]

Abstract

Many studies face the problem of comparing estimates obtained with different survey methodology, including differences in frames, measurement instruments, and modes of delivery. The problem arises in multimode surveys and in surveys that are redesigned. Major redesign of survey processes could affect survey estimates systematically, and it is important to quantify and adjust for such discontinuities between the designs to ensure comparability of estimates over time. We propose a small area estimation approach to reconcile two sets of survey estimates, and apply it to two surveys in the Marine Recreational Information Program (MRIP), which monitors recreational fishing along the Atlantic and Gulf coasts of the United States. We develop a log-normal model for the estimates from the two surveys, accounting for temporal dynamics through regression on population size and state-by-wave seasonal factors, and accounting in part for changing coverage properties through regression on wireless telephone penetration. Using the estimated design variances, we develop a regression model that is analytically consistent with the log-normal mean model. We use the modeled design variances in a Fay-Herriot small area estimation procedure to obtain empirical best linear unbiased predictors of the reconciled estimates of fishing effort (requiring predictions at new sets of covariates), and provide an asymptotically valid mean square error approximation.

Key Words: Coverage error; EBLUP; Fay-Herriot model; Log-normal model; MSE approximation; Nonsampling error.

1. Introduction

For decades, the National Marine Fisheries Service (NMFS) has conducted household surveys to count the number of recreational saltwater fishing trips from shore and private boat anglers in 17 US states along the coasts of the Atlantic Ocean and the Gulf of Mexico: Alabama, Connecticut, Delaware, Florida, Georgia, Louisiana, Maine, Maryland, Massachusetts, Mississippi, New Hampshire, New Jersey, New York, North Carolina, Rhode Island, South Carolina, and Virginia. Data collection occurs during a two-week period at the end of each two-month sample period (or "wave"), yielding six waves for each year. However, samples are not obtained for every wave in every state; for example, many states have no wave 1 sample, reflecting minimal fishing effort during January and February in those states.

Until 2017, NMFS used the Coastal Household Telephone Survey (CHTS) to collect trip data. The CHTS frame was a list of full-time residential households with landline telephone service in coastal counties. The design was stratified simple random sampling, stratified by state and county. The CHTS used random digit dialing (RDD) for landlines of households in coastal counties. RDD suffers from several shortcomings in this context, such as the inefficiency at identifying anglers (National Research Council, 2006), the declining response rate for telephone surveys (Curtin, Presser and Singer, 2005), and the undercoverage of anglers due to the increase in wireless-only households (Blumberg and Luke, 2013). Thus, after some experimentation (Andrews, Brick and Mathiowetz, 2014), NMFS implemented the new Fishing Effort Survey (FES) in 2015.

1. Teng Liu, Colorado State University, Fort Collins, U.S.A. E-mail: tristan.tju@gmail.com; F. Jay Breidt, NORC at the University of Chicago, Chicago, U.S.A; Jean D. Opsomer, Westat, Inc., Rockville, U.S.A.

Unlike CHTS, the FES is a self-administered mail survey that uses as its frame a directory of postal addresses in coastal states (not just coastal counties) serviced by the US Postal Service. In recent years, many studies have followed this same path from telephone to self-administered modes; see Olson, Smyth, Horwitz, Keeter, Lesser, Marken, Mathiowetz, McCarthy, O'Brien, Opsomer, Steiger, Sterrett, Su, Suzer-Gurtekin, Turakhia and Wagner (2020) for a recent review. The FES design is stratified simple random sampling, stratified by state, proximity to the coast, and fishing license status, with status determined by matching addresses to the National Saltwater Angler Registry. The CHTS was discontinued after 2017, so that the two surveys have overlap in 2015-2017.

The telephone-based CHTS and the mail-based FES have obvious methodological differences. The two surveys have different coverage properties because they use very different frames: RDD of landlines for CHTS versus address-based sampling, with oversampling of addresses matched to licensed anglers, for FES. They have different nonresponse patterns, with overall FES response rates nearly three times higher than CHTS response rates (Andrews et al., 2014). Finally, the measurement processes are fundamentally different, due to the differences in asking about angling activity over the phone versus a self-administered paper form.

Due at least in part to these methodological differences, there is a large discrepancy between the trip estimates from the CHTS and the FES estimates, with FES estimates consistently higher. As we do not believe that either survey reflects the true number of trips exactly, whatever the reasons for the discrepancy, it is of interest to fisheries managers and stock assessment scientists to be able to convert from the "units" of the telephone survey estimates to those of the mail survey estimates, and vice versa. This conversion is known as "calibration" in this context, and is not to be confused with the calibration method common in complex surveys (Deville and Särndal, 1992). The calibration allows construction of a series of comparable estimates across time.

The data used for the calibration exercise come from the CHTS for most states and waves from 1981 to 2017, and from the FES for states and waves from 2015 to 2018. In what follows, we work on the scale of natural logarithms of trip counts, and refer to this log count as "effort." For each survey, the data consist of estimated total effort for shore fishing and total effort for private boat fishing, along with estimated design variances and sample sizes, for each available state and wave.

As discussed below, we formulate the calibration problem as an application of area-level small area estimation, for which we briefly review some related literature. Rao and Yu (1994) propose a small area estimation model involving autoregressive random effects and sampling errors with arbitrary known covariance matrix using both time-series and cross-sectional data. Datta, Lahiri, Maiti and Lu (1999) use a random walk model for the time component, with correlated sampling errors. Pfeffermann and Tiller (2006) add benchmark constraints to the state-space model with correlated measurements. Boonstra, van den Brakel and Das (2021); Boonstra and van den Brakel (2022) develop Bayesian hierarchical models for multilevel time series in small areas. Feder (2001) reviews various time series methods on cross-sectional survey data.

The literature on combining surveys is not as extensive as it is for a single survey. Merkouris (2010) proposes a model-assisted estimator by calibrating comparable domain estimates from various non-repeated surveys sharing the same domains. Lohr and Brick (2012) adopt a dual frame survey approach and explore methods for small area estimation from two surveys when one may be biased. Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011) propose a series of Bayesian hierarchical models to combine prevalence estimates from multiple sources of data with additive biases. Wang, Holan, Nandram, Barboza, Toto and Anderson (2012) combine three surveys measured on different temporal supports and develop a Bayesian hierarchical model which produces better estimation of crop yield, with the assumptions that one of the three surveys analyzed is unbiased for the true yield. van den Brakel, Zhang and Tam (2020) review different methods to measure discontinuities due to a survey process redesign, classifying those methods according to whether there is an overlapping period between the old and new surveys, how long such a period lasts, and how the old survey switches to the new survey. For parallel data collection, where data is collected under the old and new designs alongside each other for a certain period, design-based methods in van den Brakel (2008, 2013), state-space models in van den Brakel (2008, 2010) and small area estimation models in Pfeffermann (2002, 2013) and Rao and Molina (2015) can be adopted, depending on the length of the parallel run and the sample sizes. Other related papers include Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer (2007), who combine information from two surveys to correct noncoverage and nonresponse issues through adopting a hierarchical Bayesian model assuming unbiasedness of one of the surveys, and Erciulescu, Opsomer and Breidt (2021), who establish a Bayesian hierarchical model to account for discrepancies between two sets of survey estimates and produce reliable estimates at various aggregation levels.

In Section 2.1, we build a model that assumes that both mail and telephone estimates have underlying "targets" of interest in the calibration. Both target series include a classical time series model consisting of trend, seasonal, and irregular components. This model specification supports calibration backward or forward in time. For a past time period, we can predict what the effort in "mail units" would have been by using the prior telephone estimate to predict the mail target. Similarly, for a future time period, we can predict what the effort would have been in "telephone units" by predicting the telephone target using the mail estimate.

In Section 2.3, we show that the combined model for the two sets of estimates and the underlying targets is a linear mixed model of a type that commonly appears in the context of area-level small area estimation, where it is known as the Fay-Herriot model (Fay and Herriot, 1979). In Fay-Herriot, it is standard to treat design variances as known. Our design variances are based on moderate to large sample sizes (minimum size $n = 39$) in each state and wave and so are well-estimated by the standards of small area estimation. A complication is that the original design variances are on the scale of trip counts rather than the scale of effort (log trip counts). As an alternative to standard Taylor linearization, we develop in Appendix B a novel approach to transforming the estimated design variances that ensures analytic consistency between our mean model and our variance model.

The Fay-Herriot methodology in Sections 3.1-3.2 leads to empirical best linear unbiased predictors (EBLUPs) of the mail target or the telephone target, and these constitute our calibrated effort series. Unlike the standard Fay-Herriot context, the EBLUPs require prediction at new sets of covariates. In Sections 3.3-3.4, we adapt standard Mean Square Error (MSE) approximations and estimates to this non-standard situation, and evaluate their performance via simulation in Section 4.1. In Section 4.2, we apply the methods to the problem of reconciling past telephone survey estimates to the mail survey, and conclude with a brief discussion in Section 5.

2. Model

2.1 Mean model

We fix attention on one type of fishing behavior, either shore or private boat: the model development is identical in both cases. Let $s = 1, 2, \ldots, 17$ index the US states and let $t = 1, 2, \ldots$ index time in two-month waves, starting with January-February of 1981. We assume that the telephone effort estimate $\hat{T}_{st}$ is a design-unbiased estimator of the "telephone target" T_{st}, which includes both the true effort and survey mode effects due to the telephone methodology, while the mail effort estimate $\hat{M}_{st}$ is a design-unbiased estimator of the "mail target" M_{st}, which includes both the true effort and survey mode effects due to the mail methodology.

We assume that both the telephone target and the mail target contain the true effort series, which is further assumed to contain state-specific trends, due in part to changing state population sizes, state-specific seasonal effects that vary wave to wave, and irregular terms that are idiosyncratic effects not explained by regular trend or seasonal patterns. We model state-specific trends by using annual state-level estimates of the population size from the US Census Bureau (2016) on a log scale. We model a general seasonal pattern via indicators for the two-month waves, and allow the seasonal pattern to vary from state to state. The remaining irregular terms, denoted $\{\nu_{st}\}$ below, represent real variation not explained by the regular trend plus seasonal pattern, and are modeled as independent and identically distributed (iid) random variables with mean zero and unknown variance, ψ.

The survey mode effects present in the telephone and mail targets are nonsampling errors, including potential biases due to coverage error (population $\neq$ sampling frame), nonresponse error (sample $\neq$ respondents), and measurement error (true responses $\neq$ measured responses). These effects may have their own trend and seasonality: for example, due to changes in the quality of the frame over time, changes in response rates over years or waves, changes in implementation of measurement protocols over time, etc. These nonsampling errors thus cannot be completely disentangled from the true effort series (a problem in every survey).

Given suitable covariates that explain the change in measurement error, nonresponse error, or coverage error over time, the nonsampling errors could be modeled and removed. The changing proportion of wireless-only households is a potential covariate for explaining changes in coverage error over time for the

landline-only telephone survey. In Appendix A, we describe how we constructed a set of predicted proportions of wireless-only households, $\{w_{st}\}$, for every state and wave in our data.

Either trend or seasonal could contain survey mode effects. Accordingly, we allow for the possibility that trend and seasonal are different for mail versus telephone, and in particular we allow for the possibility that either trend or seasonal can change with the level of wireless.

Our combined model for effort (natural logarithms of trip counts) then assumes

$$\begin{aligned} &\text{telephone effort estimate} && = \hat{T}_{st} = T_{st} + e_{Tst} \\ &\text{telephone target} && = T_{st} = \mathbf{x}_{Tst}^{\top}\boldsymbol{\beta} + \nu_{st} \\ &\text{mail effort estimate} && = \hat{M}_{st} = M_{st} + e_{Mst} \\ &\text{mail target} && = M_{st} = \mathbf{x}_{Mst}^{\top}\boldsymbol{\beta} + \nu_{st}, \end{aligned} \tag{2.1}$$

where

- $\boldsymbol{\beta}$ is a vector of unknown regression coefficients;
- the sampling errors $\{e_{Tst}\}$ are independent $\mathcal{N}\left(0, \sigma_{Tst}^2\right)$ random variables, with known design variances σ_{Tst}^2;
- the sampling errors $\{e_{Mst}\}$ are independent $\mathcal{N}\left(0, \sigma_{Mst}^2\right)$ random variables, with known design variances σ_{Mst}^2;
- the irregular terms $\{\nu_{st}\}$, representing real variation not explained by the regular trend plus seasonal pattern, are iid $\mathcal{N}(0, \psi)$ random variables, with unknown variance ψ;
- $\{e_{Tst}\}$, $\{e_{Mst}\}$ and $\{\nu_{st}\}$ are mutually independent.

The assumed independence of the sampling errors is justified because the sample is stratified and independent samples are drawn state-to-state and wave-to-wave. The assumed normality is justified by central limiting effects of moderate to large-size stratified samples in each state and wave (as previously noted, the minimum sample size is 39). Further, we assume that because the mail and telephone surveys are selected and conducted independently, the sampling errors $\{e_{Tst}\}$ and $\{e_{Mst}\}$ are independent of one another. We use simulation to assess the sensitivity of some of our methods to the normality assumption on the random effects in Section 4.1 below. The design variances $\{\sigma_{Tst}^2\}$ and $\{\sigma_{Mst}^2\}$ are for effort estimates (natural logarithms of trip count estimates), while the available design variance estimates $\{\hat{V}_{Tst}\}$ and $\{\hat{V}_{Mst}\}$ are for trip count estimates; we address this discrepancy in Section 2.2 below.

Let $\nu_{st}, \nu_{st}^T, \nu_{st}^M$ denote independent, zero-mean random effects where ν_{st} is measured by both surveys and represents true variation not explained by covariates, while ν_{st}^T, ν_{st}^M denote mode-specific random effects. We considered various random effect specifications, including (a) both surveys measure true variation (ν_{st}); (b) mail measures truth (ν_{st}) while telephone measures truth plus telephone random effect $(\nu_{st} + \nu_{st}^T)$; (c) telephone measures truth (ν_{st}) while mail measures truth plus mail random effect, $(\nu_{st} + \nu_{st}^M)$; (d) each survey measures truth plus its own mode-specific random effect, $(\nu_{st} + \nu_{st}^T, \nu_{st} + \nu_{st}^M)$; and (e) the two surveys have their own mode-specific random effects, independent of each other, with no true variation

(outside of fixed effects) measured by either (ν_{st}^T, ν_{st}^M). Independence is assumed in specifications (b)-(e) because any correlated effects in the two measurements should be true variation, not measurement error.

Both models (b) and (c) require specification of one model to serve as the "gold standard", which is inconsistent with the approach we have taken in this analysis. We do not pursue these models further. Model (d) is scientifically plausible, but it is the largest model considered and requires custom estimation software to account for both bivariate observations during overlapping periods and univariate observations during nonoverlapping periods. In some exploratory analysis not reported here, we found it difficult to fit this model with our limited overlapping data. Model (e) says that any random effects are purely mode-specific measurement error, unrelated to the true underlying phenomenon. While this is a possibility, the implication of such a model is that only the fixed effects are of interest for prediction, and even if we mis-specify the zero-mean random effects, we will still have reasonable estimates of the fixed effects. Consequently, we settled on model (a), in which both surveys measure a common ν_{st} that represents real variation. This is a standard specification in small area estimation and it allows us to obtain estimates using off-the-shelf software for univariate Fay-Herriot modeling, as we describe below.

The sampling errors in our application are independent due to the stratification in the design. The random effects $\{\nu_{st}\}$ in our specification do not include either spatial or temporal autocorrelation. The spatio-temporal scale of our data is state by wave, which we model with fixed effects. Any residual spatial or temporal autocorrelation is expected to be small, with temporal autocorrelation likely to be more important if any autocorrelation is present. However, our diagnostics (see Section 4.2) did not find support for residual temporal autocorrelation.

Let $\boldsymbol{\beta}^\intercal = \left[\boldsymbol{\alpha}^\intercal, \boldsymbol{\mu}^\intercal, \boldsymbol{\gamma}^\intercal\right]$. Then the fixed effects $\mathbf{x}_{Tst}^\intercal\boldsymbol{\beta}$ and $\mathbf{x}_{Mst}^\intercal\boldsymbol{\beta}$ can each be decomposed into three components,

$$\begin{aligned}\mathbf{x}_{Tst}^\intercal\boldsymbol{\beta} &= \left[\mathbf{a}_{st}^\intercal, 0\cdot\mathbf{b}_{st}^\intercal, w_{st}\mathbf{c}_{st}^\intercal\right]\boldsymbol{\beta}, = \mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + 0\cdot\mathbf{b}_{st}^\intercal\boldsymbol{\mu} + w_{st}\mathbf{c}_{st}^\intercal\boldsymbol{\gamma} \\ \mathbf{x}_{Mst}^\intercal\boldsymbol{\beta} &= \left[\mathbf{a}_{st}^\intercal, 1\cdot\mathbf{b}_{st}^\intercal, 0\cdot\mathbf{c}_{st}^\intercal\right]\boldsymbol{\beta} = \mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + 1\cdot\mathbf{b}_{st}^\intercal\boldsymbol{\mu} + 0\cdot\mathbf{c}_{st}^\intercal\boldsymbol{\gamma},\end{aligned} \tag{2.2}$$

where the known covariate vector $\mathbf{a}_{st}$ includes intercept, log(state population size), state indicators, wave indicators, and state by log(population) and state by wave interactions. In our application, the covariate vectors $\mathbf{b}_{st}$ and $\mathbf{c}_{st}$ are subvectors of $\mathbf{a}_{st}$, due to parsimony; details are provided in Section 4.2. Hence, $\mathbf{a}_{st}^\intercal\boldsymbol{\alpha}$ describes state-specific trend and seasonal variation for the telephone data, $\mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + \mathbf{b}_{st}^\intercal\boldsymbol{\mu}$ describes state-specific trend and seasonal variation for the mail data, and the wireless interaction term $w_{st}\mathbf{c}_{st}^\intercal\boldsymbol{\gamma}$ models the impact of wireless telephone penetration on the telephone trend and seasonal and has no impact on mail. It is easy to verify that all parameters in this model are identified and admit unbiased estimates from the available data.

2.2 Design variance model

Under the effort models (2.1), the variances of the sampling errors on the original scale of untransformed trip counts can be derived from the log-normal distribution as

$$V_{Tst} = \mathrm{Var}\left(\exp(\hat{T}_{st}) \mid T_{st}\right) = \left\{\exp(\sigma^2_{Tst}) - 1\right\} \exp\left\{2T_{st} + \sigma^2_{Tst}\right\} \tag{2.3}$$

and

$$V_{Mst} = \mathrm{Var}\left(\exp(\hat{M}_{st}) \mid M_{st}\right) = \left\{\exp(\sigma^2_{Mst}) - 1\right\} \exp\left\{2M_{st} + \sigma^2_{Mst}\right\}. \tag{2.4}$$

We need to estimate σ^2_{Tst} and σ^2_{Mst} (the design variances on the scale of effort, or log trip counts), incorporating the approximately design-unbiased estimates $\hat{V}_{Tst}$ and $\hat{V}_{Mst}$ of V_{Tst} and V_{Mst}, respectively.

We follow an approach related closely to generalized variance function estimation (e.g., Chapter 7 of Wolter (2007)) by developing regression models for the logarithms of the empirical CV^2 (squared coefficients of variation) and using these fitted models to generate estimates of the design variances on the log scale, σ^2_{Tst} and σ^2_{Mst}, that enforce analytical consistency between the mean model and the variance model. Details are provided in Appendix B. The sample sizes within states and waves are large in our application, so we treat these estimates as fixed and known in what follows, as is standard in the small area estimation techniques which we will apply in subsequent sections.

2.3 Fay-Herriot small area estimation model

Define

$$\mathbf{x}^{\mathsf{T}}_{st} = \begin{cases} \mathbf{x}^{\mathsf{T}}_{Tst}, & \text{if no mail estimate is available;} \\ \mathbf{x}^{\mathsf{T}}_{Mst}, & \text{if no telephone estimate is available;} \\ (\mathbf{x}_{Tst} + \mathbf{x}_{Mst})^{\mathsf{T}} / 2, & \text{otherwise.} \end{cases}$$

Then it is convenient to write

$$\begin{aligned} Y_{st} &= \begin{cases} \hat{T}_{st}, & \text{if no mail estimate is available;} \\ \hat{M}_{st}, & \text{if no telephone estimate is available;} \\ \left(\hat{T}_{st} + \hat{M}_{st}\right)/2, & \text{otherwise.} \end{cases} \\ &= \begin{cases} \mathbf{x}^{\mathsf{T}}_{Tst}\boldsymbol{\beta} + \nu_{st} + e_{Tst}, & \text{if no mail estimate is available;} \\ \mathbf{x}^{\mathsf{T}}_{Mst}\boldsymbol{\beta} + \nu_{st} + e_{Mst}, & \text{if no telephone estimate is available;} \\ (\mathbf{x}_{Tst} + \mathbf{x}_{Mst})^{\mathsf{T}}\boldsymbol{\beta}/2 + \nu_{st} + (e_{Tst} + e_{Mst})/2, & \text{otherwise.} \end{cases} \\ &= \mathbf{x}^{\mathsf{T}}_{st}\boldsymbol{\beta} + \nu_{st} + e_{st}. \end{aligned} \tag{2.5}$$

This model then follows exactly the linear mixed model structure of Fay and Herriot (1979), with direct estimates Y_{st} equal to regression model plus random effect ν_{st} plus sampling error with "known" design variance, given by

$$D_{st} = \begin{cases} \sigma^2_{Tst}, & \text{if no mail estimate is available;} \\ \sigma^2_{Mst}, & \text{if no telephone estimate is available;} \\ \frac{1}{4}\left(\sigma^2_{Tst} + \sigma^2_{Mst}\right), & \text{otherwise.} \end{cases}$$

It might seem natural to use a convex combination other than $(1/2, 1/2)$ to reflect unequal variances in the two data sources. Simple averaging could result in nontrivial loss of information for prediction of ν_{st}, but because our goal is calibration, prediction of ν_{st} is not required during the overlap period: we have both telephone and mail observations and no need for unit conversion. The only contribution of these overlap observations is therefore to the estimation of the regression parameters $\boldsymbol{\beta}$ and random effect variance ψ. Averaging the telephone and mail estimates results in a small loss of information for parameter estimation, since we are replacing two correlated observations with one observation, but this simple approach allows the use of standard software for estimation.

3. Methods

3.1 Estimation for the Fay-Herriot Model

Define $\mathcal{A} = \{(s,t)\colon Y_{st} \text{ is not missing}\}$ to be the set of all state by year-wave combinations for which we have an estimate from either survey. Let m denote the size of the set $\mathcal{A}$. Define $\mathbf{X} \coloneqq [\mathbf{x}^{\mathsf{T}}_{st}]_{(s,t)\in\mathcal{A}}$, $\mathbf{Y} \coloneqq [Y_{st}]_{(s,t)\in\mathcal{A}}$. We have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + [\nu_{st}]_{(s,t)\in\mathcal{A}} + [e_{st}]_{(s,t)\in\mathcal{A}}.$$

Then $\boldsymbol{\Sigma}(\psi) \coloneqq \mathrm{Var}(\mathbf{Y}) = \mathrm{diag}\{\psi + D_{st}\}_{(s,t)\in\mathcal{A}}$. If ψ were known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ would be

$$\tilde{\boldsymbol{\beta}}_{\psi} = \left\{\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\,\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\,\mathbf{Y}. \tag{3.1}$$

Since ψ is not known, we replace it by a consistent estimator to obtain

$$\hat{\boldsymbol{\beta}} = \left\{\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\,\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\,\mathbf{Y}. \tag{3.2}$$

We will use the Restricted Maximum Likelihood (REML) estimate $\hat{\psi}$ unless otherwise indicated.

3.2 Prediction

In the classical Fay-Herriot context, it is of interest to predict

$$\mathbf{x}^{\mathsf{T}}_{st}\boldsymbol{\beta} + \nu_{st}$$

from (2.5). In our setting, however, we seek to predict

$$\phi_{st} = \mathbf{z}_{st}^{\top}\boldsymbol{\beta} + \nu_{st}, \tag{3.3}$$

where $\mathbf{z}_{st}$ may not equal $\mathbf{x}_{st}$. As noted in the introduction, it is of interest to convert from the "units" of one mode to those of the other mode. To convert a past telephone survey estimate to mail survey units, we can use

$$\mathbf{z}_{st}^{\top} = \mathbf{x}_{Mst}^{\top} = \left[\mathbf{a}_{st}^{\top}, \mathbf{b}_{st}^{\top}, \mathbf{0}^{\top}\right]$$

to predict the mail target M_{st}. To convert a future mail survey estimate to historic telephone survey units, we may want to use

$$\mathbf{z}_{st}^{\top} = \left[\mathbf{a}_{st}^{\top}, \mathbf{0}^{\top}, \mathbf{0}^{\top}\right]$$

to predict the telephone target, corrected for the wireless effect: $T_{st} - w_{st}\mathbf{c}_{st}^{\top}\boldsymbol{\gamma} = \mathbf{a}_{st}^{\top}\boldsymbol{\alpha} + \nu_{st}$.

Let $\boldsymbol{\lambda}_{st}$ denote an $m \times 1$ vector with a one in the (s,t) th position and zero elsewhere. Under normality, it is well-known that the best mean square predictor of ϕ_{st} in (3.3) is

$$\phi_{st}\left(\boldsymbol{\beta}, \psi\right) = \mathbf{z}_{st}^{\top}\boldsymbol{\beta} + \psi\boldsymbol{\lambda}_{st}^{\top}\boldsymbol{\Sigma}^{-1}(\psi)\left(\mathbf{Y} - X\boldsymbol{\beta}\right), \tag{3.4}$$

which is feasible only if both $\boldsymbol{\beta}$ and ψ are both known. If only ψ is known, the best linear unbiased predictor (BLUP)

$$\phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right) = \mathbf{z}_{st}^{\top}\tilde{\boldsymbol{\beta}}_{\psi} + \psi\boldsymbol{\lambda}_{st}^{\top}\boldsymbol{\Sigma}^{-1}(\psi)\left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\psi}\right) \tag{3.5}$$

is obtained by plugging the BLUE from (3.1) into (3.4). Finally, if neither $\boldsymbol{\beta}$ nor ψ is known, then the empirical best linear unbiased predictor (EBLUP) can be obtained by substituting a consistent estimator of ψ into (3.5):

$$\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) = \mathbf{z}_{st}^{\top}\hat{\boldsymbol{\beta}} + \hat{\psi}\,\boldsymbol{\lambda}_{st}^{\top}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{3.6}$$

where $\hat{\boldsymbol{\beta}}$ is given by (3.2). These EBLUPs are the proposed calibrated values on the log scale.

3.3 Mean square error approximation

Mean square error approximation has been investigated widely; see Jiang and Lahiri (2006) for an excellent review. Our prediction approach is slightly nonstandard because we predict at new sets of covariates when converting our estimates from the "units" of one mode to those of the other mode. It can be shown that

$$\begin{aligned}
\mathrm{MSE}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\} &= E\left[\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\right\}^2\right] \\
&= E\left[\left\{\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right\}^2\right] + E\left[\left\{\phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right\}^2\right] \\
&\quad + E\left[\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right\}^2\right] \\
&= \dot{g}_{1st}(\psi) + \dot{g}_{2st}(\psi) + \dot{g}_{3st}(\psi) + o\left(m^{-1}\right),
\end{aligned} \tag{3.7}$$

where m is the number of small areas,

$$\dot{g}_{1st}(\psi) = \frac{\psi D_{st}}{\psi + D_{st}},$$

$$\dot{g}_{2st}(\psi) = \left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\mathsf{T}} + D_{st}\mathbf{z}_{st}^{\mathsf{T}}}{\psi + D_{st}} \right) \left[\sum_{u \in \mathcal{A}} (\psi + D_u)^{-1} \mathbf{x}_u \mathbf{x}_u^{\mathsf{T}} \right]^{-1}$$

$$\times \left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\mathsf{T}} + D_{st}\mathbf{z}_{st}^{\mathsf{T}}}{\psi + D_{st}} \right)^{\mathsf{T}},$$

and

$$\dot{g}_{3st}(\psi) = \frac{2D_{st}^2}{(\psi + D_{st})^3} \frac{1}{\sum_{u \in \mathcal{A}} (\psi + D_u)^{-2}}.$$

The terms in this MSE approximation can be obtained from results in Section 5.3 of Rao and Molina (2015).

3.4 Mean square error estimation

As in Section 5.3 of Rao and Molina (2015), an approximately unbiased estimator of the MSE approximation in (3.7) is given by

$$\text{mse}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\} = \dot{g}_{1st}(\hat{\psi}) + \dot{g}_{2st}(\hat{\psi}) + 2\dot{g}_{3st}(\hat{\psi}). \tag{3.8}$$

We assess the quality of the asymptotic approximation (3.7) and its estimator (3.8) via simulation in Section 4.1 .

3.5 Prediction on the original scale

To compute predictors on the original scale, we back-transform by exponentiating the EBLUP from (3.6) and adjust for the nonlinearity of the back-transformation using the estimated MSE from (3.8):

$$\widehat{\exp(\phi_{st})} = \exp\left[\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) + \frac{1}{2}\text{mse}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\}\right], \tag{3.9}$$

which is an estimator of the best mean square predictor under the normal model, and a standard adjustment even without the normality assumption. The back-transformation of Slud and Maiti (2006) uses the leading term, $\dot{g}_{1st}(\hat{\psi})$, of (3.8) and is nearly identical to (3.9) in our application because the impact of parameter estimation is small.

4. Empirical results

4.1 Simulation

In this section, we investigate the performance of our second-order approximation of MSE and the estimated MSE under a setting that mimics the reconciliation problem of this paper. We use a subset of the

original data consisting of shore mode for all 17 states and seven years: the three overlap years 2015-2017 and four previous years (1985, 1995, 2005, 2010). By design, there are some missing state and wave combinations (e.g., January-February in Maine). There are 607 telephone estimates and 258 mail estimates, among which 257 have the same state and wave combinations as telephone. Hence, the number of small areas in this setting is $m = 607 + 258 - 257 = 608$. We take the wireless values and population counts from the actual data.

We use the covariates stated above to construct the design matrix and multiply it by the parameter estimates from the final model fitted by all the shore mode data as described in Section 4.2 to get the fixed effects as in (3.3).

Following Datta, Rao and Smith (2005), we consider three distributions to simulate the normalized random effects:

- $\{\psi^{-1/2}\nu_{st}\}$ iid $\mathcal{N}(0,1)$;
- $\{\psi^{-1/2}\nu_{st}\}$ iid Laplace$(0,\ 1/\sqrt{2})$;
- $\{\psi^{-1/2}\nu_{st}\}$ iid centered Exponential(1) (that is, exponential random variables centered to mean zero).

Under each distribution, $\mathrm{E}[\nu_{st}] = 0$ and $\mathrm{Var}(\nu_{st}) = \psi$. We pick $\psi = 0.12$, again from the fit of the model. We get true efforts by adding the random effects to the fixed effects as in (3.3).

We consider three different patterns for the design variances $\{D_{st}\}$. First, we use the modeled design variances in Section 2.2 as true design variances to create pattern (b). We consider two additional settings, by multiplying pattern (b) by 0.5 to yield pattern (a), and multiplying pattern (b) by 2.0 to yield pattern (c). The simulated sampling errors $\{e_{st}\}$ in (2.5) are then generated independently as $\mathcal{N}(0, D_{st})$ under each pattern.

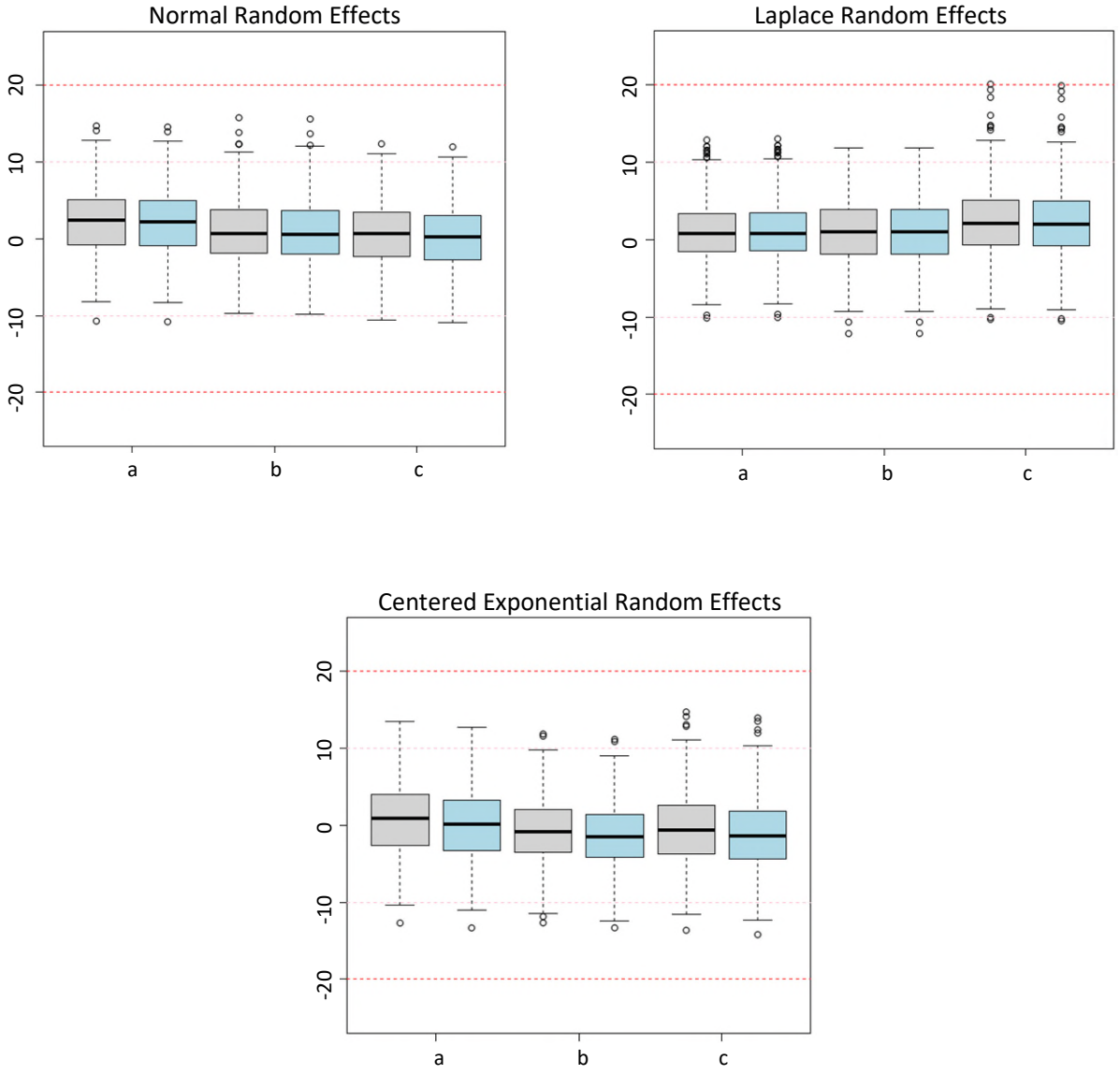
For each combination of sampling variance pattern and random effect distribution, we generate 1,000 data sets from model (2.5). For each simulated data set, we use the `R` package `sae` Molina and Marhuenda (2015) to compute $\hat{\psi}$ via REML and $\hat{\boldsymbol{\beta}}$. We compute the EBLUPs in (3.6) for the mail targets $\{M_{st}\}$, approximate their MSEs using (3.7), and estimate their MSEs using (3.8). We then compare the approximations and the estimates to the true (Monte Carlo) MSEs over the 1,000 simulated realizations.

The simulation model is similar to the final model selected in Section 4.2 below except for removing certain nonexisting state and wave combinations in the subset of covariates.

Figure 4.1 shows boxplots of the relative error (in percent) of the MSE approximation (3.7) and the Monte Carlo average of the MSE estimator (3.8) relative to Monte Carlo MSE evaluated with 1,000 replicates. Each boxplot consists of relative error from the 608 state and wave combinations for normal, Laplace, or centered exponential random effects with sampling error pattern (a), (b) or (c) described above. As expected, the MSE estimator is nearly unbiased for the MSE approximation in all cases, so the pairs of boxplots are nearly indistinguishable at each setting. Across all settings, the MSE approximation is close to

the true MSE (as measured by Monte Carlo), hence most of the relative errors are close to zero and within the ±10% reference lines, with a few outside ±10% but within ±20%.

Figure 4.1 Boxplots of the relative error (in percent) of the MSE approximation (in light gray) and the Monte Carlo average of the MSE estimator (in light blue) relative to Monte Carlo MSE evaluated with 1,000 replicates. Each boxplot consists of relative error from the 608 state and wave combinations for normal, Laplace, or centered exponential random effects with sampling error pattern (a), (b) or (c) as described in the text. Horizontal lines at ±10% and ±20% are drawn for reference.



4.2 Calibration of the CHTS and FES estimates

For the data described in Section 1, we use the `R` package `sae` (Molina and Marhuenda, 2015) to fit a number of models via maximum likelihood for both shore fishing and private boat fishing, and compare the models via their AIC values. The smallest model considered includes intercept, log(population), state indicators, wave indicators, state by log(population) interaction, and state by wave interaction. That is, the smallest model includes no differences due to survey methodology and instead drops the terms $\mathbf{b}_{st}^{\top}\boldsymbol{\mu}$ and $w_{st}\mathbf{c}_{st}^{\top}\boldsymbol{\gamma}$ from (2.1). The largest model considered adds wireless and its interactions with log(population), state indicators, wave indicators, and state by log(population), together with an indicator for presence of a mail survey estimate and the mail indicator's interactions with log(population), state indicators, and wave indicators.

The largest model adds two main effects and seven interactions to the smallest model. We considered 80 submodels between the smallest and largest, each starting from the smallest model plus wireless and mail main effects. The six two-way interactions were then included or excluded, leading to $2^6 = 64$ possible models, and an additional $2^4 = 16$ models were considered by including the three-way interaction, wireless by state by log(population), and the corresponding two-way interactions, wireless by state and wireless by log(population). This resulted in a total of 80 submodels for consideration.

We use the data before 2018 as our training data, with sample size $m =$ 3,174 for shore fishing and $m =$ 3,164 for private boat fishing. The best five models and additional reference models are given in Table 4.1 for shore fishing and Table 4.2 for private boat fishing. The tables are ordered by AIC values, with the best models at the top. The models that ignore some (largest minus all mail, largest minus all wireless) or all (smallest) of the survey mode differences are not competitive with the models that include these factors. The largest model considered is quite competitive, with the best models dropping a small number of interactions from that largest model.

While not the best model in terms of AIC for either shore or private boat, the largest model minus the mail by log(population) and mail by state interaction is fifth best in both cases. It is operationally convenient to use a common model for both reconciliations, and this particular model is further convenient because, when extrapolating back in time, it involves only wave level shifts once the effect of wireless has died out.

Using the fitted models, we conducted some diagnostics to assess the potential importance of temporal (wave to wave) autocorrelation in our random effects specification. We subtracted off the estimated fixed effects and computed empirical covariances at each of lags one through six within each state. These empirical covariances would include any covariance due to correlated random effects, but no covariance arising from the independent sampling errors. We also computed a version of a Ljung and Box (1978) statistic (normalized sum of squared autocorrelations) for shore fishing and for private boat fishing in each state. We compared each computed statistic to a null distribution (no autocorrelation) obtained by resampling the residuals. Among the 34 statistics, 11 were significant at the conventional 0.05 level, with first-order autocorrelation values ranging from -0.11 to 0.29. Because these estimated values were small

and inconsistent across states and fishing modes, we did not further pursue autocorrelated random effects in our modeling.

Table 4.1
Out-of-sample prediction MSE, AIC and number of fixed effect parameters for various models fitted to effort estimates for shore fishing. See text for description of largest model.

Model is largest minus terms below:	MSE	AIC	Parameters
mail:log(pop), mail:state, wireless:wave	0.0837	4,564.28	152
mail:state, wireless:wave	0.0899	4,564.69	153
mail:log(pop) and wireless:wave	0.1350	4,564.86	168
wireless:wave	0.1354	4,566.85	169
mail:log(pop) and mail:state	0.0840	4,570.45	157
nothing (largest)	0.1343	4,573.28	174
mail interactions	0.2104	4,580.51	152
wireless interactions	0.3694	4,719.05	136
all interactions	0.3341	4,742.84	124
all wireless	0.4745	4,758.73	145
all mail	1.9466	4,838.73	151
all mail and all wireless (smallest)	2.7443	5,106.70	122

Table 4.2
Out-of-sample prediction MSE, AIC and number of fixed effect parameters for various models fitted to effort estimates for private boat fishing. See text for description of largest model.

Model is largest minus terms below:	MSE	AIC	Parameters
nothing (largest)	0.2068	3,314.55	174
mail:log(pop)	0.2124	3,314.56	173
mail:log(pop) and wireless:wave	0.2163	3,316.42	168
wireless:wave	0.2241	3,316.47	169
mail:log(pop) and mail:state	0.2050	3,322.73	157
mail:state	0.1910	3,323.00	158
mail interactions	0.2272	3,362.27	152
all mail	0.7046	3,501.23	151
wireless interactions	0.4004	3,520.33	136
all interactions	0.4615	3,646.78	114
all wireless	0.5421	3,750.03	135
all mail and all wireless (smallest)	1.2677	3,901.82	112

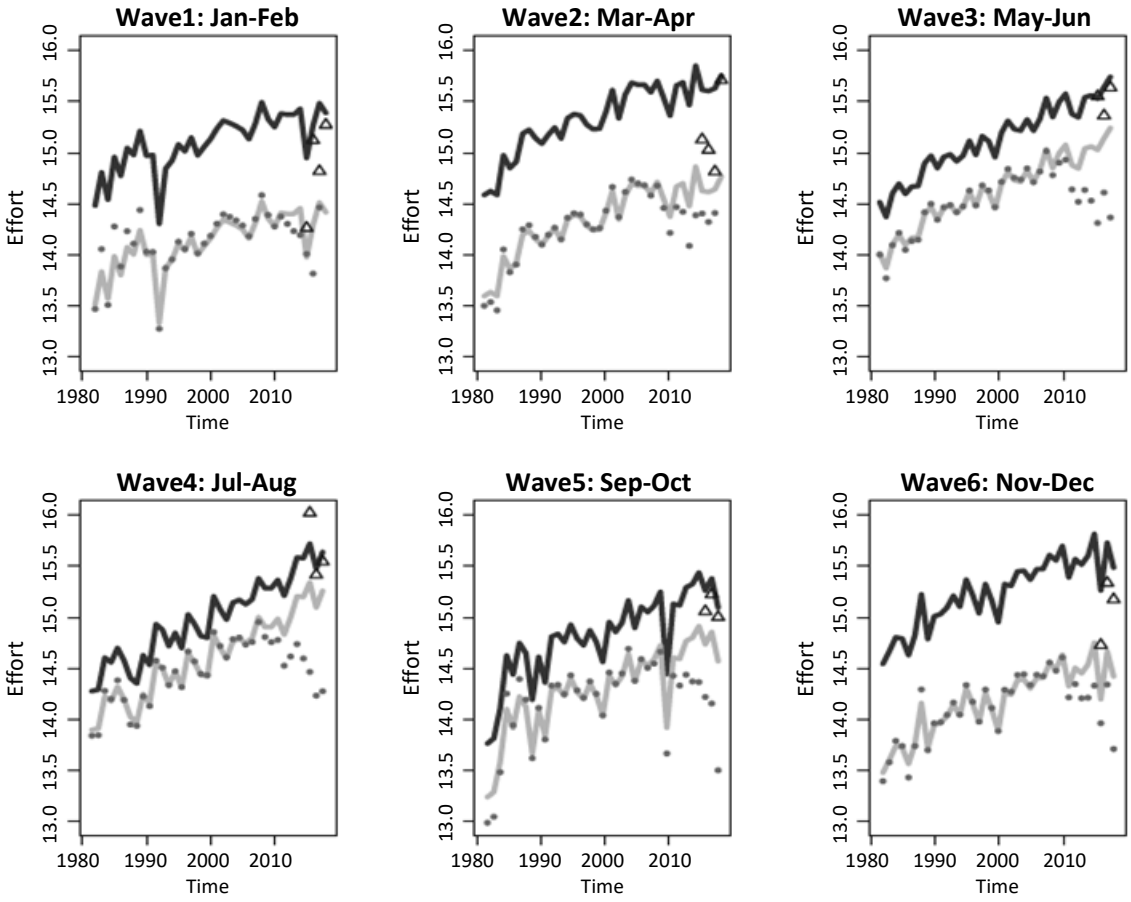
We use the first two waves of 2018 as our out-of-sample data for prediction; not every state has data in these waves, resulting in 18 out-of-sample observations each for shore fishing and private boat fishing. (The 2018 data are mail-only, and were selected for out-of-sample prediction because predicting mail is the most interesting use case.) The selected model has the lowest out-of-sample MSE for private boats and is tied (to three decimal places) with the lowest for shore fishing. Finally, the selected model is one of the most parsimonious among the top models. We therefore choose this model as the final model for both modes of

fishing, and refit it using REML to estimate the unknown variance ψ. We then compute EBLUPs of the mail target $\{M_{st}\}$ for all states and waves.

The model effectively borrows strength and reduces the variance of the direct estimates. For shore fishing, the averaged design variance is 0.0792 and the averaged in-sample MSE is 0.0445; for private boat fishing, they are 0.0789 and 0.0317, respectively.

An example of effort estimation by wave for Florida private boat fishing is shown in Figure 4.2. In each subfigure, we show the original point estimates from the telephone survey ($\hat{T}_{st}$: gray dots) and from the mail survey ($\hat{M}_{st}$: open triangles). The black curve shows the EBLUPs of the mail target, M_{st}. The gray curve shows the EBLUPs of the telephone target corrected for the wireless effect, $T_{st} - w_{st}\mathbf{c}_{st}^{\mathrm{T}}\boldsymbol{\gamma}$. The curves within each panel show the trend in effort over time according to each survey mode. The curves across panels show the seasonal pattern from wave to wave, peaking in the summer months for the telephone survey (though the seasonal pattern is not strong in Florida and is hard to see for the mail survey).

Figure 4.2 Effort estimates by wave for Florida private boat fishing. Gray dots are telephone effort estimates $\{\hat{T}_{st}\}$ and open triangles are mail effort estimates $\{\hat{M}_{st}\}$. Black curve shows the EBLUPs of the mail effort, M_{st}. Gray curve shows the EBLUPs of the telephone effort corrected for the wireless effect, $T_{st} - w_{st}\mathbf{c}_{st}^{\mathrm{T}}\boldsymbol{\gamma}$.



The EBLUPs can be seen as smoothed versions of the point estimates. The gray curve either passes through the gray dots or is a shrunken version of the gray dots prior to 2010. The gray curve diverges from the point estimates after 2010, which reflects the wireless effect on the coverage of the telephone survey. In every wave, there is a positive shift from the gray curve to the black curve, which shows the underlying difference between telephone and mail survey modes.

5. Discussion

The proposed methodology accounts for various sources of variation in the effort series from each survey, including trend, seasonality and irregular terms in the true effort series, together with survey mode effects in the two series. The model assumes that differences in measurement and nonresponse errors between the two surveys would be stable over time, while the changes in coverage error over time due to growth in wireless-only households is explicitly modeled. Further, the methodology accounts for uncertainty due to sampling error, using a novel approach to ensure analytical consistency in mapping design variances estimated on the original scale to design variances estimated on the log scale.

As formulated in this paper, the reconciliation methodology turns out to follow a standard, well-established procedure: Fay-Herriot small area estimation. This means that the calibrated values turn out to be empirical best linear unbiased predictors under a linear mixed model fitted using likelihood-based techniques. The method is flexible enough to provide optimal calibrated values for different problems: predicting mail targets for time points with telephone-only data, or predicting telephone targets for time points with mail-only data, for example.

Uncertainty is quantified via a mean square error approximation of EBLUPs at new sets of covariates that adapts existing methods from the literature. Simulation results show that the mean square error approximation and its estimator are highly accurate for the kinds of sample sizes and sampling errors present in the calibration data. The methodology is readily implemented with standard software.

As the data collection continues with the new mail methodology, there will be more data to explore other possible specifications of the calibration model. Of particular interest would be other random effects specifications, such as model (d) as described in the introduction, temporal autocorrelation, hetero-skedasticity across states, or some combination of such features. The basic calibration approach would not be fundamentally altered with such alternative specifications, however.

Acknowledgements

We thank NOAA Fisheries scientists Rob Andrews and John Foster for assistance with the problem formulation and all aspects of the data compilation. NOAA affiliate Ryan Kitts-Jensen and Westat statistician Mike Brick provided useful discussion.

Appendix

A. Proportions of wireless-only households

The changing proportion of wireless-only households is a potential covariate for explaining changes in coverage error over time for the landline-only telephone survey. Here, we describe how we constructed a set of predicted proportions of wireless-only households, $\{w_{st}\}$, for every state and wave in our data.

These proportions were approximately zero in every state prior to the year 2000, but have been steadily increasing over time. While these proportions are not available in every wave, the best available data are June and/or December wireless-only proportion estimates for each state from 2007-2014 from the National Health Interview Survey, conducted by the National Center for Health Statistics (Blumberg and Luke, 2013). We transformed these proportions via empirical logits and fitted the transformed values as state-specific, continuous, piecewise linear functions with a slope change in 2010. While the uncertainty in covariates that are themselves survey estimates could be formally addressed (e.g., Ybarra and Lohr (2008); Bell, Chung, Datta and Franco (2019)), the wireless estimates are precise and smoothly varying in this application, as partly reflected by the adjusted R^2 value of 0.9948 for the fitted model. We therefore ignore sampling uncertainty in the wireless estimates in what follows. Transforming back to proportions and extrapolating backward in time yields a set of predicted proportions of wireless-only households, $\{w_{st}\}$, for every state and wave in our data.

B. Design variance model

We need to estimate σ^2_{Tst} and σ^2_{Mst} in the sampling error variance models (2.3) and (2.4), incorporating the approximately design-unbiased estimates $\hat{V}_{Tst}$ and $\hat{V}_{Mst}$ of V_{Tst} and V_{Mst}, respectively. Modeling or smoothing design variances prior to incorporation in the Fay-Herriot methodology is standard practice; see, for example (You and Chapman (2006); You (2021); You and Hidiroglou (2023)). We follow an approach related closely to generalized variance function estimation (e.g., Chapter 7 of Wolter (2007)).

Let n_{Tst} denote the telephone sample size in state s and wave t, if non-zero, and let n_{Mst} denote the mail sample size, if non-zero. Assume that given T_{st} and M_{st}, the empirical squared coefficients of variation are log-normally distributed, independent of the effort estimates $\hat{T}_{st}$ and $\hat{M}_{st}$:

$$\ln\left(\frac{\hat{V}_{Tst}}{\exp(2\hat{T}_{st})}\right) = \mathbf{d}_{Tst}^{\mathrm{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \xi_{Tst}, \quad \xi_{Tst} \sim \mathcal{N}(0, \tau_T^2) \tag{B.1}$$

where $\mathbf{d}_{Tst}$ is a vector of known covariates (state, wave, and state by wave interaction) and $\boldsymbol{\delta}_{T0}$, δ_{T1} are unknown regression coefficients, and assume that

$$\ln\left(\frac{\hat{V}_{Mst}}{\exp(2\hat{M}_{st})}\right) = \mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0} + \delta_{M1}\ln(n_{Mst}) + \xi_{Mst}, \quad \xi_{Mst} \sim \mathcal{N}(0, \tau_M^2), \tag{B.2}$$

where $\mathbf{d}_{Mst}$ is a vector of known covariates (state, wave, and state by wave interaction) and $\boldsymbol{\delta}_{M0}$, δ_{M1} are unknown regression coefficients. These models can be rewritten as regression models for the design variance estimates, with known offsets:

$$\ln\left(\hat{V}_{Tst}\right) = 2\hat{T}_{st} + \mathbf{d}_{Tst}^{\mathsf{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \xi_{Tst}, \quad \xi_{Tst} \sim \mathcal{N}(0, \tau_T^2) \tag{B.3}$$

and

$$\ln\left(\hat{V}_{Mst}\right) = 2\hat{M}_{st} + \mathbf{d}_{Mst}^{\mathsf{T}}\boldsymbol{\delta}_{M0} + \delta_{M1}\ln(n_{Mst}) + \xi_{Mst}, \quad \xi_{Mst} \sim \mathcal{N}(0, \tau_M^2). \tag{B.4}$$

Empirically, each of these models fits very well: 94.54% adjusted R^2 value for telephone, and 98.01% adjusted R^2 value for mail.

These empirical models may be of independent interest as generalized variance functions for variance estimation on the original scale: by plugging the point estimate, state, wave, and sample size into the fitted versions of (B.3) or (B.4), one obtains excellent point estimates of the log design variance.

Assuming that $\hat{V}_{Tst}$ is exactly unbiased for V_{Tst}, we then have from the log-normal CV model (B.1) and the assumed conditional independence of $\hat{V}_{Tst}$ and $\hat{T}_{st}$ given T_{st} that

$$\begin{aligned}
\exp\left\{\mathbf{d}_{Tst}^{\mathsf{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \frac{\tau_T^2}{2}\right\} &= \mathrm{E}\left[\frac{\hat{V}_{Tst}}{\exp\left(2\hat{T}_{st}\right)} \,\middle|\, T_{st}\right] \\
&= \mathrm{E}\left[\hat{V}_{Tst} \,\middle|\, T_{st}\right]\mathrm{E}\left[\exp\left(-2\hat{T}_{st}\right)\middle|\, T_{st}\right] \\
&= V_{Tst}\exp\left(-2T_{st} + 2\sigma_{Tst}^2\right),
\end{aligned} \tag{B.5}$$

and similarly

$$\begin{aligned}
\exp\left\{\mathbf{d}_{Mst}^{\mathsf{T}}\boldsymbol{\delta}_{M0} + \delta_{M1}\ln(n_{Mst}) + \frac{\tau_M^2}{2}\right\} &= \mathrm{E}\left[\frac{\hat{V}_{Mst}}{\exp\left(2\hat{M}_{st}\right)} \,\middle|\, M_{st}\right] \\
&= \mathrm{E}\left[\hat{V}_{Mst}\middle|M_{st}\right]\mathrm{E}\left[\exp\left(-2\hat{M}_{st}\right)\middle|\, M_{st}\right] \\
&= V_{Mst}\exp\left(-2M_{st} + 2\sigma_{Mst}^2\right).
\end{aligned} \tag{B.6}$$

Thus, we have from (2.3) and (B.5) that

$$\begin{aligned}
\exp\left\{\mathbf{d}_{Tst}^{\mathsf{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \frac{\tau_T^2}{2}\right\} &= \left\{\exp(\sigma_{Tst}^2) - 1\right\}\exp\left\{2T_{st} + \sigma_{Tst}^2\right\}\exp\left(-2T_{st} + 2\sigma_{Tst}^2\right) \\
&= \exp(4\sigma_{Tst}^2) - \exp\left(3\sigma_{Tst}^2\right)
\end{aligned} \tag{B.7}$$

and from (2.4) and (B.6) that

$$\exp\left\{\mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0}+\delta_{M1}\ln(n_{Mst})+\frac{\tau_M^2}{2}\right\} = \left\{\exp(\sigma_{Mst}^2)-1\right\}\exp\left\{2M_{st}+\sigma_{Mst}^2\right\}\exp\left(-2M_{st}+2\sigma_{Mst}^2\right) \quad \text{(B.8)}$$

$$= \exp(4\sigma_{Mst}^2)-\exp\left(3\sigma_{Mst}^2\right).$$

The left-hand-side parameters of (B.7) can be estimated from (B.1) and the left-hand-side parameters of (B.8) can be estimated from (B.2). The resulting estimates of σ_{Tst}^2 and σ_{Mst}^2 can then be obtained by solving the equations (B.7) and (B.8), which are quartic polynomials in $\exp(\sigma_{Tst}^2)$ and $\exp(\sigma_{Mst}^2)$. Using Descartes' rule of signs, it can be shown that each of these quartic equations has one negative real root, two complex conjugate roots, and one positive real root. The solutions for σ_{Tst}^2 and σ_{Mst}^2 are then the logarithms of the unique, positive real roots, which can be obtained via standard numerical procedures. While these solutions are in fact estimates, we will treat them as fixed and known in what follows, as is standard in the small area estimation techniques which we will apply in subsequent sections.

The resulting design variances on the log scale, σ_{Tst}^2 and σ_{Mst}^2, have strong correlations (0.798 and 0.803, respectively) with the variance approximations from Taylor linearization, $\hat{V}_{Tst}\exp\left(-2\hat{T}_{st}\right)$ and $\hat{V}_{Mst}\exp\left(-2\hat{M}_{st}\right)$. But they are not identical (see Figure B.1), and the method described forces analytical consistency between the mean model and the variance model and does some smoothing of the design variances. Further, the method produces sensible estimates for some cases in which the design variances have been artificially thresholded at a maximum value, as shown in the left panel of Figure B.1.

Figure B.1 Estimated design variances for effort (log trips) via Taylor linearization versus solution of the quartic polynomial equations (B.7) for telephone (top panel) and (B.8) for mail (bottom panel).

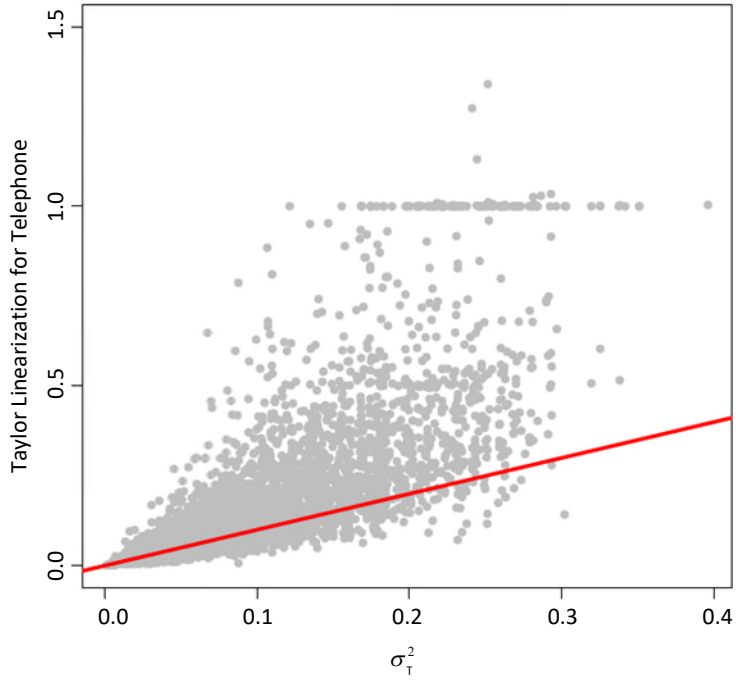
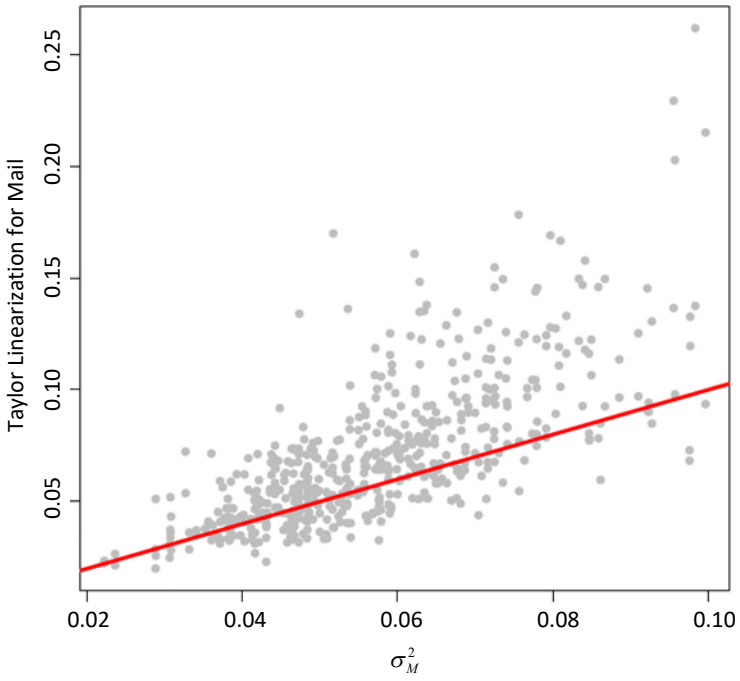


Figure B.1(continued) Estimated design variances for effort (log trips) via Taylor linearization versus solution of the quartic polynomial equations (B.7) for telephone (top panel) and (B.8) for mail (bottom panel).



References

Andrews, R., Brick, J.M. and Mathiowetz, N.A. (2014). Development and testing of recreational fishing effort surveys: Testing a mail survey design. Technical report, National Marine Fisheries Service. https://www.st.nmfs.noaa.gov/pims/main/public?method=DOWNLOAD_FR_PDF&record_id=1179.

Bell, W.R., Chung, H.C., Datta, G.S. and Franco, C. (2019). Measurement error in small area estimation: Functional versus structural versus naïve models. *Survey Methodology*, 45, 1, 61-80. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00005-eng.pdf.

Blumberg, S., and Luke, J. (2013). Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2012. Technical report, National Center for Health Statistics. http://www.cdc.gov/nchs/nhis.htm.

Boonstra, H.J. and van den Brakel, J. (2022). Multilevel time-series models for small area estimation at different frequencies and domain levels. *The Annals of Applied Statistics*, 16(4), 2314-2338.

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.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69(1), 87.

Datta, G., Lahiri, P., Maiti, T. and Lu, K. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, 94, 1074-1082.

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(1), 183-196.

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

Erciulescu, A.L., Opsomer, J.D. and Breidt, J.F. (2021). A bridging model to reconcile statistics based on data from multiple surveys. *The Annals of Applied Statistics*, 15(2), 1068-1079.

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

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55(2), 182-199.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, 15(1), 1.

Ljung, G.M., and Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.

Lohr, S.L., and Brick, J.M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canadian Journal of Statistics*, 40(4), 679-696.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. and Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1), 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 27-48.

Molina, I., and Marhuenda, Y. (2015). `sae`: An `R` package for small area estimation. *The R Journal*, 7/1, 81-98.

National Research Council (2006). *Review of Recreational Fisheries Survey Methods*. Washington, DC: The National Academies Press.

Olson, K., Smyth, J.D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N.A., McCarthy, J.S., O'Brien, E., Opsomer, J.D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z.T., Turakhia, C. and Wagner, J. (2020). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *International Statistical Review*, 70(1), 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Pfeffermann, D., and Tiller, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101(476), 1387-1397.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. and Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102(478), 474-486.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology. Wiley.

Rao, J.N.K., and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4), 511-528.

Slud, E.V., and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2), 239-257.

US Census Bureau (2016). *State Population Totals Datasets: 2010-2016*, 2016. https://www.census.gov/data/datasets/2016/demo/popest/state-total.html.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3), 581-613.

van den Brakel, J.A. (2010). Sampling and estimation techniques for the implementation of new classification systems: The change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys. *Survey Research Methods*, 4, 103-119.

van den Brakel, J.A. (2013). Design-based analysis of factorial designs embedded in probability samples. *Survey Methodology*, 39, 2, 323-349. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11870-eng.pdf.

van den Brakel, J., Zhang, X. and Tam, S.-M. (2020). Measuring discontinuities in time series obtained with repeated sample surveys. *International Statistical Review*, 88(1), 155-175.

Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C. and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(1), 84-106.

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

Ybarra, L.M., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

You, Y. (2021). Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling. *Survey Methodology*, 47, 2, 361-370. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00007-eng.pdf.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.

You, Y., and Hidiroglou, M.A. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, 39(4), 571-590.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 347-373
Statistics Canada, Catalogue No. 12-001-X

Fully synthetic data for complex surveys

Shirley Mathur, Yajuan Si and Jerome P. Reiter[1]

Abstract

When seeking to release public use files for confidential data, statistical agencies can generate fully synthetic data. We propose an approach for making fully synthetic data from surveys collected with complex sampling designs. Our approach adheres to the general strategy proposed by Rubin (1993). Specifically, we generate pseudo-populations by applying the weighted finite population Bayesian bootstrap to account for survey weights, take simple random samples from those pseudo-populations, estimate synthesis models using these simple random samples, and release simulated data drawn from the models as public use files. To facilitate variance estimation, we use the framework of multiple imputation with two data generation strategies. In the first, we generate multiple data sets from each simple random sample. In the second, we generate a single synthetic data set from each simple random sample. We present multiple imputation combining rules for each setting. We illustrate the repeated sampling properties of the combining rules via simulation studies, including comparisons with synthetic data generation based on pseudo-likelihood methods. We apply the proposed methods to a subset of data from the American Community Survey.

Key Words: Bootstrap; Confidentiality; Disclosure; Privacy; Weights.

1. Introduction

Many national statistics agencies, survey organizations, and researchers – henceforth all called agencies – disseminate microdata, i.e., data on individual units, to the public. Wide dissemination of microdata greatly benefits society, enabling broad subsets of the research community to access and analyze the collected data (Reiter, 2009). Often, however, agencies cannot release microdata as collected, because doing so could reveal survey respondents' identities or values of sensitive attributes, thereby failing to satisfy ethical or legal requirements to protect data subjects' confidentiality (Reiter and Raghunathan, 2007).

To manage these risks, several agencies have implemented or are considering synthetic data approaches, as first proposed by Rubin (1993). In this approach, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. These are called fully synthetic data sets (Drechsler, 2011; Raghunathan, 2021). Releasing fully synthetic data can preserve confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values (Reiter and Drechsler, 2010). Methods for inferences from these multiply-imputed data files have been developed for a variety of statistical inference tasks (Raghunathan, Reiter and Rubin, 2003; Reiter, 2002, 2005a,b; Drechsler and Reiter, 2010; Si and Reiter, 2011).

While prominent applications of fully synthetic data exist for censuses or administrative data (e.g., Kinney, Reiter, Reznek, Miranda, Jarmin and Abowd, 2011), many research data sets are based on surveys

1. Shirley Mathur, Department of Statistics, B-313 Padelford Hall, University of Washington, Seattle, WA 98195-4322; Yajuan Si, Survey Research Center, Institute for Social Research, University of Michigan, Rm 4014, 426 Thompson St., Ann Arbor, MI 48104. E-mail: yajuan@umich.edu; Jerome P. Reiter, Department of Statistical Science, 214a Old Chemistry Building, Duke University, Durham, NC 27708-0251.

collected with sampling designs that use unequal probabilities of selection. Previous research on multiple imputation for missing data suggests that imputation models should account for the survey design features, such as stratification, clustering, and survey weights (Reiter, Raghunathan and Kinney, 2006). Similarly, when using multiple imputation for synthetic data, the models also should account for the survey design (Mitra and Reiter, 2006; Fienberg, 2010; Kim, Drechsler and Thompson, 2021). The key challenge is properly incorporating weights in the synthesis models, which relates to the long-standing debate about the role of survey weights in model-based inferences (Pfeffermann, 1993, 2011; Little, 2004).

Researchers have proposed a variety of approaches for generating fully synthetic data in complex surveys. The suggestion in early work (Rubin, 1993; Raghunathan et al., 2003; Reiter, 2002) was to take a Bayesian finite population inference approach, in which the agency (i) builds predictive models for the survey variables conditional on design features like stratum/cluster indicators or size measures, which are assumed known by the agency for every unit in the population, (ii) imputes the missing survey variables for the nonsampled units in the population, and (iii) takes a simple random sample from the completed population to release as one synthetic data set. A related approach uses the weighted finite population Bayesian bootstrap (WFPBB) (Dong, Elliott and Raghunathan, 2014), in which the agency generates completed populations by replicating individuals from the confidential data in proportion to their survey weights and then releases the completed populations, forgoing the step of simple random sampling. More recently, it has been suggested to build synthetic data models that account for the sampling design directly, so that they estimate the joint distribution of the population data. For example, the agency can use a pseudo-likelihood approach (Pfeffermann, 1993; Savitsky and Toth, 2016), in which each individual's contribution to the likelihood function of a synthesis model is raised to a power that is a function of the survey weights (Kim et al., 2021). Departing from the proposal of Rubin (1993), a completely different approach is to create and attach new weights to synthetic data records simulated from models that are agnostic to the survey weights (United Nations Economic Commission for Europe, 2022). Here, the goal is to allow users to use weighted estimates that scale up to the finite population. The new weights can be created by treating the survey weights as a variable in the synthesis, so that the agency specifies a predictive model for the weights. The simulated weights may be adjusted by raking or calibration before inclusion in the released file.

Each of these methods has its potential drawbacks. The Bayesian finite population inference approach, while theoretically principled, requires completing full populations, which can be cumbersome, and the availability of design variables for all records in the population, which may not be the case in some surveys. The WFPBB releases (multiple copies of) individuals' genuine data records, which creates obvious disclosure risks. Pseudo-likelihood approaches may not estimate sampling variability correctly (Williams and Savitsky, 2021), and it is not clear how easily they can be implemented with machine learning synthesizers like classification and regression trees (Reiter, 2005c), which are commonly used in practical synthetic data projects (Raab, Nowok and Dibben, 2018). With synthesized weights, secondary analysts are expected to use the simulated weights to approximate design-based inference. This approximation does not have a theoretical basis; as such, it is unclear whether the synthetic weights approach facilitates accurate inferences in general.

In this article, we propose an approach to generate fully synthetic data from complex samples in the spirit of the original proposal of Rubin (1993), i.e., the agency releases simple random samples that do not require users to perform survey-weighted analyses with the synthetic data. To do so, we build on the WFPBB approach of Dong et al. (2014) by first creating pseudo-populations that account for the survey weights. We then take simple random samples (SRSs) from each pseudo-population, estimate synthesis models from each SRS, and generate draws from these models to create multiply-imputed, fully synthetic public use files. The latter step provides confidentiality protection, as the agency is not releasing genuine records. We consider two processes for the last step of generating the synthetic data. In *Synrep-R*, we generate multiple synthetic data sets from each SRS. In *SynRep-1*, we generate one synthetic data set from each SRS. *SynRep-R* releases more data sets than *SynRep-1*, which can result in reduced variances. However, the additional data sets can increase the overhead for the agency and secondary analysts, and they provide additional information for adversaries seeking disclosures. For both approaches, we derive multiple imputation combining rules that enable the estimation of variances. Using simulation studies, we illustrate the repeated sampling performances of the combining rules and compare them to fully synthetic data generated while disregarding the sampling design entirely. We also compare them against approaches that use synthesis models estimated with weighted pseudo-likelihoods (Kim et al., 2021). Finally, we illustrate the proposed methods using a subset of the American Community Survey (ACS) data. Code for the simulation studies and the ACS illustration is available at https://github.com/yajuansi-sophie/SynRep.

The remainder of the article is organized as follows. Section 2 describes the two synthetic data generation processes in detail and presents the new combining rules. Section 3 presents the simulation studies. Section 4 presents the illustration with the ACS data. Section 5 suggests topics for future research.

2. Proposed methods for generating fully synthetic survey data

Let $\mathcal{D}$ be a probability sample of size n randomly drawn from a finite population comprising N units. For $i = 1, \ldots, N,$ let π_i be the selection probability for unit $i,$ and let $w_i = 1/\pi_i$ be the unit's survey weight. Here, we are agnostic as to whether w_i is potentially adjusted, e.g., for normalization, calibration or nonresponse, although in our simulation studies we use pure design weights. For $i = 1, \ldots, N,$ let Y_i be the $p \times 1$ vector of survey variables. Hence, $\mathcal{D} = \{(w_i, Y_i) : i = 1, \ldots, n\}.$ For simplicity of exposition, we suppose that $p = 1,$ so that Y_i is a scalar. *SynRep-R* and *SynRep-1*, and their corresponding inferential methods, can be used with multivariate survey data as well.

In Section 2.1, we describe the processes of generating synthetic data. In Section 2.2, we describe the inferential methods. As mentioned in Section 1 and following the proposal in Rubin (1993), we take as a goal allowing secondary users to analyze the released data sets as if they were simple random samples from the population.

2.1 Data generation process

Figure 2.1 and Figure 2.2 display the processes of generating synthetic data for *SynRep-R* and *SynRep-1*, respectively. We now describe these steps in detail.

Figure 2.1 Process for generating synthetic data with multiple data sets per simple random sample (SRS), which we call *SynRep-R*.

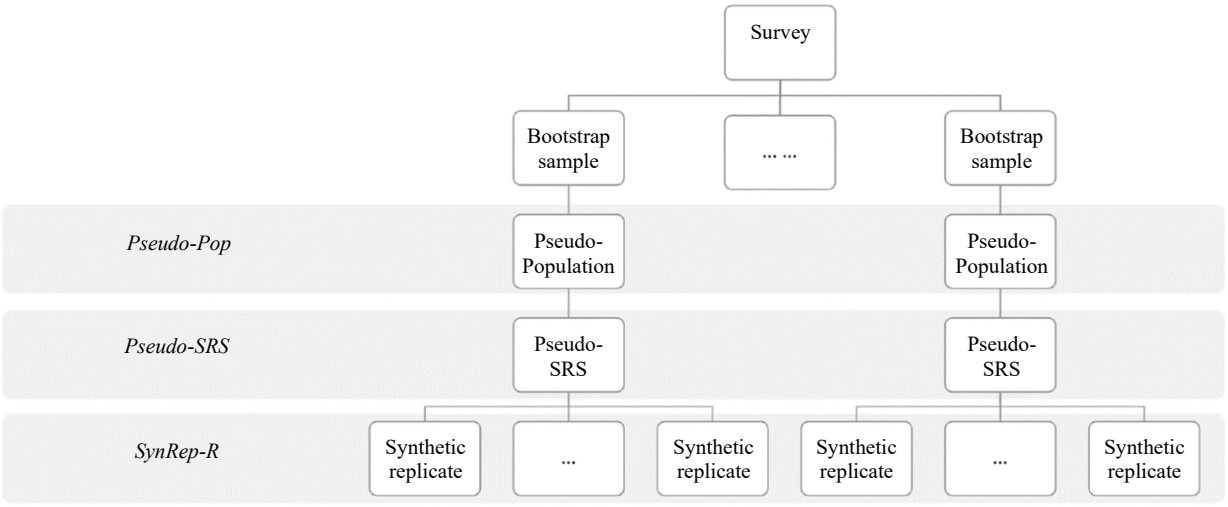
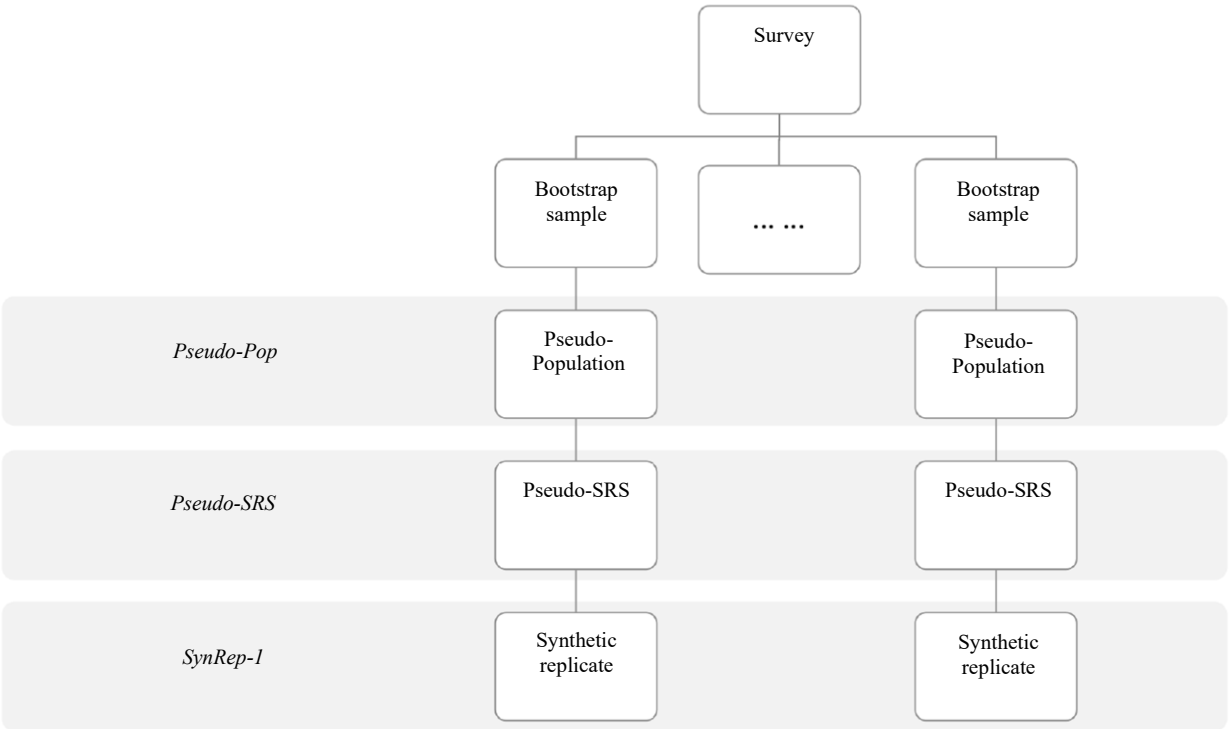


Figure 2.2 Process for generating synthetic data with one data set per simple random sample (SRS), which we call *SynRep-1*.



In either process, the first step is to generate pseudo-populations using the WFPBB (Dong et al., 2014). The WFPBB generates pseudo-populations by "undoing" the complex sampling design and accounting for the sampling weights. The idea is to draw from the posterior predictive distribution of non-observed data (Y_{nob}) given the observed data (Y_{obs}) and the survey weights, i.e., drawing from $P(Y_{\text{nob}} \mid Y_{\text{obs}}, w_1, \ldots, w_n)$. This distribution supposes that the population is comprised of the unique values of $Y_i \in \mathcal{D}$, and that the corresponding counts for each value in the population follow a multinomial distribution. With a non-informative Dirichlet prior distribution on the multinomial probabilities, the Pólya distribution can be used to draw the predictive samples in place of the Dirichlet-multinomial distribution.

With this in mind, the process of generating the synthetic data is described below.

1. **Resample via Bayesian bootstrap**: To inject sufficient sampling variability, using the data from the "parent" sample $\mathcal{D}$, we generate M samples, $(\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(M)})$, each of size n using independent Bayesian bootstraps (Rubin, 1981). For each $\mathcal{S}^{(m)}$ and for $i = 1, \ldots, n$, let $w_i^{(m)} = c w_i r_i^{(m)}$, where $r_i^{(m)}$ is the number of times that element i from $\mathcal{D}$ appears in $\mathcal{S}^{(m)}$. The c is a normalizing constant to ensure that the new weights sum to the population size N. Thus, in each $\mathcal{S}^{(m)}$, for $i = 1, \ldots, n$, we create $w_i^{(m)} = \left(N w_i r_i^{(m)}\right) \Big/ \left(\sum_k w_k r_k^{(m)}\right)$.
2. **Use the WFPBB to make pseudo-populations**: For each $\mathcal{S}^{(m)}$, we construct an initial Pólya urn using the set of $\left\{Y_i, w_i^{(m)}\right\}$. We then draw $N - n$ units using probabilities $\left(p_1^{(m)}, \ldots, p_n^{(m)}\right)$ determined from

$$p_i^{(m)} = \frac{w_i^{(m)} - 1 + l_{i,k-1}^{(m)} (N-n)/n}{N - n + (k-1)(N-n)/n}, \tag{2.1}$$

   for the k th draw, $k \in \{1, \ldots, N-n\}$, where $l_{i,k-1}^{(m)}$ is the number of bootstrap selections of Y_i among the elements present in the urn at the $k-1$ draw. The $N-n$ draws combined with the data in $\mathcal{S}^{(m)}$ comprise one pseudo-population, $\mathcal{P}^{(m)}$. We repeat this for $m = 1, \ldots, M$ to create $\mathcal{P}_{\text{pseudo}} = \{\mathcal{P}^{(m)} : m = 1, \ldots, M\}$. When N is very large, we can save memory and computational costs by creating a pseudo-population that is large enough to be practically the same for inference as a population of size N, which we operationalize by generating $50n$ rather than $N-n$ records.
3. **Draw SRS from each pseudo-population**: For $m = 1, \ldots, M$, take a simple random sample $\mathcal{D}^{(m)}$ of size n from $\mathcal{P}^{(m)}$. Let $\mathcal{D}_{\text{srs}} = \{\mathcal{D}^{(m)} : m = 1, \ldots, M\}$.
4. **Generate synthetic data replicates:** For $m = 1, \ldots, M$, estimate a synthesis model using $\mathcal{D}^{(m)}$, and draw from the predictive distributions to form synthetic data replicates using either Step 4a or Step 4b.

   4a. *SynRep-R:* For $m = 1, \ldots, M$, draw $R > 1$ synthetic replicates $\mathcal{D}_{\text{syn}}^{(m,r)}$ of size n, where $r = 1, \ldots, R$, using each $\mathcal{D}^{(m)}$. We release $\mathcal{D}_{\text{syn}} = \{\mathcal{D}_{\text{syn}}^{(m,r)} : m = 1, \ldots, M;\ r = 1, \ldots, R\}$ including indicators of which m each $\mathcal{D}_{\text{syn}}^{(m,r)}$ belongs to.

   4b. *SynRep-1:* For $m = 1, \ldots, M$, draw one synthetic data sample $\mathcal{D}_{\text{syn}}^{(m)}$ of size n from each $\mathcal{D}^{(m)}$. Release $\mathcal{D}_{\text{syn}} = \{\mathcal{D}_{\text{syn}}^{(m)} : m = 1, \ldots, M\}$.

The synthesis model for each $\mathcal{D}^{(m)}$ can utilize plug-in values of model parameters, e.g., their maximum likelihood estimates. It is not necessary to use posterior distributions at this stage of the process (Reiter and Kinney, 2012).

As these two processes for generating synthetic data differ from those of Raghunathan et al. (2003), as well as from other synthetic data scenarios such as those of Reiter (2003, 2004), we require new methods for inferences, to which we now turn.

2.2 Inferences for *SynRep-R* and *SynRep-1*

To derive the inferential methods, we follow the general strategy of multiple imputation (Rubin, 1987) and use a Bayesian inference approach. For any population quantity Q, such as the population mean $Q \equiv \bar{Y}$, we seek the posterior distribution $P(Q \mid \mathcal{D}_{\text{syn}})$. Following Raghunathan et al. (2003), we compute the following integral based upon each level of the data synthesis process from Figure 2.1 or Figure 2.2.

$$P(Q \mid \mathcal{D}_{\text{syn}}) = \iiint P(Q \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}, \mathcal{P}_{\text{pseudo}}, \mathcal{D})\, P(\mathcal{D} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}, \mathcal{P}_{\text{pseudo}}) \\ P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}})\, P(\mathcal{D}_{\text{srs}} \mid \mathcal{D}_{\text{syn}})\, d\mathcal{D} d\mathcal{P}_{\text{pseudo}}\, d\mathcal{D}_{\text{srs}}. \tag{2.2}$$

When we condition on $\mathcal{D}$, the values of $(\mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}, \mathcal{P}_{\text{pseudo}})$ do not provide any additional information about Q. Thus, we can simplify $P(Q \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}, \mathcal{P}_{\text{pseudo}}, \mathcal{D}) = P(Q \mid \mathcal{D})$. When we condition on $\mathcal{P}_{\text{pseudo}}$, the values of $(\mathcal{D}_{\text{rep}}, \mathcal{D}_{\text{syn}})$ provide no additional information about $\mathcal{D}$. Thus, we simplify $P(\mathcal{D} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}, \mathcal{P}_{\text{pseudo}}) = P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})$. When we condition on $\mathcal{D}_{\text{srs}}$, the value of $\mathcal{D}_{\text{syn}}$ provides no information about $\mathcal{P}_{\text{pseudo}}$. Hence, $P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{srs}}) = P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{srs}})$. With some re-arrangement to aid interpretation, we re-express (2.2) as

$$P(Q \mid \mathcal{D}_{\text{syn}}) = \int \left[\int \left[\int P(Q \mid \mathcal{D})\, P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})\, d\mathcal{D} \right] P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{srs}})\, d\mathcal{P}_{\text{pseudo}} \right] \\ P(\mathcal{D}_{\text{srs}} \mid \mathcal{D}_{\text{syn}})\, d\mathcal{D}_{\text{srs}}. \tag{2.3}$$

We begin with $P(Q \mid \mathcal{P}_{\text{pseudo}}) = \int P(Q \mid \mathcal{D})\, P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})\, d\mathcal{D}$. We assume that, for large M, this is approximately a normal distribution. This should be reasonable in large samples, which are typical in settings where agencies want to release public use data. We only require the posterior distribution of Q to be normal, not the distribution of the survey variables themselves; indeed, the underlying data can be categorical. We note that the inferential methods are not intended for quantities like medians or other quantiles; inferential methods for such quantities is a topic for additional research.

We only require means and variances to characterize normal sampling distributions. Thus, we focus on estimating the distributions of the first two moments. For $m = 1, \ldots, M$, let $Q^{(m)}$ be the computed value of Q if we had access to $\mathcal{P}^{(m)}$. Rubin (1987) shows that

$$(Q \mid \mathcal{P}_{\text{pseudo}}) \sim t_{M-1}\left(\bar{Q}, (1 + M^{-1})\, B\right), \tag{2.4}$$

where $\bar{Q} = \sum_m Q^{(m)}/M$ and $B = \sum_m (Q^{(m)} - \bar{Q})^2/(M-1)$. Here $t_\nu(\mu, \sigma^2)$ denotes a t-distribution with ν degrees of freedom, location μ, and variance σ^2. In the derivations, for convenience we approximate the t-distribution in (2.4) as a normal distribution, which should be reasonable for somewhat large M.

We next turn to $P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{srs}})$. Here, we only need $P(\bar{Q}, B \mid \mathcal{D}_{\text{srs}})$. For $m = 1, \ldots, M$, let $q^{(m)}$ be the estimate of $Q^{(m)}$ and $v^{(m)}$ be the estimate of the sampling variance associated with $q^{(m)}$; we could compute these if we had access to $\mathcal{D}^{(m)}$. We assume that $\{q^{(m)}, v^{(m)} : m = 1, \ldots, M\}$ are valid in the following sense.

1) For each m, $q^{(m)}$ is approximately unbiased for $Q^{(m)}$ and asymptotically normally distributed, with respect to repeated sampling from the pseudo-population $\mathcal{P}^{(m)}$ with sampling variance $V^{(m)}$. That is, we have $(q^{(m)} \mid \mathcal{P}^{(m)}) \sim N(Q^{(m)}, V^{(m)})$.
2) The sampling variance estimate $v^{(m)}$ is approximately unbiased for $V^{(m)}$, and the sampling variability in $v^{(m)}$ is negligible. That is, $(v^{(m)} \mid \mathcal{P}^{(m)}) \approx V^{(m)}$.
3) The variation in $V^{(m)}$ across the M pseudo-populations is negligible; that is, $V^{(m)} \approx V \approx \bar{v}$, where $\bar{v} = \sum_m v^{(m)}/M$.

Using standard Bayesian arguments based on these sampling distributions, it follows that

$$(Q^{(m)} \mid q^{(m)}, \bar{v}) \sim N(q^{(m)}, \bar{v}) \tag{2.5}$$

$$(\bar{Q} \mid \bar{q}, \bar{v}) \sim N(\bar{q}, \bar{v}/M), \tag{2.6}$$

where $\bar{q} = \sum_m q^{(m)}/M$.

To obtain the distribution of $(Q \mid \mathcal{D}_{\text{srs}})$, we integrate the distribution in (2.4), which we approximate as a normal distribution, with respect to the distributions of $\bar{Q}$ and B. We only need the first two moments since the resulting distribution is a normal distribution. We have

$$E(Q \mid \mathcal{D}_{\text{srs}}) = E(E(Q \mid \bar{Q}) \mid \mathcal{D}_{\text{srs}}) = E(\bar{Q} \mid \mathcal{D}_{\text{srs}}) = \bar{q}. \tag{2.7}$$

We also have

$$\begin{aligned} \text{Var}(Q \mid \mathcal{D}_{\text{srs}}) &= E(\text{Var}(Q \mid \mathcal{P}_{\text{pseudo}}) \mid \mathcal{D}_{\text{srs}}) + \text{Var}(E(Q \mid \mathcal{P}_{\text{pseudo}}) \mid \mathcal{D}_{\text{srs}}) \\ &= (1 + M^{-1}) E(B \mid \mathcal{D}_{\text{srs}}) + \bar{v}/M. \end{aligned} \tag{2.8}$$

This is the variance estimator in Raghunathan et al. (2003), which analysts would use if the agency releases $\mathcal{D}_{\text{srs}}$ as the public use files. However, since we take an additional step of replacing each $\mathcal{D}^{(m)}$ with simulated values, we need to average over the distributions of $(\bar{q}, \bar{v}, B)$. The result depends on whether we use *SynRep-R* or *SynRep-1*, as we now describe.

2.2.1 Derivation with *SynRep-R*

For each $\mathcal{D}_{\text{syn}}^{(m,r)}$, let $q_{\text{syn}}^{(m,r)}$ be the point estimate of Q, and let $v_{\text{syn}}^{(m,r)}$ be the estimate of the variance associated with $q_{\text{syn}}^{(m,r)}$. The analyst computes $q_{\text{syn}}^{(m,r)}$ and $v_{\text{syn}}^{(m,r)}$ acting as if $\mathcal{D}_{\text{syn}}^{(m,r)}$ is the collected data obtained

via a simple random sample of size n taken from the population. The analyst needs to compute the following quantities.

$$\bar{q}_{\text{syn}}^{(m)} = \sum_{r=1}^{R} q_{\text{syn}}^{(m,r)} / R \tag{2.9}$$

$$\bar{q}_{\text{syn}} = \sum_{m=1}^{M} \bar{q}_{\text{syn}}^{(m)} / M \tag{2.10}$$

$$b_{\text{syn}} = \sum_{m=1}^{M} (\bar{q}_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2 / (M-1) \tag{2.11}$$

$$w_{\text{syn}}^{(m)} = \sum_{r=1}^{R} (q_{\text{syn}}^{(m,r)} - \bar{q}_{\text{syn}}^{(m)})^2 / (R-1) \tag{2.12}$$

$$\bar{w}_{\text{syn}} = \sum_{m=1}^{M} w_{\text{syn}}^{(m)} / M \tag{2.13}$$

$$\bar{v}_{\text{syn}} = \sum_{m=1}^{M} \sum_{r=1}^{R} v_{\text{syn}}^{(m,r)} / MR. \tag{2.14}$$

We now complete the derivation of the posterior distribution for $(Q \mid \mathcal{D}_{\text{syn}})$ in the *SynRep-R* approach. To do so, we assume large-sample normal approximations for the sampling distributions of the point estimates. Specifically, for all (m, r), we assume that

$$q_{\text{syn}}^{(m,r)} \sim N(q^{(m)}, W^{(m)}), \tag{2.15}$$

where $W^{(m)}$ is the sampling variance for $q_{\text{syn}}^{(m,r)}$ over draws of synthetic data from $\mathcal{D}^{(m)}$. The normality should be reasonable when n is large. Assuming diffuse prior distributions and conditioning on $W^{(m)}$, we have

$$(q^{(m)} \mid \mathcal{D}_{\text{syn}}^{(m,1)}, \ldots, \mathcal{D}_{\text{syn}}^{(m,R)}, W^{(m)}) \sim N(\bar{q}_{\text{syn}}^{(m)}, W^{(m)}/R) \tag{2.16}$$

$$(\bar{q} \mid \mathcal{D}_{\text{syn}}, \bar{W}) \sim N(\bar{q}_{\text{syn}}, \bar{W}/MR), \tag{2.17}$$

where $\bar{W} = \sum_m W^{(m)}/M$.

Having now determined distributions for the point estimators, we put everything together for the posterior distribution of Q. Since all the components are normal distributions, $P(Q \mid \mathcal{D}_{\text{syn}})$ is a normal distribution. Thus, for the expectation, we use (2.7) and (2.17) to obtain

$$E(Q \mid \mathcal{D}_{\text{syn}}) = (E(Q \mid \mathcal{D}_{\text{srs}}) \mid \mathcal{D}_{\text{syn}}) = E(\bar{q} \mid \mathcal{D}_{\text{syn}}) = \bar{q}_{\text{syn}}. \tag{2.18}$$

For the variance, we first write the variance in terms of $(B, \bar{v}, \bar{W})$ and then plug in point estimates of these terms. To emphasize the use of $(B, \bar{v}, \bar{W})$, we write

$$\begin{aligned} \text{Var}(Q \mid \mathcal{D}_{\text{syn}}, B, \overline{v}_M, \overline{W}) &= E(((1+M^{-1})\,B + \overline{v}/M) \mid \mathcal{D}_{\text{syn}}, B, \overline{v}, \overline{W}) \;+\; \text{Var}(\overline{q} \mid \mathcal{D}_{\text{syn}}, B, \overline{v}, \overline{W}) \\ &= (1+M^{-1})\,B + \overline{v}/M + \overline{W}/MR. \end{aligned} \tag{2.19}$$

We now define the estimates for $(B, \overline{v}, \overline{W})$, which we plug into (2.19). For $\overline{v}$, we assume that $\overline{v}_{\text{syn}} \approx \overline{v}$. This assumption follows from the rationale in Raghunathan et al. (2003), who argue this is the case when the synthetic data are generated from the same underlying distribution as the data used to fit the models.

For $\overline{W}$, we note that (2.15) implies that, for $m = 1, \ldots, M$,

$$\frac{(R-1)\,w_{\text{syn}}^{(m)}}{W^{(m)}} \;\sim\; \chi^2_{R-1}. \tag{2.20}$$

We further assume that each $W^{(m)} \approx \overline{W}$. This assumption is in line with a similar assumption provided in Reiter (2004) regarding the variability of posterior variances. Essentially, as stated in Reiter (2004), this assumption stems from the observation that variability amongst posterior variances is generally smaller in magnitude than variability in posterior expectations. With this assumption and utilizing (2.20), we have

$$\sum_{m=1}^{M} \frac{(R-1)\,w_{\text{syn}}^{(m)}}{\overline{W}} \;\sim\; \chi^2_{M(R-1)}. \tag{2.21}$$

Thus, we have

$$\text{E}\left(\sum_{m=1}^{M} \frac{(R-1)\,w_{\text{syn}}^{(m)}}{\overline{W}} \right) = M\,(R-1). \tag{2.22}$$

Utilizing a methods of moments approach to approximate $\overline{W}$, we obtain $\overline{W} \approx \overline{w}_{\text{syn}}$.

For approximating B, we note that the sampling distribution of a randomly generated $\overline{q}_{\text{syn}}^{(m)}$ over all steps in the data generation process is $N(Q, B + \overline{v} + \overline{W}/R)$. Using this fact, we have

$$\frac{\sum_{m=1}^{M} (\overline{q}_{\text{syn}}^{(m)} - \overline{q}_{\text{syn}})^2}{B + \overline{v} + \overline{W}/R} \;\sim\; \chi^2_{M-1}, \tag{2.23}$$

so that

$$\text{E}\left(\frac{\sum_{m=1}^{M} (\overline{q}_{\text{syn}}^{(m)} - \overline{q}_{\text{syn}})^2}{B + \overline{v} + \overline{W}/R} \right) = M - 1. \tag{2.24}$$

Using a method of moments approach and the definition of b_{syn} in (2.11), and the plug-in estimate $\overline{w}_{\text{syn}}$ for $\overline{W}$, we have $b_{\text{syn}} \approx B + \overline{v}_{\text{syn}} + \overline{w}_{\text{syn}}/R$, so that $B \approx b_{\text{syn}} - \overline{v}_{\text{syn}} - \overline{w}_{\text{syn}}/R$.

Putting all together, we can approximate $\text{Var}(Q \mid \mathcal{D}_{\text{syn}})$ with the estimate T_r, where

$$\begin{aligned} T_r &= (1+M^{-1})\left(b_{\text{syn}} - \overline{v}_{\text{syn}} - \overline{w}_{\text{syn}}/R\right) + \overline{v}_{\text{syn}}/M + \overline{w}_{\text{syn}}/MR \\ &= (1+M^{-1})\, b_{\text{syn}} - \overline{v}_{\text{syn}} - \overline{w}_{\text{syn}}/R. \end{aligned} \tag{2.25}$$

We compute approximate 95% intervals for Q as $\overline{q}_{\text{syn}} \pm t_{0.975,\, M-1}\sqrt{T_r}$. The t-distribution is a simple approximation based on the degrees of freedom in (2.4). As with the variance estimator in Raghunathan et al. (2003), the estimate T_r can be negative, particularly for small M. As an *ad hoc* adjustment when $T_r < 0$, we recommend replacing B with $\overline{v}$ in (2.19) and using $T_r^* = \left(1 + 2/M\right)\overline{v}_{\text{syn}} + \overline{w}_{\text{syn}}/MR$.

2.2.2 Derivation with *SynRep-1*

With large M and R, *SynRep-R* results in many synthetic data sets, which may be undesirable from the perspective of the agency and secondary data analysts. Instead, agencies may want to use *SynRep-1*. To obtain inferences for Q in this setting, we leverage the methodology of Raab et al. (2018), who observed that when the source data come from a simple random sample, as is the case for each $\mathcal{D}^{(m)}$, we can obtain valid variance estimates with single implicates with adjustments of the combining rules. We now describe this derivation.

For $m = 1, \ldots, M$, let $q_{\text{syn}}^{(m)}$ be the point estimate of Q computed using $\mathcal{D}_{\text{syn}}^{(m)}$, and let $v_{\text{syn}}^{(m)}$ be the estimated variance associated with $q_{\text{syn}}^{(m)}$. The analyst computes each $(q_{\text{syn}}^{(m)}, v_{\text{syn}}^{(m)})$ by acting is if $\mathcal{D}_{\text{syn}}^{(m)}$ is a SRS of size n from the population. We require the following quantities for inferences. To economize on notation, we re-use some of the notation introduced in Section 2.2.1.

$$\overline{q}_{\text{syn}} = \sum_{m=1}^{M} q_{\text{syn}}^{(m)}/M \tag{2.26}$$

$$b_{\text{syn}} = \sum_{m=1}^{M} (q_{\text{syn}}^{(m)} - \overline{q}_{\text{syn}})^2 / (M-1) \tag{2.27}$$

$$\overline{v}_{\text{syn}} = \sum_{m=1}^{M} v_{\text{syn}}^{(m)}/M. \tag{2.28}$$

The pairs of equations (2.26) and (2.10), (2.27) and (2.11), and (2.28) and (2.14) can be viewed as equivalent when $R = 1$.

To complete the derivation for *SynRep-1*, we follow the logic in Raab et al. (2018) and assume that $q_{\text{syn}}^{(m)} \sim N(q^{(m)}, V^{(m)})$. Assuming $V^{(m)} \approx \overline{v}$ for all m, we have

$$(q^{(m)} \mid \mathcal{D}_{\text{syn}}^{(m)}, \overline{v}) \sim N(q_{\text{syn}}^{(m)}, \overline{v}) \tag{2.29}$$

$$(\overline{q} \mid \mathcal{D}_{\text{syn}}, \overline{v}) \sim N(\overline{q}_{\text{syn}}, \overline{v}/M). \tag{2.30}$$

We note, however, that one should not assume that $B \approx \overline{v}$ as well. As $\mathcal{D}$ is a complex sample, it yields sampling variances that could differ from the simple random sampling variances associated with $\mathcal{D}_{\text{srs}}$.

Since all the components are approximately normal distributions, $P(Q|\mathcal{D}_{\text{syn}})$ also is approximately a normal distribution. For its expectation, we use (2.7) and (2.30) to obtain

$$E(Q|\mathcal{D}_{\text{syn}}) = E(E(Q|\mathcal{D}_{\text{srs}})|\mathcal{D}_{\text{syn}}) = E(\bar{q}|\mathcal{D}_{\text{syn}}) = \bar{q}_{\text{syn}}. \tag{2.31}$$

For its variance, as with *SynRep-R*, we write the variance in terms of $(B, \bar{v})$ and then plug in point estimates of these terms. We have

$$\begin{aligned}\text{Var}(Q|\mathcal{D}_{\text{syn}}, B, \bar{v}) &= E((1+M^{-1})B + \bar{v}/M)|\mathcal{D}_{\text{syn}}, B, \bar{v}) + \text{Var}(\bar{q}|\mathcal{D}_{\text{syn}}, B, \bar{v}) \\ &= (1+M^{-1})B + \bar{v}/M + \bar{v}/M = (1+M^{-1})B + 2\bar{v}/M.\end{aligned} \tag{2.32}$$

We now define the estimates for $(B, \bar{v})$ to plug into (2.32). For $\bar{v}$, we use $\bar{v}_{\text{syn}}$ defined in (2.28). This should be reasonable since we are replacing the entire set of each $\mathcal{D}^{(m)}$ with synthetic values. To approximate B, we note that the sampling distribution of a randomly generated $q_{\text{syn}}^{(m)}$ over all steps in the data generation process is $N(Q, B + 2\bar{v})$. Using this fact, we have

$$\frac{\sum_{m=1}^{M}(q_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2}{B + 2\bar{v}} \sim \chi^2_{M-1}, \tag{2.33}$$

so that

$$\text{E}\left(\frac{\sum_{m=1}^{M}(q_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2}{B + 2\bar{v}}\right) = M - 1. \tag{2.34}$$

Using a method of moments approach and the definition of b_{syn} in (2.27), we have $b_{\text{syn}} \approx B + 2\bar{v}_{\text{syn}}$, so that $B \approx b_{\text{syn}} - 2\bar{v}_{\text{syn}}$.

Thus, we can approximate $\text{Var}(Q|\mathcal{D}_{\text{syn}})$ with the estimate T_m, where

$$T_m = \left(1 + M^{-1}\right) b_{\text{syn}} - 2\bar{v}_{\text{syn}}. \tag{2.35}$$

We compute approximate 95% intervals for Q as $\bar{q}_{\text{syn}} \pm t_{0.975, M-1}\sqrt{T_m}$. When $T_m < 0$, as an *ad hoc* variance estimate we replace B by $\bar{v}$ in (2.32) and use $T_m^* = \left(1 + 3/M\right)\bar{v}_{\text{syn}}$.

3. Simulation studies

In this section, we present simulation studies to illustrate the repeated sampling properties of the inferential methods in Section 2.2 for various finite population quantities.

3.1 Simulation design

We construct a finite population based on data from the Public Use Microdata Sample of the 2021 American Community Survey (United States Bureau of the Census, 2021). This file comprises 3,252,599

individuals, which we treat as a population of size N. The file also has person-level weights (named "PWGTP" in the data file). We do not treat these as survey weights, per se; rather, we treat them as size variables x_i, where $i = 1, \ldots, N$, for use in probability proportional to size (PPS) sampling. We also use these constructed size measures to generate two survey variables, (y_{i1}, y_{i2}), where $i = 1, \ldots, N$. Specifically, we let each y_{i1} be a binary variable sampled from a Bernoulli distribution with probability $\Pr(y_{i1} = 1) = \exp(-7 + 2\log x_i) / \left(1 + \exp(-7 + 2\log x_i)\right)$. We let each y_{i2} be a continuous variable sampled from a normal distribution with mean $20 + 50 y_{i1}$ and standard deviation 50. We estimate the finite population proportion $\bar{Y}_1 = \sum_{i=1}^{N} y_{i1} / N \approx 0.765$; the finite population mean $\bar{Y}_2 = \sum_{i=1}^{N} y_{i2} / N \approx 58.2$; and, the finite population regression coefficient of Y_1 in the linear regression of Y_2 on Y_1, which is $\beta \approx 50$.

From this population, we sample $\mathcal{D}$ using a PPS sample of size $n = 500$ survey units, setting $\pi_i = n x_i \Big/ \sum_{i=1}^{N} x_i$ and using the function "ppss" in the R package "pps" (Gambino, 2021). Under this PPS sampling design, we expect that unweighted inferences using $\mathcal{D}$ should be badly biased for $(\bar{Y}_1, \bar{Y}_2)$ but perhaps not so for β. We repeat the sampling process to create 1,000 independent realizations of $\mathcal{D}$.

For each $\mathcal{D}$, we implement *SynRep-R* and *SynRep-1* with various (M, R). Specifically, we examine $(M = 4, R = 5), (M = 10, R = 5), (M = 50, R = 5), (M = 10, R = 10), (M = 10, R = 25)$, and $(M = 10, R = 50)$. The choice of R only affects *SynRep-R*. We implement the WFPBB using the "polyapost" package in R (Meeden, Lazar and Geyer, 2020), creating pseudo-populations $(\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)})$ each comprising 25,000 individuals. From each $\mathcal{P}^{(m)}$ where $m = 1, \ldots, M$, we take a simple random sample of size n to make a corresponding $\mathcal{D}^{(m)}$. To make each synthetic data replicate stemming from each $\mathcal{D}^{(m)}$, we sample n synthetic values for Y_1 using a Bernoulli distribution with probability set to the empirical proportion of Y_1 in $\mathcal{D}^{(m)}$. We sample the corresponding synthetic values of Y_2 from normal distributions with means equal to the predicted values from the regression of Y_2 on Y_1, computed using the synthetic values of Y_1 and the unbiased estimates of the coefficients computed with $\mathcal{D}^{(m)}$, and variance equal to the unbiased estimate of the regression variance computed with $\mathcal{D}^{(m)}$.

To assist in evaluating the repeated sampling performances of *SynRep-1* and *SynRep-R*, we also use results computed with $\mathcal{P}_{\text{pseudo}}$ and $\mathcal{D}_{\text{srs}}$. Specifically, in each of the 1,000 simulation runs, we define *Pseudo-Pop* as the procedure that uses a point estimator of $\bar{Q}$ and variance estimator of $(1 + 1/M)\, B$ computed with the WFPBB-generated pseudo-populations $(\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)})$. We define *Pseudo-SRS* as the procedure that uses a point estimator of $\bar{q}$ and variance estimator of Raghunathan et al. (2003) computed with $(\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(M)})$. As a comparison against what happens if we disregard the sampling design entirely, we define *SRSsyn* as the procedure that generates synthetic data by using (i) the unweighted sample proportion for Y_1 as the Bernoulli probability to generate n synthetic values of Y_1 and (ii) the unweighted estimates of parameters in the regression of Y_2 on Y_1 as the parameters of the normal distribution to generate the corresponding n synthetic values of Y_2.

We also evaluate the repeated sampling performances of pseudo-likelihood approaches to making fully synthetic data. For each synthesis model, i.e., the Bernoulli and linear regression models, we start with a likelihood function defined as the product of the contributions from each individual in $\mathcal{D}$. We create the

pseudo-likelihood by raising each individual's contribution to a power defined by the individual's survey weight. We use these weighted pseudo-likelihoods to estimate synthesis model parameters. We implement this approach using the software *Stan* (Stan Development Team, 2024), which can generate posterior samples of model parameters based on user-specified likelihood functions. We run *Stan* to create four chains of 4,000 iterations and discard the first 2,000 iterations as burn-in. We randomly sample one of the resulting draws and use its parameter values in the Bernoulli and linear regression models to generate the synthetic data. We repeat this process M times and apply the inference rules in Raghunathan et al. (2003). We call this method *Wtreg*. We note that that Kim et al. (2021) use the variance estimator in (2.8) from Raghunathan et al. (2003) with $\overline{v} = 0$. Kim et al. (2021) release synthetic populations (where $\overline{v} = 0$) rather than synthetic samples (where $\overline{v} > 0$).

We also consider a modification of *Wtreg* to address potential underestimation of variability in the parameter draws. We call this method *Wtreg-Boot*. First, we take a bootstrap sample of size n from $\mathcal{D}$. We construct the pseudo-likelihood functions using the bootstrapped data and the calibrated survey weight for each resampled individual. Using this pseudo-likelihood function, we then generate and analyze synthetic data following the steps described for *Wtreg*.

Finally, we define *Direct* as using the unweighted sample mean and standard deviation from $\mathcal{D}$, i.e., ignoring the survey weights, and *HT* as using the Horvitz and Thompson (1952) estimator and its estimated variance using $\mathcal{D}$. We use these latter two procedures to assess the importance of accounting for the sampling design in inferences with $\mathcal{D}$.

Let superscript s index the results from simulation run s, where $s = 1, \ldots, 1{,}000$. For any estimator $\hat{q}$ for any of the methods we examine, we compute the percent bias, $100\sum_{s=1}^{1{,}000}(\hat{q}^s - Q)\,/\,(1{,}000Q)$. We compute the proportion of the 1,000 95% confidence intervals based on $\hat{q}$ and its corresponding variance estimate that cover Q. We also compute the ratio of the empirical variance of the 1,000 values of $\hat{q}$ to the empirical variance of the 1,000 values of the *HT* point estimator. To investigate the accuracy of variance estimators, for each method we compute the ratio of the average of the 1,000 variance estimates over its corresponding empirical variance. Finally, to examine the stability of the variance estimator for each method, we compute the standard deviation of the 1,000 variance estimates. We present results for the first four quantities in the main text and for the last quantity in the Appendix.

3.2 Results

We first investigate the properties of *SynRep-R* and *SynRep-1* for the various settings of (M, R). Figure 3.1, Figure 3.2, and Figure 3.3 display results for $\overline{Y}_1$, $\overline{Y}_2$, and β, respectively, for these two methods as well as for *Pseudo-Pop* and *Pseudo-SRS*. All four methods offer approximately unbiased point estimates of the three finite population quantities, with simulated percent biases generally around 1% or lower. These small biases originate primarily from the step of completing populations, as the biases in *Pseudo-Pop* are close to the biases in the other three methods. As expected, compared to the variance for *HT*, the simulated variances are increasingly inflated as M decreases. Holding $M = 10$ constant, decreasing R tends to

increase the simulated variances, although the effects are less pronounced than those from decreasing M. The variability in *SynRep-1* results with fixed M reflects Monte Carlo error. Taken together, these results suggest it is preferable to increase M rather than R when keeping MR constant. For example, when we compare *SynRep-R* with $(M = 10, R = 5)$ to *SynRep-1* with $M = 50$, the latter tends to result in smaller empirical variance with closer-to-nominal coverage rates. Similar benefits appear when comparing *SynRep-R* with $(M = 10, R = 25)$ to *SynRep-R* with $(M = 50, R = 5)$. This finding accords with results from Reiter (2008), who considered a similar trade-off for nested multiple imputation for partially synthetic and missing data. We note that using larger values of M also offers smaller variability in the estimated variances, as shown in the Appendix.

Figure 3.1 Repeated sampling properties of *SynRep-1* and *SynRep-R* for $\bar{Y}_1$ under different numbers of synthetic samples (M) and replicates (R) under a probability proportional to size design.

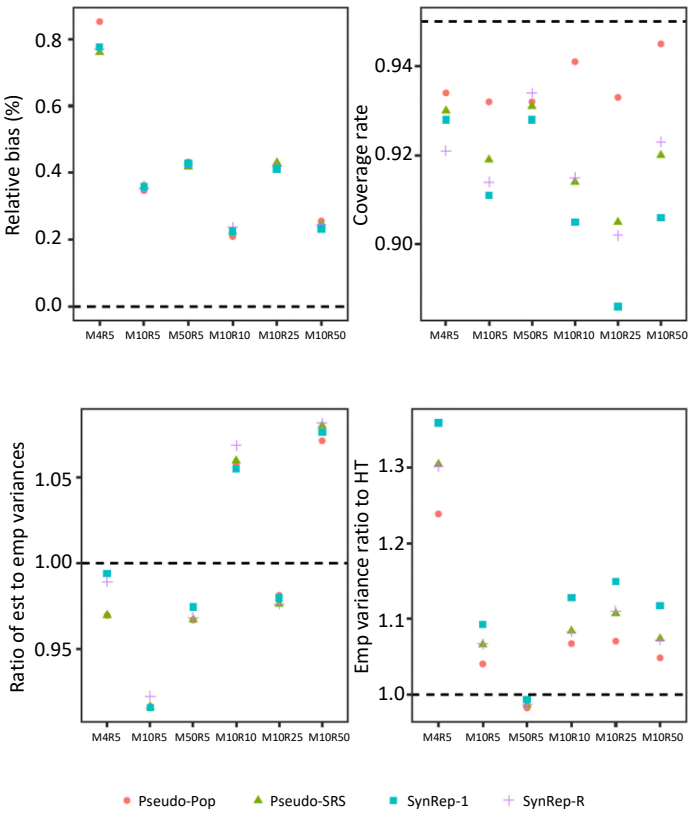
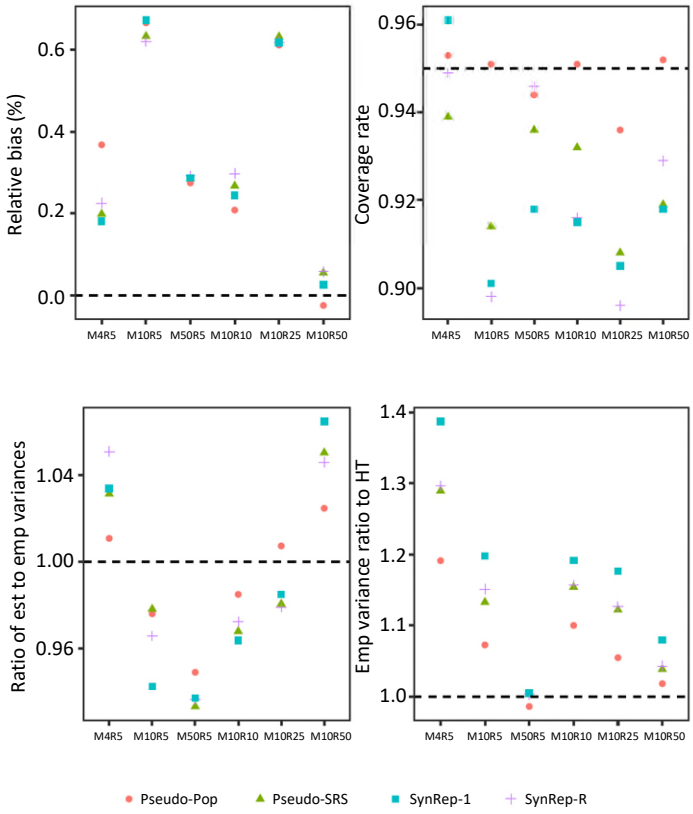


Figure 3.2 Repeated sampling properties of *SynRep-1* and *SynRep-R* for $\bar{Y}_2$ under different numbers of synthetic samples (M) and replicates (R) under a probability proportional to size design.



By comparing the ratios of the empirical variances to the variances for *HT*, we can see the effect on efficiency of the steps in the synthesis process. The variances generally increase as we go from *Pseudo-Pop* to *Pseudo-SRS* to *SynRep-R* or *SynRep-1*; that is, they increase as we add more steps that involve randomness. The variances for *SynRep-R* generally are slightly smaller than those for *SynRep-1*, reflecting the benefit for efficiency of the additional information from MR rather than M synthetic data sets. We note that the variance inflation from using synthetic data procedures versus *HT* largely disappears when $M = 50$.

Across all four synthetic data methods, the average variance estimates are reasonably similar to the empirical variances. Disparities from ratios of one apparently stem, once again, mainly from the step of completing the populations. The confidence interval coverage rates range from a low of 88% to a high of 96%, with most slightly below nominal. Coverage rates for *SynRep-R* and *SynRep-1* tend to be highest when $M = 50,$ further reflecting the benefits of using a larger M. For $M \geq 10,$ the coverage rates for *SynRep-R* tend to be higher than those for *SynRep-1*, although the difference is typically only a point or two.

The combining rules in (2.35) and (2.25) do result in negative variance estimates, as evident in Table 3.1. In the simulations, we use T_r^* and T_m^* to make confidence intervals when needed. As M increases, the number of negative variance estimates decreases. In fact, when $M = 50,$ all of the variance estimates are

positive, offering additional support for making M large. The estimates of b_{syn} become less variable as M increases, which helps avoid the negative variances. Negative variance rates tend to be lower for *SynRep-R* than for *SynRep-1*, reflecting the benefits of increased datasets to estimate variance parameters. Although not shown in Table 3.1, the negative variance rates when $M = 10$ do not change much as we increase $R \geq 5$. We note that the negative variance rates for *SynReg-R* are similar to those for *Pseudo-SRS*. Evidently, when MR is large, the information available in $\mathcal{D}_{syn}$ to estimate b_{syn} is on par with the information available in $\mathcal{D}_{srs}$.

Figure 3.3 Repeated sampling properties of *SynRep-1* and *SynRep-R* for β under different numbers of synthetic samples (M) and replicates (R) under a probability proportional to size design.

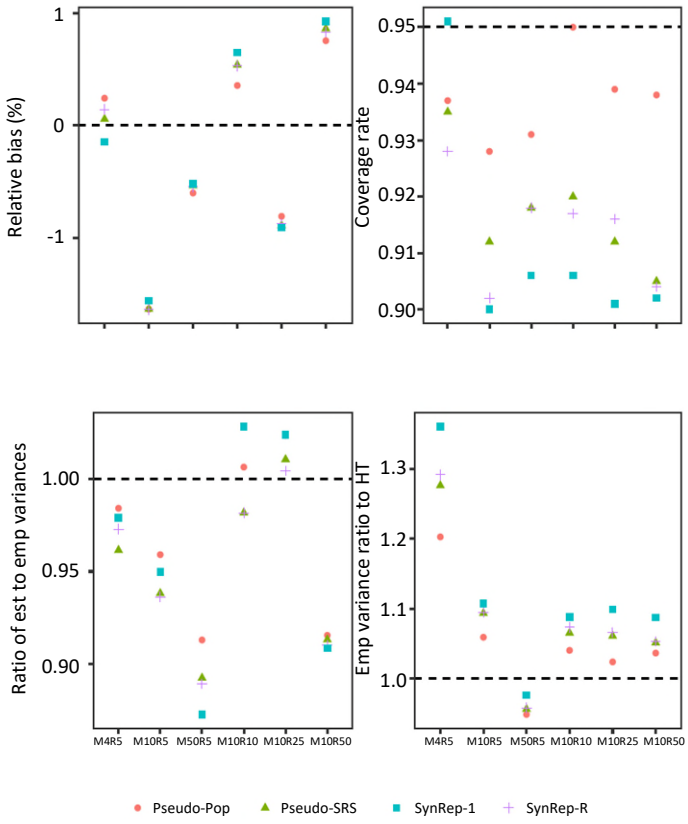
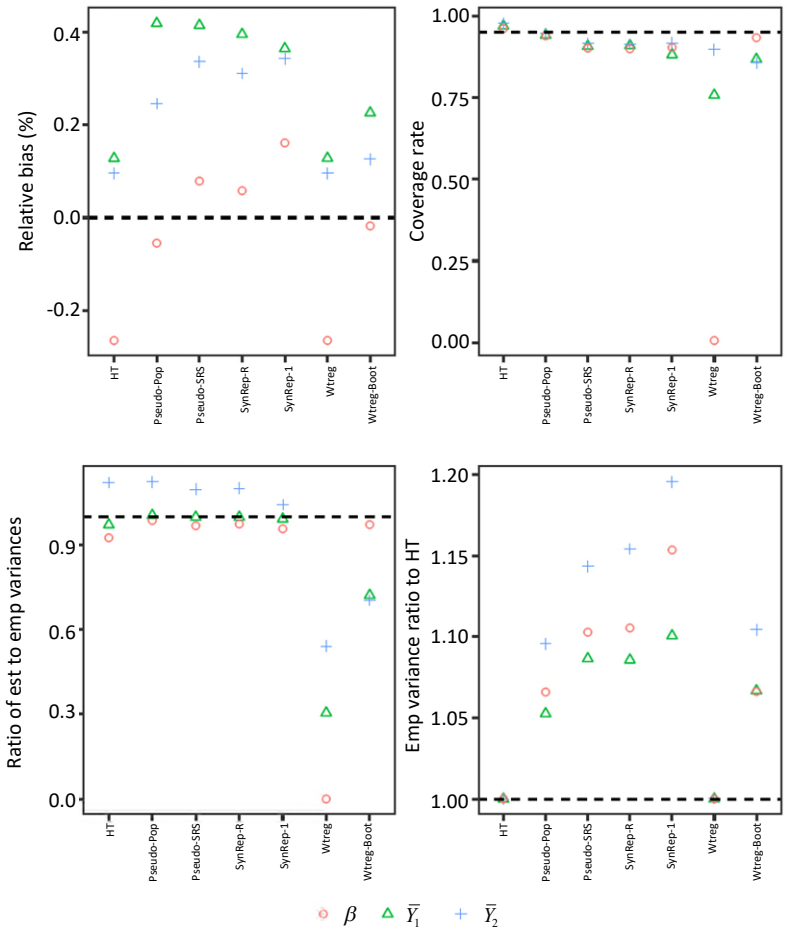


Table 3.1
Proportion of negative variance estimates in the PPS simulation studies. When $M = 50$, all variance estimates are positive.

(M, R)	Method	$\bar{Y}_1$	$\bar{Y}_2$	β
M4R5	*Pseudo-SRS*	0.09	0.15	0.13
M4R5	*SynRep-R*	0.11	0.17	0.13
M4R5	*SynRep-1*	0.17	0.26	0.22
M10R5	*Pseudo-SRS*	0.01	0.02	0.02
M10R5	*SynRep-R*	0.01	0.03	0.02
M10R5	*SynRep-1*	0.04	0.09	0.07

We next turn to compare *SynRep-R* and *SynRep-1* with other approaches, particularly *Wtreg*, *Wtreg-Boot*, and *SRSsyn*. Here, we set $M = 10$ and, where relevant, $R = 10$, and draw 500 repeated samples. Figure 3.4 summarizes the repeated sampling performances of the methods that account for survey weights. For all these methods, the point estimators have simulated percent biases that typically are negligible. For *SynRep-R* and *SynRep-1*, the average variance estimates are close to their corresponding empirical variances, and the coverage rates are close to nominal. For *Wtreg* and *Wtreg-Boot*, the variance estimators can underestimate the corresponding empirical variances severely, especially for $\bar{Y}_1$ and $\bar{Y}_2$, resulting in confidence interval coverage rates that can be substantially lower than the nominal 95% level. The bootstrap step in *Wtreg-Boot* results in more reliable variance estimates compared to *Wtreg*, but *Wtreg-Boot* is not as well calibrated as *SynRep-R* and *SynRep-1*, which have closer to nominal coverage rates. As expected, *HT* results in accurate estimates with near nominal coverage rates. We note that Figure 3.4 does not display results for *Direct* and *SRSsyn* because they perform poorly for $\bar{Y}_1$ and $\bar{Y}_2$. For these two methods, the simulated biases for $\bar{Y}_1$ and for $\bar{Y}_2$ are around 16% and 11%, respectively, with coverage rates near 0 and near 30%, respectively. These results emphasize the importance of accounting for informative designs when generating fully synthetic data that can be analyzed as simple random samples.

Figure 3.4 Repeated sampling properties of different quantities and procedures with $M = 10$ synthetic samples and $R = 10$ replicates under a probability proportional to size design.



Overall, the simulation studies suggest that *SynRep-R* and *SynRep-1* can provide approximately valid inferences, and they are superior inferentially to fully synthetic data that ignore the complex design. The Appendix also includes results of simulation studies where we sample $\mathcal{D}$ via simple random samples. These confirm that the combining rules offer reasonable performance even without unequal probabilities of selection.

4. Illustration with ACS data

We illustrate *SynRep-R* and *SynRep-1* by letting $\mathcal{D}$ be a subset of data from the 2021 ACS Public Use Microdata Sample for $n =$ 84,128 individuals from the state of Michigan. The variables for our illustration include each participant's person-level weight, age, and total income. To mimic the variables in the simulations, we create a binary indicator Y_1 from age that equals one when someone is at least 65 years old; we refer to this indicator as senior status. For purposes of synthesis, we transform income by taking its cubic root. The synthesis models are then a Bernoulli distribution for Y_1 and a linear regression of the cubic root of total income on Y_1. After synthesizing values of the cubic root of income, we raise them to the third power to get incomes on the original scale. We implement each method following the procedures from Section 3. For *SynRep-R* and *SynRep-1*, we set $M = 10$ and $R = 10$.

As population quantities, we estimate the population proportion of senior status individuals $\bar{Y}_1$, the population mean of the income values, $\bar{Y}_2$, and the coefficient β of Y_1 in the linear regression model of the cubic-root transformed income on senior status.

Figure 4.1 presents the point estimates and 95% confidence intervals for the three population quantities. Since *Direct* and *SRSsyn* ignore the sample design, they result in relatively inaccurate results, especially for $\bar{Y}_1$. In contrast, the point estimates for the synthetic data methods that account for survey weights are closer to the *HT* point estimates. Additionally, the 95% confidence intervals for these methods largely overlap with the *HT* confidence intervals. We note, however, that *Wtreg* appears to suffer from underestimation of variance, particularly for β. Additionally, the confidence intervals for the pseudo-likelihood approaches can be narrower than those for *HT*, *SynRep-R*, and *SynRep-1*.

We also can examine potential disclosure risks for the synthetic data methods. Here, we mimic an attack scenario described by Kim et al. (2021), in which we consider an adversary who uses the synthetic data to estimate the largest income value in $\mathcal{D}$. Specifically, we examine differences between the maximum synthetic income in each synthetic dataset and the maximum income in $\mathcal{D}$. This evaluation is not intended to illustrate a rigorous and thorough process for assessing disclosure risks. Rather, we use this attack scenario mainly to compare the different synthesis procedures.

Table 4.1 presents the distributions of the differences for the synthesis methods that account for the survey design. Overall, the results are reasonably similar across the methods, suggesting they offer similar levels of protection in this scenario. All result in substantial differences between the largest synthetic and

observed incomes. The results suggest that an adversary taking this attack strategy is not likely to estimate the largest income accurately.

Figure 4.1 Point estimates and 95% confidence intervals for $\bar{Y}_1$, $\bar{Y}_2$, and β in the ACS data illustration. Results based on $M = 10$ synthetic samples and $R = 10$ replicates.

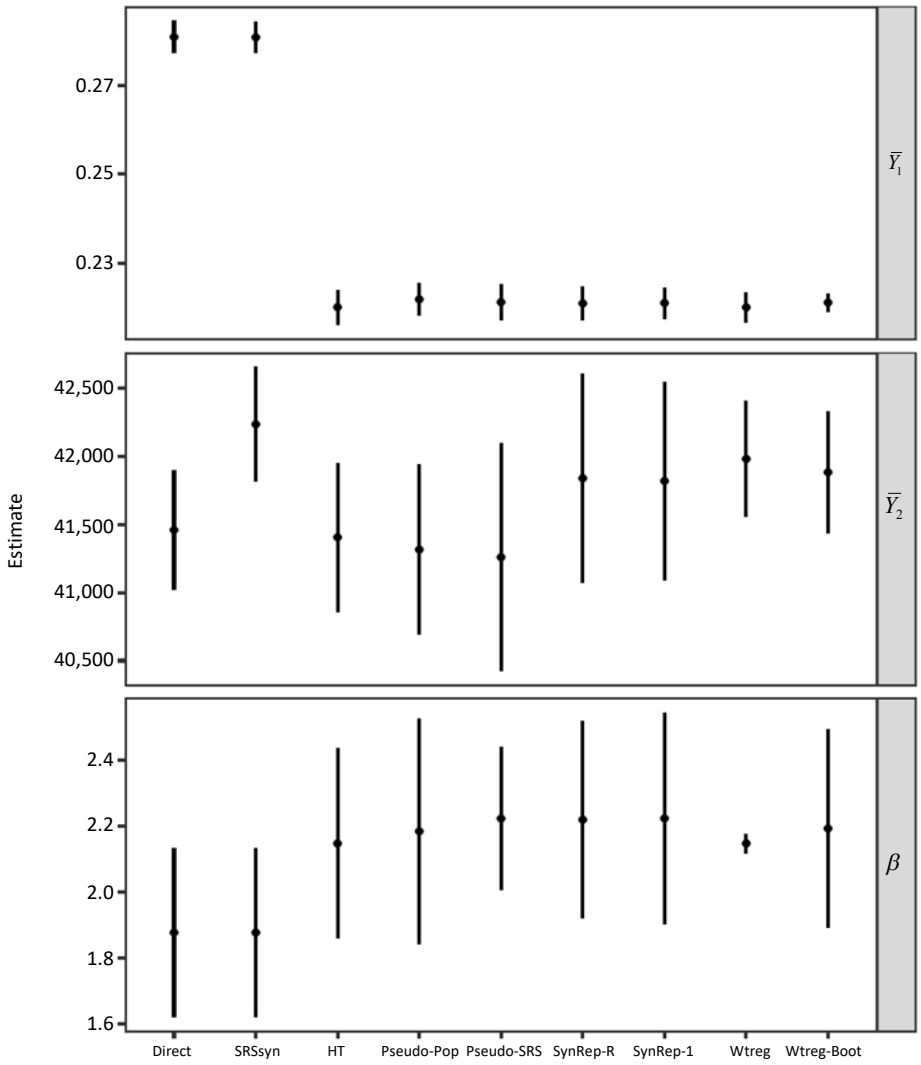


Table 4.1
Summaries of the differences ($) in the largest income value in the synthetic and American Community Survey data. The actual largest value is $1,029,000.

Method	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
SynRep-R	-424,323	-298,230	-252,984	-214,476	-139,874	465,380
SynRep-1	-371,466	-297,180	-287,199	-268,711	-267,428	-40,405
Wtreg	-440,253	-297,689	-242,095	-218,810	-159,766	707,411
Wtreg-Boot	-410,354	-275,398	-209,513	-174,444	-139,109	133,759

5. Discussion

SynRep-R and *SynRep-1* represent a general strategy for constructing fully synthetic data that account for complex sample designs: use the WFPBB to "undo" the design, then replace the confidential values with simulated values. Releasing multiple synthetic data sets, i.e., setting $MR > 1$, can increase statistical efficiency and facilitate variance estimation. However, agencies also can use the WFPBB approach with $MR = 1$. Although releasing a single synthetic data set may not enable approximately valid variance estimation for complex surveys, it still can be useful in certain settings, e.g., when the synthetic data are intended for code training or exploratory analyses where variance estimation is not essential.

As noted by a reviewer, several agencies implementing synthetic data approaches also provide means for users to check the quality of their synthetic data inferences. For example, users can submit their code to the agency that released the synthetic data, which then can run the code and report back disclosure-protected outputs to the user. This is known as validation of results (Barrientos, Bolton, Balmat, Reiter, de Figueiredo, Machanavajjhala, Chen, Kneifel and DeLong, 2018). Alternatively, users can submit queries to a server that computes an analysis of the confidential and synthetic data, and reports back measures of similarity of the two analysis results, e.g., the overlap in the confidence intervals (Karr, Kohnen, Oganian, Reiter and Sanil, 2006). This is known as verification of results (Barrientos et al., 2018). With validation or verification, users of *SynRep-R* and *SynRep-1* may face an additional burden. If the agency directly runs the users' submitted analysis code, the user may need to specify a survey-weighted version of the code for validation, even though they have used a simple random sample analysis for synthetic data. Of course, for many analyses, e.g., regression modeling, some users forego weighted analyses, in which case the issue is moot. It is also possible for the agency to automate validation or verification, in which case it may be able to turn users' submitted queries into survey-weighted versions automatically in the background; this is an area for future research.

We chose to develop methods that enable agencies to follow the idea in Rubin (1993): release data that can be analyzed as simple random samples. This can make analyses easier for users, as they do not have to figure out how to deal with any weights on the file, e.g., in variance estimation. Releasing simple random samples could also help mitigate disclosure risks that may arise from releasing survey weights. For example, if the weights released on the synthetic files are sampled directly from the weight values in $\mathcal{D}$ without alteration, the weights may reveal information about data subjects that is considered an unacceptable disclosure risk (Fienberg, 2010). Finally, releasing simple random samples avoids the need to estimate relationships between the weights and the outcome variables, which could be complicated in practice. Nonetheless, it would be interesting to compare risk and utility profiles of these approaches with those developed here.

There are many other topics related to the general strategy worth further investigation. First, in practice, survey weights can be highly variable and may not be strongly related to the survey variables of interest; this can cause survey-weighted estimates to have inflated variances. This can be remedied somewhat, for example, by using model-based approaches to smooth the weights (Beaumont, 2008; Xia and Elliott, 2016; Si, Trangucci, Gabry and Gelman, 2020). Synthetic data generation based on the WFPBB (or any other

approach) is not immune to these weighting issues. Thus, it would be interesting to examine if and how the synthesis model can reduce the effects of variance inflation from extreme weights.

Second, we focus on developing the fully synthetic data framework and corresponding combining rules, using simple settings and synthesis models to illustrate the methods. Conceptually, agencies can apply *SynRep-R* and *SynRep-1* to multivariate data and for various estimands of interest, e.g., subdomain means and multiple regression coefficients. In such cases, it may be advantageous to use flexible modeling approaches, such as tree-based models or other machine learning algorithms. Future work could investigate the performance of these synthesizers in combination with the pseudo-population and pseudo-SRS generation steps.

Third, we derive the combining rules assuming the original survey data are complete. Agencies could impute missing survey data and generate synthetic replicates simultaneously, possibly accounting for the complex design in the imputation model and synthesis approach. This strategy may necessitate new combining rules akin to those in Reiter (2004).

Fourth, we present *ad hoc* adjustments to deal with negative values of the variance estimates. We may be able to improve on those adjustments. For example, we may be able to adapt the strategy in Si and Reiter (2011), who develop inferential methods for fully synthetic data based on sampling from the distributions used in the derivations of the combining rules. Additionally, as pointed out by a reviewer, it may be beneficial to use the insight of Raab et al. (2018) for the sampling and synthesis components of the derivation in *SynRep-R*. This results in an alternative variance estimator, $(1+M^{-1})\,b_{\text{syn}} - (1+R^{-1})\,\overline{v}_{\text{syn}}$. Future work can investigate the performance of these alternative inference methods.

Fifth, it would be informative to generalize the implementation of *SynRep-R* and *Syn-Rep-1* to other complex designs, such as the stratified multi-stage cluster sampling designs that are common in practice. Zhou, Elliott and Raghunathan (2016) have extended the WFPBB to account for strata, clustering, and survey weights in synthetic population generation. We expect that one could take simple random samples from these pseudo-populations and generate synthetic replicates, possibly using synthesis models that capture design information as suggested in Reiter (2002), and extend the combining rules presented here. It would be a natural extension to comprehensively assess the repeated sampling performances of *SynRep-R* and *Syn-Rep-1* in such multi-stage complex samples.

Lastly, it would be useful to develop principled approaches to measuring disclosure risks for these methods. For *SynRep-R* and *SynRep-1*, conceptually one could estimate an adversary's posterior distribution for confidential data values given the released synthetic values, e.g., as described for simple settings in Reiter, Wang and Zhang (2014) and Hu, Reiter and Wang (2015). However, this would be computationally challenging in practice. One would need to account for the entire synthetic data generation process – including the bootstrapping, sampling, and synthesis – when computing this posterior distribution. Indeed, as far as we are aware, agencies that release synthetic data use *ad hoc* approaches to assessing disclosure risks, such as comparing the similarity of outlier values in the confidential and synthetic data as we illustrated here (Kinney, Reiter and Miranda, 2014). Developing disclosure risk methods is a major area for future research for all approaches to generating fully synthetic data.

Acknowledgements

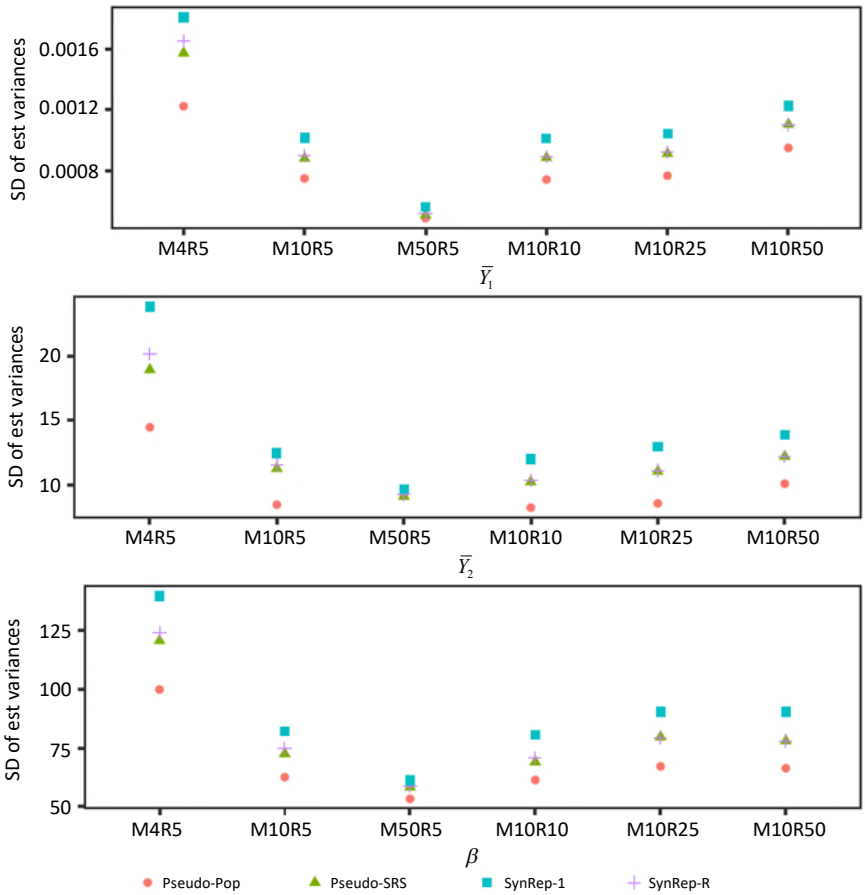
The work was funded by the U.S. National Science Foundation grant (SES 2217456) and a pilot project from the Michigan Center on the Demography of Aging with funding from the National Institute on Aging (P30 AG012846).

Appendix

A. Additional simulation results

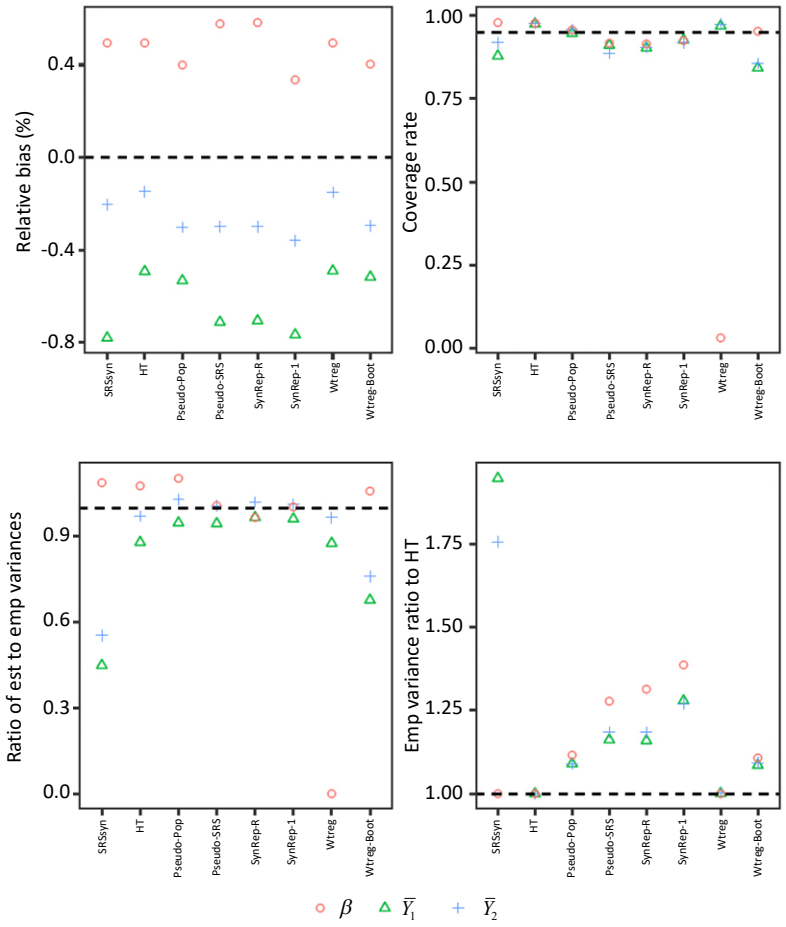
Figure A.1 displays the variability of the 1,000 values of estimated variances of the point estimators for $\beta, \bar{Y}_1,$ and $\bar{Y}_2$ for the simulation with the PPS design. The variability tends to decrease with M. Increasing R when M is held constant seems not to have much impact on the stability of the results. We see increased variability as the procedures introduce more steps that involve randomness; that is, as we go from *Pseudo-Pop* to *Pseudo-SRS* to *SynRep-R* and *SynRep-1*. The variability tends to be largest for *SynRep-1*.

Figure A.1 Standard deviation (SD) of estimated (est) variances of different population quantities with different procedures for different numbers of synthetic samples (M) and replicates (R) under a probability proportional to size design.



As another check of the validity of the combining rules, we repeat the simulations from Section 3 using a SRS in place of a PPS design. Specifically, we use the population described in Section 3.1, but we use a SRS of $n = 500$ records for each $\mathcal{D}$. Figure A.2 displays the results. Overall, the performances of *SynRep-R* and *SynRep-1* mirror the patterns seen for the PPS design in Section 3.

Figure A.2 Repeated sampling properties of different quantities and procedures with $M = 10$ synthetic samples and $R = 10$ replicates under a SRS design.



References

Barrientos, A.F., Bolton, A., Balmat, T., Reiter, J.P., de Figueiredo, J.M., Machanavajjhala, A., Chen, Y., Kneifel, C. and DeLong, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Annals of Applied Statistics*, 12, 1124-1156.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 3, 539-553.

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.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer.

Drechsler, J., and Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

Fienberg, S.E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality*, 1, 183-195.

Gambino, J.G. (2021). R package pps: PPS Sampling. https://cran.r-project.org/web/packages/pps/index.html.

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

Hu, J., Reiter, J.P. and Wang, Q. (2015). Disclosure risk evaluation for fully synthetic data. In *Privacy in Statistical Databases*, (Ed., J. Domngo-Ferrer), 185-199. Heidelberg: Springer.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. and Sanil, A.P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60, 224-232.

Kim, H.J., Drechsler, J. and Thompson, K.J. (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of Royal Statistical Society, Series A*, 184, 255-281.

Kinney, S.K., Reiter, J.P. and Miranda, J. (2014). Synlbd 2.0: Improving the Synthetic Longitudinal Business Database. *Statistical Journal of the International Association for Offcial Statistics*, 30, 129-135.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S. and Abowd, J.M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, 79, 363-384.

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

Meeden, G., Lazar, R. and Geyer, C.J. (2020). R package polyapost: Simulating from the Polya posterior. https://cran.r-project.org/web/packages/polyapost/index.html.

Mitra, R., and Reiter, J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer and L. Franconi), 177-188. New York: Springer-Verlag.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2), 317-337.

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

Raab, G.M., Nowok, B. and Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3), 67-97.

Raghunathan, T.E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, 8, 129-140.

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

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Offcial Statistics*, 18, 531-544.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6785-eng.pdf.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 2, 235-242. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7755-eng.pdf.

Reiter, J.P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, 185-205.

Reiter, J.P. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Offcial Statistics*, 21, 441-462.

Reiter, J.P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters*, 78, 15-20.

Reiter, J.P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review*, 77, 179-195.

Reiter, J.P., and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 20, 405-422.

Reiter, J.P., and Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Offcial Statistics*, 28, 583-590.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

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

Reiter, J.P., Wang, Q. and Zhang, B. (2014). Bayesian estimation of disclosure risks in multiply imputed, synthetic data. *Journal of Privacy and Confidentiality 6*, Article 2.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 130-134.

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

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Offcial Statistics*, 9, 462-468.

Savitsky, T.D., and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10, 1677-1708.

Si, Y., and Reiter, J.P. (2011). A comparison of posterior simulation and inference by combining rules for multiple imputation. *Journal of Statistical Theory and Practice*, 5, 335-347.

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.

Stan Development Team (2024). Stan: A C++ library for probability and sampling. http://mc-stan.org.

United Nations Economic Commission for Europe (2022). Synthetic Data for National Statistical Organizations. https://statswiki.unece.org/display/SDS/Synthetic+Data+Sets+public?preview=%2F282330193%2F330369384%2FHLG-MOS+Synthetic+Data+Guide.docx. Accessed: 2022-01-12.

United States Bureau of the Census (2021). Accessing American Community Survey PUMS data. https://www.census.gov/programs-surveys/acs/microdata/access.html.

Williams, M.R., and Savitsky, T.D. (2021). Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *International Statistical Review*, 89, 72-107.

Xia, X., and Elliott, M.R. (2016). Weight smoothing for generalized linear models using a Laplace prior. *Journal of Offcial Statistics*, 32, 507-539.

Zhou, H., Elliott, M.R. and Raghunathan, T.E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Offcial Statistics*, 32, 231-256.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 375-408
Statistics Canada, Catalogue No. 12-001-X

Models of linkage error for capture-recapture estimation without clerical reviews

Abel Dasylva, Arthur Goussanou and Christian-Olivier Nambeu[1]

Abstract

The capture-recapture method can be applied to measure the coverage of administrative and big data sources, in official statistics. In its basic form, it involves the linkage of two sources while assuming a perfect linkage and other standard assumptions. In practice, linkage errors arise and are a potential source of bias, where the linkage is based on quasi-identifiers. These errors include false positives and false negatives, where the former arise when linking a pair of records from different units, and the latter arise when not linking a pair of records from the same unit. So far, the existing solutions have resorted to costly clerical reviews, or they have made the restrictive conditional independence assumption. In this work, these requirements are relaxed by modeling the number of links from a record instead. The same approach may be taken to estimate the linkage accuracy without clerical reviews, when linking two sources that each have some undercoverage.

Key Words: Big data; Data integration; Data matching; Dual system estimation; Quality; Record linkage.

1. Introduction

The capture-recapture method is an important tool for estimating the coverage of administrative and big data sources that are increasingly used in official statistics (Zhang, 2015). In its simplest form, it estimates the coverage of two sources on the same finite population, by identifying the units selected in both sources, i.e., their intersection, under standard assumptions that include a perfect linkage. Then, the estimated coverage is based on the well-known estimator by Petersen (1896) and Lincoln (1930). However, linkage errors may arise because the linkage is often based on quasi-identifiers such as names and dates. These errors may bias the coverage estimate, which must be corrected.

Regarding this accuracy, a linkage error is defined as a *false negative* or a *false positive*, where a false negative is failing to link records from the same unit, and a false positive is linking records from different units. In connection with these concepts, a record pair is called *matched* if its records are from the same unit (Fellegi and Sunter, 1969; Herzog, Scheuren and Winkler, 2007). Otherwise, it is called *unmatched*. The linkage accuracy may be measured by clerical review, a statistical model, or a combination of both approaches. Clerical reviews consist in the visual inspection of a probability sample of record pairs to determine if they are matched (Dasylva, Abeysundera, Akpoué, Haddou and Saïdi, 2016). They are very flexible and apply regardless of the linkage details. However, they are costly. The alternative to clerical reviews is fitting a statistical model of which quite a few have been proposed, including log-linear mixtures (Fellegi and Sunter, 1969; Thibaudeau, 1993; Winkler, 1993; Daggy, Xu, Hui, Gamache and Grannis, 2013; Chipperfield, Hansen and Rossiter, 2018; Winglee, Valliant and Scheuren, 2005; Chipperfield and Chambers, 2015; Haque, Mengersen and Stern, 2021; Haque and Mengersen, 2022), models of a pair

1. Abel Dasylva, Arthur Goussanou and Christian-Olivier Nambeu, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: abel.dasylva@statcan.gc.ca, arthur.goussanou@statcan.gc.ca, christianolivier.nambeu@statcan.gc.ca.

probabilistic linkage weight (Belin and Rubin, 1995; Sariyar, Borg and Pommerening, 2011), Bayesian models (Fortini, Liseo, Nuccitelli and Scanu, 2001; Tancredi and Liseo, 2011; Sadinle, 2017; Steorts, Hall and Fienberg, 2016), and models based on the number of links from a given record (Blakely and Salmond, 2002; Dasylva and Goussanou, 2022). This latter modeling approach is of special interest in this work, because it is not limited to probabilistic linkages and implicitly accounts for all the interactions among the linkage variables. The modeling approach is not as costly as clerical reviews, but it is less flexible as it relies on assumptions about the linkage procedure. It is also quite challenging when the linkage is constrained to have at most one or exactly one link per record (Lahiri and Larsen, 2005, page 226). Chipperfield and Chambers (2015), and Sadinle (2017) have addressed this issue. However, the proposed methodologies are computer intensive and depend on the restrictive assumption that the linkage variables are conditionally independent, i.e., they are independent given that a pair is matched or unmatched. Indeed, this assumption is a potential source of bias, according to Newcombe (1988, Chapter E.6, page 149), Belin and Rubin (1995) and Blakely and Salmond (2002). Following Larsen and Rubin (2001), it is also possible to combine clerical reviews and statistical modeling to take advantage of the flexibility of the former and the low costs of the latter. However, the overall costs remain beyond the budget of many studies. From the point of view of the capture-recapture method, linkage errors are detrimental because they may bias the estimated coverage. Indeed, a false negative may lead to underestimating the coverage, while a false positive may produce a bias in the opposite direction. Of course, this bias must be removed to accurately estimate the coverage.

Many error correction methods have been described, which make the standard capture-recapture assumptions except for the imperfect linkage, i.e., a closed population, independent units that are selected independently by each source, a homogeneous capture probability by at least one source, and no duplicates or out-of-scope units in either source. They include solutions that require clerical estimates of the linkage accuracy (Ding and Fienberg, 1994; Di Consiglio and Tuoto, 2015; de Wolf, van der Laan and Zult, 2019; Brown, Bycroft, Di Cecco, Elleouet, Powell, Račinskij, Smith, Tam, Tuoto and Zhang, 2020), and other solutions that rely on a statistical model under the conditional independence assumption (Tancredi and Liseo, 2011; Račinskij, Smith and van der Heijden, 2019). Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) and de Wolf et al. (2019) describe three closely related solutions of the former kind, where they constrain the linkage to have at most one link per record, and assume that the false positive probability is negligible for units that are captured by both sources. However, they estimate the linkage accuracy through clerical reviews, which are costly but the only practical solution, given the linkage constraints. Brown et al. (2020) describe a different solution, which also relies on clerical estimates of the linkage accuracy, where the two sources must be linked twice with different linkage procedures, under the assumption that the related link indicators are independent in each matched pair. Instead, Tancredi and Liseo (2011), and Račinskij et al. (2019) use statistical models to jointly estimate the linkage accuracy and the coverage without clerical reviews. However, they make the restrictive conditional independence assumption.

This work aims to jointly estimate the coverage and linkage accuracy without clerical reviews, while relaxing the assumption that the linkage variables are conditionally independent. To that end, a new

methodology is described, which extends a previous model of linkage error (Dasylva and Goussanou, 2022), under the standard capture-recapture assumptions except for the perfect linkage assumption. In this model-based approach, the coverage is estimated by linking the records with a sufficiently high recall, or by specifying the interactions in the matched pairs, while allowing arbitrary interactions in the unmatched pairs. The same models may be used to estimate the recall and precision when linking two sources that each have some undercoverage.

The remaining sections comprise the notations and assumptions, background, proposed methodology, simulations, and conclusion, in this order.

2. Notations and assumptions

In the basic version of the capture-recapture problem, the coverage of a list from a finite population must be estimated by exploiting a second list from the same finite population, under standard assumptions, which include a closed population, independent units that are selected independently by each list, homogeneous capture by at least one list, no duplicates or out-of-scope units in either list and a perfect linkage of the two lists. In general, these lists correspond to probability samples with unknown selection probabilities. In what follows, it is assumed that the standard capture-recapture assumptions hold except for the perfect linkage assumption that is relaxed. Each list is modeled by a Bernoulli sample, where each included unit is associated with a record that possibly contains typos. For example, with person lists, this record may contain the last name and the birth date. The record values are assumed to be independent across the units, but no assumption is made regarding the dependence of these values for records that are from the same unit, or the dependence of the variables within these records. To satisfy the homogeneous capture assumption, it is assumed that a unit is captured in the first list independently of the associated record values (e.g., the recorded last name and birth date). However, capture in the second list may depend on these values. In what follows, denote the cardinality of a set s by $|s|$, and for a tuple $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ let $|\mathbf{x}| = |x_1| + \ldots + |x_d|$.

2.1 Finite population and data sources

The units are from a finite population U with N units that are labeled from 1 to N. The units are selected according to two Bernoulli samples that are denoted by S_A and S_B and identified by two subsets of $\{1, \ldots, N\}$, where it is assumed that the inclusion probability $P(i \in S_A)$ does not depend on N. For example, S_A may be the census of population and S_B may be a coverage survey. In each list where it is included, unit i is associated with a record, whose value is its defining characteristic. In what follows, we let the term record also refer to this value where it is clear from the context. The record value is assumed to live in the record space $\mathcal{V}_N$, which is finite but possibly large. For example, the record space may comprise all the strings written with no more than 32 alphabetical characters, if linking with the surname. In S_B, this record is denoted by V_i. For S_A, the labeling of the records depends on a uniformly random permutation

$\Pi(.)$ of $\{1, \ldots, N\}$, to model the complete lack of information about the records associated with the same unit. In this list, unit i is associated with the record $V'_{\Pi(i)}$. The use of a random unknown permutation is a common device in the record linkage literature (Lahiri and Larsen, 2005; Chambers, 2009). It is mathematically convenient to view the two lists as samples drawn from conceptual registers A and B, where each unit is associated with a record. Then, the recording and list inclusion processes are assumed to be such that the observations

$$\left[\left(I(i \in S_A), I(i \in S_B), V_i, V'_{\Pi(i)}\right)\right]_{1 \le i \le N}$$

are independent and identically distributed and independent of the random permutation $\Pi(.)$. This means that the two lists are labeled independently and that one may consider only the case where the permutation $\Pi(.)$ is the identity (i.e., conditioning on $\Pi(.)$ being the identity in what follows), without loss of generality. Then, unit i is associated with V_i in S_B and V'_i in S_A. In the discussion that is to follow, all the arguments are conditional on $\Pi(.)$ being equal to the identity. However, this information is omitted to simplify the notation. To satisfy the homogeneous capture assumption, it is assumed that capture in S_A is independent of V_i and V'_i. However, capture in S_B may depend on these records, i.e. we may have

$$P\left(i \in S_B \,\middle|\, V_i, V'_i\right) \neq P(i \in S_B).$$

For example, the capture probability may vary across post-strata based on V_i.

2.2 Record linkage and related errors

The indicator of a link between V_i and V'_j is denoted by L_{ij} and called *linkage decision* for the pair (i, j). Let n_i denote the number of links from V_i in S_B, i.e.,

$$n_i = \sum_{j \in S_A} L_{ij}. \tag{2.1}$$

The linkage is assumed to be such that L_{ij} is only a function of V_i and V'_j, i.e., the decision to link two records does not involve other records. In the current setup, this assumption precisely means that

$$E\left[L_{ij} \,\middle|\, \begin{matrix} \left[\left(I(k \in S_A), I(k \in S_B), V_k, V'_k\right)\right]_{k \in \{1, \ldots, N\} \setminus \{i, j\}}, \\ (i, j) \in S_A \times S_B, V_i, V'_j \end{matrix} \right] = E\left[L_{ij} \,\middle|\, (i, j) \in S_A \times S_B, V_i, V'_j \right].$$

This condition covers a wide range of practical linkage strategies that may be implemented with the probabilistic, deterministic or hierarchical methods. However, it excludes linkages that constrain the number of links per record (e.g., exactly one or at most one) even if such linkages may be built from simpler ones, which meet the condition. A record pair is denoted by an element (i, j) of $S_A \times S_B$. As mentioned before, a pair is called *matched* if the two records are from the same unit. Otherwise, it is called *unmatched*. To discuss the linkage errors, a *false negative* is a matched pair that is not linked, a *false positive* is an

unmatched pair that is linked, and a *true positive* is a matched pair that is linked. For completeness, define a *true negative* as an unmatched pair that is not linked. For convenience, let FN, FP, TP and TN denote the total numbers of false negatives, false positives, true positives and true negatives, respectively. It is common to represent the different kinds of record pairs in a 2×2 table called a confusion matrix where the off-diagonal cells represent the errors, as shown in Table 2.1. The linkage accuracy is typically measured by the recall and the precision, where the *recall* is the proportion of matched pairs that are linked (i.e. $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$) and the *precision* is the proportion of linked pairs that are matched (i.e. $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$). It is also measured by the *false negative rate*, which is the proportion of matched pairs that are not linked (i.e., $\mathrm{FN}/(\mathrm{TP}+\mathrm{FN})$), and the *false positive rate* (FPR), which is the proportion of unmatched pairs that are linked (i.e., $\mathrm{FP}/(\mathrm{TN}+\mathrm{FP})$). With a perfect linkage, the precision and recall are equal to 1.0, while the FPR is null. In this ideal situation, the population size is estimated according to Petersen (1896) and Lincoln (1930) by

$$\hat{N}=\frac{|S_A||S_B|}{|S_A\cap S_B|}.$$

Consequently, the estimated coverage of S_A is given by $|S_A\cap S_B|/|S_B|$, while that of S_B is given by $|S_A\cap S_B|/|S_A|$. With linkage errors, the intersection of the two lists is not directly observed. Instead, the size of this intersection must be inferred from the observed links and the linkage accuracy that can be estimated by modeling the number of links from a given record.

Table 2.1
Confusion matrix.

	Link	No link
Matched	TP	FN
Unmatched	FP	TN

3. Background

This section provides some background on the error model (Dasylva and Goussanou, 2022), which is to be adapted for the problem in hand. This model applies when S_A is a census (i.e., $S_A=U$) and the linkage is such that the decision to link two given records involves no other record. In this case the linkage accuracy may be estimated by modeling the distribution of the number of links from a given record, without assumptions about the dependence of the linkage variables.

3.1 How the errors relate to the number of links from a record

In general, there is a strong connection between the number of links from a given record and the related linkage errors. This connection is described in Table 3.1 when S_A is a census (Dasylva and Goussanou,

2020). When $n_i = 0,$ it is known that there is no false positive but one false negative because S_A is a census. When $n_i = 1, \ldots, N-1,$ there is either no or one false negative and thus n_i or $n_i - 1$ false positives, according to whether the record is linked to the matched census record or not, because S_A is a census and it has no duplicate records. When $n_i = N,$ it is known that there are no false negatives and $N-1$ false positives, for the same reasons. As an illustration, consider the example shown in Table 3.2, where $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\},$ $S_B = \{2, 3\}$ and the nature of each record pair is also shown. In this example there are four links including $(2,1), (2, 2), (3, 2)$ and $(3, 4).$ It can be verified that n_i is related to the linkage errors according to Table 3.2. When $n_i = 1, \ldots, N-1,$ the errors are not fully known and may be predicted with a model.

Table 3.1
Connection between n_i and errors when S_A is a census.

n_i	False negatives	False positives
0	1	0
$1 \le n_i \le N-1$	0 or 1	$n_i - 1$ or n_i
N	0	$N-1$

Table 3.2
Example, where $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\}$ and $S_B = \{2, 3\}$ and the links are indicated by the check marks.

	$j =$							
	1	2	3	4	5	n_i	# FN	# FP
$i = 2$	√ FP	√ TP	TN	TN	TN	2	0	1
3	TN	√ FP	FN	√ FP	TN	2	1	2

3.2 A model for homogeneous records

Blakely and Salmond (2002) model n_i by the sum of a Bernoulli variable (for the true positives) with an independent binomial variable (for the false positives) and they estimate the related parameters through a quadratic equation. However, the n_i distribution must be the same for all the records. Otherwise, the estimator may be biased or fail to exist (Dasylva and Goussanou, 2022) if the quadratic equation has no solution. In practice, this issue may arise when linking with names or other characteristics, which occur with different frequencies in the population.

3.3 A model for heterogeneous records

To address the problem, Dasylva and Goussanou (2022) have extended the model from (Blakely and Salmond, 2002) into a finite mixture, which applies when N gets large under regularity conditions. To describe these conditions, let

$$\mathcal{V}_N^* = \left\{ v \in \mathcal{V}_N \text{ s.t. } P\left(V_i = v \middle| i \in S_B\right) > 0 \right\}. \tag{3.1}$$

In other words, $\mathcal{V}_N^*$ is the subset of record values that may be observed in S_B, with a positive probability. At this point, it is useful to consider the subset of all record values (from $\mathcal{V}_N$) that are linked to a particular record value with a positive probability, as well as a superset of this set, which is called *neighborhood* and denoted by $\mathcal{B}_N(v)$ for the value $v \in \mathcal{V}_N^*$. Thus,

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \middle| i \in S_B, \left(V_i, V_j'\right) = (v, v') \right] > 0 \right\}. \tag{3.2}$$

Informally, the neighborhood of a particular record value is a subset of record values that look like this value according to some criterion. For example, consider linking records based on the last name in capital letters and suppose that two records are linked if they agree exactly on this variable. In this case the record value (v) "JARO" may be associated with the singleton neighborhood $(\mathcal{B}_N(v) = \{v\})$ {"JARO"}. To refine this example, suppose now that two records are linked if the last names are identical, or they have the same length and differ by a single letter. In this case, the value "JARO" may be associated with the neighborhood

$$\begin{aligned} &\{\text{“AARO”, “BARO”, …, “ZARO”}\} \cup \{\text{“JARO”, “JBRO”, …, “JZRO”}\} \cup \\ &\{\text{“JAAO”, “JABO”, …, “JAZO”}\} \cup \{\text{“JARA”, “JARB”, …, “JARZ”}\}. \end{aligned}$$

The concept of neighborhood is useful when characterizing the discriminating power of the linkage variables, and when articulating regularity conditions for the consistent estimation of the recall and precision without clerical reviews. To describe these conditions, define the functions $p_N(.)$, $\lambda_N(.)$ and $\lambda_N^{(0)}(.)$, which give the true positive probability, the false positive probability and the probability that an unmatched record is in the neighborhood, i.e.

$$p_N(v) = E\left[L_{ii} \middle| i \in S_B, V_i = v \right], \tag{3.3}$$

$$\lambda_N(v) = E\left[L_{ij} \middle| i \in S_B, V_i = v \right], j \neq i, \tag{3.4}$$

$$\lambda_N^{(0)}(v) = P\left(V_j' \in \mathcal{B}_N\left(V_i\right) \middle| i \in S_B, V_i = v\right). \tag{3.5}$$

Then, the first two regularity conditions are given by the following equations, where Λ is positive and finite, F is a bivariate distribution with support contained in $[0,1] \times [0, \Lambda]$ and neither depends on N.

$$\sup_{v \in \mathcal{V}_N^*} (N-1)\, \lambda_N^{(0)}(v) \leq \Lambda, \tag{3.6}$$

$$\left(p_N(V_i), (N-1)\, \lambda_N(V_i) \right) \middle| \{ i \in S_B \} \xrightarrow{d} F. \tag{3.7}$$

The first condition implies that the expected number of false positives is bounded above for each record. When the true positive probability is bounded below by δ (Dasylva and Goussanou, 2020, equation 6), it also implies that the precision is no less than $\delta / (\delta + \Lambda)$ overall and for any post-stratum, which is defined based on V_i (Dasylva and Goussanou, 2020). The second condition means that the joint distribution of the

expected number of true positives and the expected number of false positives is approximately given by F when N is large. When F is discrete with G atoms, these two conditions imply the following convergence in distribution (Dasylva and Goussanou, 2022, Lemma 1).

$$n_i \mid \{i \in S_B\} \xrightarrow{d} \sum_{g=1}^{G} \alpha_g \text{ Bernoulli}(p_g) * \text{Poisson}(\lambda_g), \tag{3.8}$$

where $*$ denotes the convolution operator. This means that, in the limit, a record belongs to one of G latent classes, where α_g is the probability of class g, and p_g and λ_g are the expected numbers of true positives and false positives for the records in this class. The model parameters may be estimated by maximizing the composite likelihood of the n_i's. They are related to the linkage accuracy through the expected numbers of true positives and false positives per record in S_B, which are given by $\overline{p} = \sum_{g=1}^{G} \alpha_g p_g$ and $\overline{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$. Indeed, the recall and precision converge in probability to $\overline{p}$ and $\overline{p}/(\overline{p} + \overline{\lambda})$, respectively, under the following two additional regularity conditions (Dasylva and Goussanou, 2022, Corollary 1), where $i \neq i'$ and c is a positive finite constant not depending on N.

$$NP\left(\mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing \mid \{i, i'\} \subset S_B\right) \leq c, \tag{3.9}$$

$$NP\left(V_i' \in \mathcal{B}_N(V_i) \mid \{i, i'\} \subset S_B, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing\right) \leq c. \tag{3.10}$$

These two conditions mean that records from different units are very likely to have disjoint neighborhoods (3.9), and that matched records are not far apart (3.10). With the other regularity conditions (i.e., (3.6)-(3.7)), they also imply the consistency of the composite maximum likelihood estimators (Dasylva and Goussanou, 2022, Theorem 3).

Overall, this methodology has several advantages over alternative model-based solutions, because it seamlessly accounts for the interactions among the linkage variables and the records' heterogeneity. Besides, the model fit may be tested using the procedure described by Dasylva and Goussanou (2024), to account for the correlation of the n_i's. When S_A is a census, the methodology may also serve to estimate the false negatives generated by the blocking procedure (Dasylva and Goussanou, 2021). However, it must be adapted when S_A has some undercoverage.

4. Methodology

The proposed methodology is based on two extensions of the model described in the previous section, which is subsequently called the *univariate neighbor model* or simply neighbor model. The first extension accounts for the undercoverage of S_A and only changes the interpretation of some model parameters. It is used to estimate the coverage when linking with a sufficiently high recall. The second extension is more substantial, as it replaces n_i by a vector of such variables; each representing the number of links for a distinct linkage rule. The resulting model is called the *multivariate neighbor model*, which is used to estimate the

coverage by specifying the interactions in the matched pairs, while allowing arbitrary interactions in the unmatched pairs. The following paragraphs discuss the linkage strategy before delving into the details of the various extensions and how they are used to estimate the coverage.

4.1 Linking the sources

In their solutions, Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) and de Wolf et al. (2019) constrain each record to have at most one link. Yet, this constraint greatly complicates the modeling of the linkage errors as mentioned before. Here, it is instead proposed to link the records without this constraint, with a rule such that the decision to link two records involves no other record. Thus, a record may have zero, one or many links. Such a linkage rule may be implemented with the deterministic, hierarchical or probabilistic methods of record linkage.

4.2 Extending the univariate neighbor model

When S_A is not a census, the connection between n_i and the errors is according to Table 4.1, which differs from Table 3.1 when $n_i = 0$ and $n_i = |S_A|$. When $n_i = 0$, there is no certainty about the occurrence of a false negative because it is not known if the corresponding unit is in S_A, unlike what happens in Table 3.1. When $n_i = |S_A|$, the number of false positives is not known with certainty for the same reason. To account for the undercoverage of S_A, redefine $\mathcal{B}_N(v)$ as a subset of $\mathcal{V}_N$ such that

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \,\middle|\, (i,j) \in S_B \times S_A, \left(V_i, V_j'\right) = (v, v') \right] > 0 \right\}. \tag{4.1}$$

Table 4.1
Connection between n_i and errors when S_A is not a census.

n_i	False negatives	False positives
0	0 or 1	0
$1 \le n_i \le \lvert S_A \rvert - 1$	0 or 1	$n_i - 1$ or n_i
$n_i = \lvert S_A \rvert$	0	$\lvert S_A \rvert - 1$ or $\lvert S_A \rvert$

Also, redefine $p_N(.)$, $\lambda_N(.)$ and $\lambda_N^{(0)}(.)$ as

$$p_N(v) = E\left[I\left(i \in S_A\right) L_{ii} \,\middle|\, i \in S_B, V_i = v \right], \tag{4.2}$$

$$\lambda_N(v) = E\left[I\left(j \in S_A\right) L_{ij} \,\middle|\, i \in S_B, V_i = v \right], \; j \neq i, \tag{4.3}$$

$$\lambda_N^{(0)}(v) = P\left(j \in S_A, V_j' \in \mathcal{B}_N\left(V_i\right) \,\middle|\, i \in S_B, V_i = v \right), \tag{4.4}$$

where $p_N(v)$ is the joint probability of including i in S_A and having a true positive, $\lambda_N(v)$ is still the false positive probability and $\lambda_N^{(0)}(v)$ is the probability of having an unmatched record in the neighborhood. With these updated definitions, it is easily shown that (3.8) applies, when $N \to \infty$ under the regularity conditions given by (3.6)-(3.7) and F is discrete with G atoms. Indeed, the proof for the census case still applies with n_i based on (2.1). See Dasylva and Goussanou (2022, Lemma 1) for the details. The parameters α_g and λ_g have the same interpretation, but p_g now corresponds to the product of the inclusion probability $P(i \in S_A)$ by the probability of a true positive for a record in class g. As before, let $\overline{p} = \sum_{g=1}^{G} \alpha_g p_g$ and $\overline{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$, where $\overline{p}$ is the expected number of true positives per record, which is also equal to $E\left[I(i \in S_A)\, L_{ii} \,\middle|\, i \in S_B \right]$, and $\overline{\lambda}$ is the expected number of false positives per record. Thus, $\overline{p}$ is a useful lower-bound on $P(i \in S_A)$. The model parameters may be estimated by maximizing the composite likelihood of the n_i's when G is given, and by selecting this latter parameter through the minimization of Akaike's information criterion as in the census case (Dasylva and Goussanou, 2022). Let $\hat{\overline{p}}$ and $\hat{\overline{\lambda}}$ denote the resulting maximum likelihood estimators. In Appendix A, it is stated that the recall and precision (two finite population parameters) converge in probability to $P(i \in S_A)^{-1} \overline{p} = E\left[L_{ii} \middle| i \in S_A \cap S_B \right]$ and $\overline{p} / (\overline{p} + \overline{\lambda})$, respectively, under the regularity conditions. Under the same conditions, $\hat{\overline{p}}$ and $\hat{\overline{\lambda}}$ are also consistent estimators of $\overline{p}$ and $\overline{\lambda}$, so that $P(i \in S_A)^{-1} \hat{\overline{p}}$ and $\hat{\overline{p}} / (\hat{\overline{p}} + \hat{\overline{\lambda}})$ are consistent estimators of the recall and precision, respectively.

4.3 Estimating the coverage with a high recall

From the above discussion, it follows that a consistent estimator of the coverage $P(i \in S_A)$ may be obtained as

$$\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \right)^{-1} \hat{\overline{p}},$$

if the recall (i.e., $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$) is known. In particular, $\hat{\overline{p}}$ is consistent, if the recall is known to be perfect, i.e., $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) = 1.0$, which is equivalent to having no false negatives. However, it must be noted that the proposed model is of interest only where the linkage is not perfect, i.e., if either the precision, the recall or both are smaller than 1.0. Otherwise, the standard capture-recapture estimator would apply, including in the ideal situation where the linkage key is an error-free unique identifier, i.e., a perfect linkage key. Besides, the neighbor model is not advised with such a key, because some of the model assumptions may not hold.

To use the above approach in practice, one would want to design the linkage rule such that it generates very few false negatives if any, and ideally with a false negative rate smaller than $\min\left(P(i \in S_A), 1 - P(i \in S_A)\right)$ by an order of magnitude. This may be inspired from blocking procedures, which are used in probabilistic linkages to select a small subset of the Cartesian product with the majority of the matched pairs. Christen (2012) provides a good review of these procedures, which are indispensable when the sources are large. However, achieving a sufficiently high recall may come at the expense of tolerating a very low precision, which can prevent the estimation of the coverage with the required accuracy. In such cases, it is

proposed to estimate the coverage by specifying the interactions in the matched pairs. However, this requires a multivariate extension of the neighbor model.

4.4 Multivariate neighbor model

The multivariate extension concerns finite collections of simple linkage rules that are also mutually exclusive, i.e., for each rule, the decision to link two records involves no other records, and each pair is linked by at most one rule. The need for this extension is best explained with an example. For simplicity, suppose that S_A is known to be a complete census and that the two sources are to be linked with the given name, last name and birth date. To do so, seven rules are to be evaluated, which are shown in Table 4.2, where γ lives in the finite set $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\},$ and the components of γ indicate whether there is an exact agreement on the last name, given name and birth date, in this order. For rule $\gamma,$ denote by $n_i^{(\gamma)}$ the corresponding number of links for the sample record $i,$ e.g., $n_i^{(0,0,1)}$ is the number of links when linking based on having the same names but a different birth date. A simple way to evaluate the different rules is to fit a model of the form

$$n_i^{(\gamma)} \Big| \{i \in S_B\} \sim \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \text{Bernoulli}\left(p_g^{(\gamma)}\right) * \text{Poisson}\left(\lambda_g^{(\gamma)}\right), \tag{4.5}$$

separately for each $\gamma,$ where the expected numbers of true positives and false positives per record are $\overline{p}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} p_g^{(\gamma)}$ and $\overline{\lambda}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \lambda_g^{(\gamma)},$ respectively. Note that we necessarily have the constraint $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1,$ because the rules are mutually exclusive. However, the resulting estimators of the recall and precision may be inefficient because important information is ignored, such as the constraint $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1,$ or the correlation among the numbers of links from the different rules for the same sample record, which is not exploited either. Also, when choosing the number of classes $G^{(\gamma)}$ according to Akaike's information criterion, the result $\hat{G}^{(\gamma)}$ may vary across the different rules, which is counterintuitive. Besides, even in the best case where $\hat{G}^{(\gamma)}$ is the same for all the rules, the classes may correspond to different latent partitions of the sample records across the different rules, which is also counterintuitive and undesirable. Furthermore, the above limitations apply in the more general situation where S_A is not a census and its coverage is unknown.

Table 4.2
Mutually exclusive rules based on the given name, last name and birth date.

Rule index $\gamma = (\gamma_1, \gamma_2, \gamma_3)$	Same last name	Same given name	Same birth date
(0,0,1)	x	x	√
(0,1,0)	x	√	x
(0,1,1)	x	√	√
(1,0,0)	√	x	x
(1,0,1)	√	x	√
(1,1,0)	√	√	x
(1,1,1)	√	√	√

The solution is to model the joint distribution of the vector of counts $\left[n_i^{(\gamma)}\right]_{\gamma\in\Gamma}$ with a multivariate extension of the neighbor model, as follows, with further details in Appendix B. To describe this extension, it is convenient to define the following multivariate distributions. The first distribution is the joint distribution of mutually independent variables that are indexed over a finite set Γ, where variable γ follows the $\text{Poisson}\left(\lambda^{(\gamma)}\right)$ distribution. Thus, the joint distribution is simply the product distribution. For notational convenience, we define $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$ and denote this distribution by $\text{PPoisson}\left(\boldsymbol{\lambda}\right)$, where the first "P" stands for product. The second distribution corresponds to the joint distribution of the cell counts in a multinomial experiment with n trials, where the last cell is excluded, the other cells are indexed over Γ, and the probability of observing cell γ is denoted by $p^{(\gamma)}$, such that $\sum_{\gamma\in\Gamma} p^{(\gamma)} \le 1$. In this case, we define $\mathbf{p} = \left[p^{(\gamma)}\right]_{\gamma\in\Gamma}$ and denote the joint distribution by $\text{IMultinomial}\left(n, \mathbf{p}\right)$, where the "I" stands for incomplete. In general, the multivariate extension may be considered for modeling the joint distribution of the numbers of links, which result from the application of mutually exclusive simple linkage rules that are indexed over some finite set Γ, where $n_i^{(\gamma)}$ denotes the number of links from rule γ for the sample record i and $\mathbf{n}_i = \left[n_i^{(\gamma)}\right]_{\gamma\in\Gamma}$. The multivariate model is a finite mixture of $|\Gamma|$-variate discrete distributions, where each component is the convolution of an incomplete multinomial distribution with a product of independent Poisson distributions, i.e.,

$$\mathbf{n}_i \,\big|\, \{i \in S_B\} \sim \sum_{g=1}^{G} \alpha_g \,\text{IMultinomial}(1, \mathbf{p}_g) * \text{PPoisson}\left(\boldsymbol{\lambda}_g\right), \tag{4.6}$$

where G is the number of record classes, α_g is the probability that a sample record is from class g, and $\mathbf{p}_g = \left[p_g^{(\gamma)}\right]_{\gamma\in\Gamma}$ and $\boldsymbol{\lambda}_g = \left[\lambda_g^{(\gamma)}\right]_{\gamma\in\Gamma}$ are the vectors of expected numbers of true positives and false positives for a record in the class. Furthermore, $p_g^{(\gamma)}$ is the expected number of true positives and $\lambda_g^{(\gamma)}$ is the expected number of false positives, under rule γ. Then, given G, the model is parametrized by $\left[\left(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\right)\right]_{1\le g\le G}$. When the records are homogeneous,

$$\mathbf{n}_i \,\big|\, \{i \in S_B\} \sim \text{IMultinomial}(1, \mathbf{p}) * \text{PPoisson}\left(\boldsymbol{\lambda}\right), \tag{4.7}$$

where $\mathbf{p} = \left[p^{(\gamma)}\right]_{\gamma\in\Gamma}$ and $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$. Furthermore, if $\min_{\gamma\in\Gamma}\lambda^{(\gamma)} > 0$ and $\mathbf{t} = \left[t^{(\gamma)}\right]_{\gamma\in\Gamma}$, we have

$$\begin{aligned} P\left(\mathbf{n}_i = \mathbf{t} \,\middle|\, i \in S_B\right) &= I\left(|\mathbf{t}| = 0\right)\left(1 - |\mathbf{p}|\right)e^{-|\boldsymbol{\lambda}|} \\ &\quad + I\left(|\mathbf{t}| > 1\right)\Bigg(\left(1 - |\mathbf{p}|\right)\prod_{\gamma\in\Gamma}\frac{e^{-\lambda^{(\gamma)}}\left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}}}{t^{(\gamma)}!} \\ &\qquad + \sum_{\gamma\in\Gamma: t^{(\gamma)}>0} p^{(\gamma)}\,\frac{e^{-\lambda^{(\gamma)}}\left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}-1}}{\left(t^{(\gamma)}-1\right)!}\prod_{\gamma'\in\Gamma\setminus\{\gamma\}}\frac{e^{-\lambda^{(\gamma')}}\left(\lambda^{(\gamma')}\right)^{t^{(\gamma')}}}{t^{(\gamma')}!}\Bigg). \end{aligned} \tag{4.8}$$

Like before, the model is motivated by the convergence in distribution of the vector of counts $\mathbf{n}_i = \left[n_i^{(\gamma)}\right]_{\gamma\in\Gamma}$, when $N \to \infty$, as stated by Lemma 2 in Appendix B. From the multivariate mixture, it follows that the marginal distribution of $n_i^{(\gamma)}$ is still given by (4.5), except that $G^{(\gamma)}$ and $\alpha_g^{(\gamma)}$ are the same for all

the rules, as desired. Also, the record classes correspond across all the rules now. A restricted model is obtained when $\mathbf{p}_g = \varrho\left(\boldsymbol{\beta}_g\right)$ for each class, where $\varrho(.)$ is a known injective function and $\boldsymbol{\beta}_g$ is a vector of regression coefficients of dimension smaller than $|\Gamma|$. This means that the $|\Gamma|$ true positive probabilities for the different rules are not free but bound by fewer regression coefficients, for each record class. Such a restriction is useful when exploiting the information about the variables'interactions in the matched pairs, through a log-linear specification. Then, a multivariate mixture with G components is parametrized by $\left[\left(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\right)\right]_{1\le g\le G}$. According to Theorem 2 in Appendix B, the model parameters (for each proposed parameterization) is estimated consistently by maximizing the composite likelihood of the observed vectors of counts $\left[\mathbf{n}_i\right]_{i\in S_B}$, when G is known or unknown. In the latter case, G may be selected according to the minimum Akaike's information criterion as before.

4.5 Estimating the coverage through the correlation structure

When the records are linked with a perfect recall, the coverage may be estimated with the univariate mixture model as discussed before. Otherwise, the coverage may be estimated with the multivariate neighbor model, where the true positive probabilities are constrained according to the correlation structure of the linkage variables through a log-linear specification. In detail, the proposed multivariate model is based on a collection of simple linkage rules (i.e., each rule is such that the decision to link two records involves no other record), which are themselves based on a first set of simple linkage rules that are divided into K groups. In group k, there are H_k mutually exclusive rules (i.e., a pair is linked by at most one rule from the group) and $L_{ij}^{(k,h)}$ indicates whether the pair (i, j) is linked by rule h. For example, each group may be based on a single variable, and the rules may correspond to different levels of agreement on this variable. For example, for the last name, these levels may comprise exact agreement, typo agreement (excluding exact agreement) and SOUNDEX agreement (excluding exact and typo agreements). However, a group may also involve many variables. A second set of rules is obtained by combining the rules from the first set as follows. Let the index set be $\Gamma = \{0,\ldots,H_1\} \times \ldots \times \{0,\ldots,H_K\} - \mathbf{0}_K$, and for $\gamma = (\gamma_1,\ldots,\gamma_K) \in \Gamma$, let $L_{ij}^{(\gamma)}$ denote the indicator that rule γ links the pair (i, j), where $L_{ij}^{(\gamma)} = 1$ only if $L_{ij}^{(k,\gamma_k)} = 1$ for each k such that $\gamma_k \ge 1$, and $\sum_{h=1}^{H_k} L_{ij}^{(k,h)} = 0$ for each k such that $\gamma_k = 0$. The proposed model is the special case of the multivariate neighbor model (4.6), where $p_g^{(\gamma)}$ has the following form, for a vector of covariates $\mathbf{z}^{(\gamma)}$ and regression coefficients $\mathbf{u}_g$.

$$p_g^{(\gamma)} = \frac{P(i \in S_A)\exp\left(\mathbf{z}^{(\gamma)^\top}\mathbf{u}_g\right)}{1+\sum_{\gamma'\in\Gamma}\exp\left(\mathbf{z}^{(\gamma')^\top}\mathbf{u}_g\right)}. \tag{4.9}$$

For a given number of classes G, the model parameters include $\left[\left(\alpha_g, \mathbf{u}_g, \boldsymbol{\lambda}_\mathbf{g}\right)\right]_{1\le g\le G}$ and $P(i \in S_A)$. The specific form of $\mathbf{z}^{(\gamma)}$ and $\mathbf{u}_g$ depends on the model. This is illustrated in the following example, where the model includes all the main terms and no interactions. This is also saying that the components of γ are independent in the matched pairs. In this case, the coefficient corresponding to the event $\{\gamma_k = l_k\}$ is denoted

by $u_{g,k(l_k)}$. By the dummy coding convention, the coefficient is set to zero if $l_k = 0$. The covariate corresponding to this coefficient is the indicator $I(\gamma_k = l_k)$ so that

$$\mathbf{z}^{(\gamma)} = \left[I(\gamma_1 = 1) \dots I(\gamma_1 = H_1) \dots I(\gamma_K = 1) \dots I(\gamma_K = H_K) \right]^\top, \tag{4.10}$$

$$\mathbf{u}_g = \left[u_{g,1(1)} \dots u_{g,1(H_1)} \dots u_{g,K(1)} \dots u_{g,K(H_K)} \right]^\top. \tag{4.11}$$

In the next example, the model includes all the main terms and second-order interactions, but no higher-order interactions. For $1 \le k_1 < k_2 \le K$, the coefficient of the interaction between the events $\{\gamma_{k_1} = l_{k_1}\}$ and $\{\gamma_{k_2} = l_{k_2}\}$ is denoted by $u_{g,k_1k_2(l_{k_1}l_{k_2})}$. By the same coding convention, the coefficient is set to zero if $l_{k_1} = 0$ or $l_{k_2} = 0$. The covariate associated with the coefficient is the indicator $I\left(\left(\gamma_{k_1}, \gamma_{k_2}\right) = \left(l_{k_1}, l_{k_2}\right)\right)$. In this case, $\mathbf{z}^{(\gamma)}$ and $\mathbf{u}_g$ have more complex expressions. In order to write them in a manner that is tidy, define

$$\mathbf{z}_{1k}^{(\gamma)} = \left[I(\gamma_k = 1) \dots I(\gamma_k = H_k) \right]^\top.$$

Also, denote the right-hand side of (4.10) by $\mathbf{z}_1^{(\gamma)}$, the right-hand side of (4.11) by $\mathbf{u}_{g1}^{(\gamma)}$, and, for $k = 1, \dots, K-1$, further define

$$\mathbf{z}_{2k}^{(\gamma)} = \left(\left[\mathbf{z}_{1(k+1)}^{(\gamma)\top} \dots \mathbf{z}_{1K}^{(\gamma)\top} \right] \otimes \mathbf{z}_{1k}^{(\gamma)\top} \right)^\top,$$

$$\mathbf{u}_{g2k} = \Bigg[\overbrace{u_{g,k(k+1)(11)} \dots u_{g,k(k+1)(H_k 1)}}^{\substack{\text{interaction terms between} \\ \text{level 1 of } \gamma_{k+1} \text{ and} \\ \text{all levels of } \gamma_k}} \dots \overbrace{u_{g,k(k+1)(1H_{k+1})} \dots u_{g,k(k+1)(H_k H_{k+1})}}^{\substack{\text{interaction terms between} \\ \text{level } H_{k+1} \text{ of } \gamma_{k+1} \text{ and} \\ \text{all levels of } \gamma_k}} \dots$$

$$\overbrace{u_{g,kK(11)} \dots u_{g,kK(H_k 1)}}^{\substack{\text{interaction terms between} \\ \text{level 1 of } \gamma_K \text{ and} \\ \text{all levels of } \gamma_k}} \dots \overbrace{u_{g,kK(1H_K)} \dots u_{g,kK(H_k H_K)}}^{\substack{\text{interaction terms between} \\ \text{level } H_K \text{ of } \gamma_K \text{ and} \\ \text{all levels of } \gamma_k}} \Bigg]^\top,$$

where $\otimes$ is the Kronecker product and $\mathbf{u}_{g2k}$ are the interaction terms between γ_k and $\gamma_{k+1}, \dots, \gamma_K$. Then, we have

$$\mathbf{z}^{(\gamma)} = \left[\mathbf{z}_1^{(\gamma)\top} \ \mathbf{z}_{21}^{(\gamma)\top} \dots \mathbf{z}_{2(K-1)}^{(\gamma)\top} \right]^\top,$$

$$\mathbf{u}_g = \left[\mathbf{u}_{g1}^\top \ \mathbf{u}_{g21}^\top \dots \mathbf{u}_{g2(K-1)}^\top \right]^\top.$$

In general, for $t \le K$, the coefficient of the interaction between the events $\{\gamma_{k_1} = l_{k_1}\}, \dots, \{\gamma_{k_t} = l_{k_t}\}$ is denoted by $u_{g,k_1 \dots k_t(l_{k_1} \dots l_{k_t})}$. As before, the coefficient is set to zero by convention if $\min(l_1, \dots, l_t) = 0$. The covariate associated with the coefficient is the indicator $I\left(\left(\gamma_{k_1}, \dots, \gamma_{k_t}\right) = \left(l_{k_1}, \dots, l_{k_t}\right)\right)$. When the model includes all the main terms and interactions of order d or smaller, we have

$$\mathbf{z}^{(\gamma)\top}\mathbf{u}_g = \sum_{t=1}^{d} \sum_{1 \le k_1 < \dots < k_t \le K} \sum_{l_{k_1}=1}^{H_{k_1}} \dots \sum_{l_{k_t}=1}^{H_{k_t}} I\left(\left(\gamma_{k_1}, \dots, \gamma_{k_t}\right) = \left(l_{k_1}, \dots, l_{k_t}\right)\right) u_{g,k_1 \dots k_t(l_{k_1} \dots l_{k_t})}.$$

According to Theorem 2 in Appendix B, this implies that the parameters of the limiting mixture (including the coverage $P(i \in S_A)$) can be estimated consistently by maximizing the likelihood of the $\mathbf{n}_i$'s, under the stated conditions. For example, this methodology may be of interest in the following simple setup, where the linkage is based on exact comparisons of the last name (first group), given name (second group) and birth date (third group), with $K = 3$, $H_1 = H_2 = H_3 = 1$, $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$ and $|\Gamma| = 7$. The coverage may be estimated by maximizing the likelihood of the $\mathbf{n}_i$'s, if the different agreements have no interactions of the third order in matched pairs for each possible value of a record in S_B, i.e., all main terms and second order interactions may be included (this adds up to six parameters besides the coverage $P(i \in S_A)$). In particular, this is true if the agreements are independent in matched pairs. In this case, the solution is related to that described by Račinskij et al. (2019). It is also related to the solution described by Brown et al. (2020) except that it does not resort to clerical reviews. Beyond this special case, the true positives distribution is associated with seven unknown parameters ($P(i \in S_A)$ and the six log-linear parameters) and seven equations (one for $p^{(\gamma)}$ for each γ), for each mixture component.

A few remarks are in order. The first remark is that the proposed model implicitly accounts for all the interactions among the linkage variables in the unmatched pairs, while accounting for all the interactions of order $K-1$ or smaller in the matched pairs, within each record class. Thus, it offers a far greater modeling flexibility than classical log-linear mixtures, while retaining the identification property (see Lemmas 3 and 4) and the ability to consistently estimate the related parameters (see Theorem 2). This is best seen in the simpler case where the true positives distribution is homogeneous across the records and $H_1 = \ldots = H_K = 1$, i.e., K dichotomous comparisons. In this case, it is clear that one cannot use a two-component K-variate log-linear mixture to model the record pairs, while accounting for all the interactions of order $K-1$ or smaller in the matched pairs. Indeed, this entails having at least 2^K free parameters, including $2^K - 2$ parameters for the matched pairs, one parameter for the mixing proportion and at least one parameter for the unmatched pairs. However, there are only 2^K observable patterns and thus only $2^K - 1$ equations to determine the parameters. The same problem occurs when $K = 2$, even when there are no interactions in the distribution of the matched pairs. The second remark is that the added modeling flexibility greatly facilitates the design of linkage rules, which meet the conditions for the consistent estimation of the coverage (see Theorem 2 and Lemma 4).

4.6 Heterogeneous capture and incomplete records

The methodology may be applied when the capture probability varies across post-strata according to covariates that are recorded without errors in each sample, so long as the stated assumptions (see Sections 2.1, 3.3 and the Appendix) hold within each post-stratum. In this case, S_A and S_B correspond to the subsets of records from a post-stratum, where the capture probability may be estimated using one of the neighbor models. Of course, the construction of the post-strata is an important practical question, which is deferred to future work.

Another practical concern is the occurrence of incomplete records in either sample. To discuss the issue without burdening the notation, let S_A and S_B now denote the two samples within a post-stratum, let S'_A and S'_B denote the corresponding subsamples of complete records, and let $P(i \in S'_A)$ denote the coverage of S'_A within the post-stratum, where i is a unit located therein. Little and Rubin (1987, Chapter 1.2) have described different strategies for conducting a statistical analysis in the presence of incomplete records. A first option is to use only the complete records, with or without weighting them to account for the incomplete ones. Two other options include imputing the missing values and the model-based approach, which consists in maximizing the likelihood of the incomplete data. Here, the first option may be considered without reweighting the complete records, when the stated assumptions apply within each post-stratum, including the fact that the inclusion in S'_A and that in S'_B are independent and the inclusion probability in S'_A is uniform. The main idea is to treat the missingness as a second stage of selection and estimate the coverage in two steps as follows. First, apply one of the two proposed methodologies within the post-stratum to obtain an estimate $\hat{P}(i \in S'_A)$ of the coverage of S'_A. Second, estimate the coverage of S_A (within the post-stratum) by $|S_A| / (|S'_A| / \hat{P}(i \in S'_A))$.

5. Simulations

The proposed methodology is evaluated with Monte Carlo simulations comprising 100 repetitions. In a repetition, a finite population is generated with 100,000 individuals, where each individual is assigned a surname and birth date. From this population two Bernoulli samples are drawn, where the surname and birth date are possibly recorded with typos. Then, the samples are linked and the coverage is estimated with the proposed models and according to Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) and Račinskij et al. (2019), for comparison. Different scenarios are considered with different linkage rules, including some where the conditional independence assumption applies and the recall is perfect. The following paragraphs provide more details.

5.1 Finite population and data sources

For the surname and birth date, the frequencies are based on crossing the surname and age distributions from the 2010 US census of population (US Census Bureau, 2020, 2016). For the surname, the relative frequency is computed after excluding the observation "all other surnames". For simplicity, the month is uniformly drawn from $\{1, \ldots, 12\}$ and the day is independently and uniformly drawn from $\{1, \ldots, 30\}$. Consequently, the surname and date components are mutually independent in the population. Two complete registers are created, where the variables are perturbed in the second register. This perturbation is described in terms of exact agreement on the surname, birth day, or birth month, and a baseline criterion, which is defined as having the same surname SOUNDEX and birth year, as well as an absolute difference that is smaller than 2 for both the day and the month. To be more specific, let γ_1, γ_2 and γ_3 denote the indicator variables, which correspond to the satisfaction of the baseline criterion in addition to having the same

surname, birth day or birth month, as shown in Table 5.1. For example, when $\gamma_1 = 1,$ the baseline criterion is satisfied and the surname is identical. When $\gamma_1 = 0,$ the baseline criterion is not met or it is met and the surname is different. In each case, the birth day and birth month may be identical or different. For a given individual, the related records in the first and second registers are hereafter called first and second records, respectively. The second record is obtained by first drawing $\gamma = (\gamma_1, \gamma_2, \gamma_3),$ and then choosing the record value according to the value of the first record and γ. For example, when $\gamma = (0, 1, 1),$ the second record is such that it has the same birth date as the first record but a different surname with the same SOUNDEX code. Using the dummy coding convention, we can write the distribution of γ in log-linear form as

$$P(\gamma) = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1} \gamma_{k_2} u_{k_1 k_2 (11)} + \gamma_1 \gamma_2 \gamma_3 u_{123(111)} \right), \tag{5.1}$$

where the intercept u is a function of the main terms and the interaction terms because $\sum_{\gamma \in \{0,1\}^3} P(\gamma) = 1.$ For simplicity, we choose the parameters such that $u_{1(1)} = u_{2(1)} = u_{3(1)}$ and $u_{12(11)} = u_{13(11)} = u_{23(11)}.$ When $\gamma_1 = 0,$ the surname in the second record is drawn from the other census surnames with the same SOUNDEX code, according to their frequencies. When $\gamma_2 = 0,$ the day in the second record is obtained by randomly increasing or decreasing the day of the first record by 1, with probability 1/2 for each alternative, except when the day is 1 or 30 in the first record, in which case the day is increased by 1 or decreased by 1, respectively. Likewise, when $\gamma_3 = 0,$ the month in the second record is obtained by randomly increasing or decreasing the month of the first record by 1, with probability 1/2 for each alternative, except when the month is 1 or 12 in the first record, in which case the month is increased by 1 or decreased by 1, respectively. From each register, an independent Bernoulli sample is drawn with an inclusion probability of 0.9, which is the actual coverage.

Table 5.1
Indicators of the perturbations in a record.

Indicator	Baseline criterion	Same surname	Same birth day	Same birth month
γ_1	√	√	?	?
γ_2	√	?	√	?
γ_3	√	?	?	√

5.2 Linkage

Two linkage rules are considered, where the first rule links the pairs that meet the baseline criterion, or the subset of these pairs where there is at least one exact agreement on the surname, day of birth or month of birth, depending on the scenario. The resulting links are used to estimate the coverage with the univariate and multivariate neighbor models, and a classical log-linear mixture model that incorporates the conditional independence assumption as described by Račinskij et al. (2019). For a given pair, the vector of outcomes is based on the indicators of exact agreement for the surname, day of birth and month of birth, e.g., $(1, 1, 1)$

for a perfect agreement on the surname and the two date components. In order to estimate the coverage with the methodologies proposed by Ding and Fienberg (1994), and Di Consiglio and Tuoto (2015), a second linkage rule is required, with at most one link per record as well as clerical estimates of the resulting linkage accuracy. This rule is derived from the first one by deleting a link, if at least one involved record has many links. The clerical estimates of the linkage accuracy are based on drawing a simple random sample of 1,000 record pairs, which satisfy the baseline criterion, and using the truth deck. Note that this procedure ignores the false negatives generated by the baseline criterion (akin to a blocking criterion), where they exist.

5.3 Scenarios

Five scenarios are considered. In the first scenario, the conditional independence assumption is satisfied based on $u_{1(1)} = 1,\ u_{12(11)} = 0$ and $u_{123(111)} = 0,$ and the first linkage rule is based on the baseline criterion and it has a perfect recall. In the second scenario, there is a departure from conditional independence due to interactions of the second order, based on $u_{1(1)} = 1$ and $u_{12(11)} = 1$ and $u_{123(111)} = 0,$ but the first linkage rule is still based on the baseline criterion. In the third scenario, there is also a departure from conditional independence due to interactions of the second and third order, based on $u_{1(1)} = 1,\ u_{12(11)} = 1$ and $u_{123(111)} = 1/4,$ but there is no change to the first linkage rule. The fourth scenario is identical to the second scenario, except that the first linkage rule now links the pairs that meet the baseline criterion and have at least one exact agreement on the surname, day of birth or month of birth. Finally, the fifth scenario is identical to the fourth scenario excepts that a third order interaction is added based on $u_{123(111)} = 1/4.$ The characteristics of the different scenarios are summarized in Tables 5.2 and 5.3. In the latter table, the figures are based on averages across the repetitions.

Table 5.2
Simulation scenarios.

Scenario	Log-linear parameters			Conditional independence
	$u_{k(1)}$	$u_{k_1k_2(11)}$	$u_{123(111)}$	
1	1	0	0	√
2 & 4	1	1	0	x
3 & 5	1	1	1/4	x

Table 5.3
Empirical averages of the rates of linkage error.

Scenario	Linkage	Recall	Precision	False positive rate (FPR) $\times 10^{-9}$	Perfect recall
1	1	1.000	0.952	498.92	√
	2	0.944	1.000	4.15	x
2 & 3	1	1.000	0.953	495.03	√
	2	0.950	1.000	4.17	x
4 & 5	1	0.996	0.964	371.84	x
	2	0.957	1.000	3.48	x

5.4 Estimators

The neighbor models are applied under a homogeneous distribution of the true positives, to reflect the current setup and the situation in practice, where the heterogeneity of the false positives distribution is expected to be the dominant source of heterogeneity for the n_i distribution (Dasylva and Goussanou, 2022). This means that the probability p_g and the vector $\mathbf{p}_g = \left[p_g^{(\gamma)} \right]_{\gamma \in \Gamma}$ are the same for all the classes. It also means that the parameter $\boldsymbol{\beta}_g$ is the same across the classes, if $\mathbf{p}_g$ is of the form $\varrho\left(\boldsymbol{\beta}_g\right)$ for some known function $\varrho(.)$. For convenience, let p, $\mathbf{p}$ and $\boldsymbol{\beta}$ denote the common values of p_g, $\mathbf{p}_g$ and $\boldsymbol{\beta}_g$, respectively. Also let

$$r^{(\gamma)} = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1} \gamma_{k_2} u_{k_1 k_2 (11)} \right), \gamma \in \Gamma = \{0,1\}^3, \tag{5.2}$$

where $\sum_{\gamma \in \{0,1\}^3} r^{(\gamma)} = 1$, the intercept u is a function of the other parameters, and the right-hand side only includes interactions of the second order unlike that of (5.1). Then

$$\boldsymbol{\beta} = \left(u_{1(1)}, u_{2(1)}, u_{3(1)}, u_{12(11)}, u_{13(11)}, u_{23(11)} \right),$$

and the mapping $\varrho(.)$ is characterized by

$$p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}, \gamma \in \Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}.$$

The estimates are computed by maximizing the likelihood numerically in R, where the number of classes is chosen by minimizing Akaike's information criterion. For the univariate model, the estimates are based on capping the n_i's by 10 (i.e., replacing n_i by $\min(10, n_i)$) and maximizing the likelihood of the resulting observations, as described by Dasylva and Goussanou (2022). With the multivariate model, the coverage is estimated when only including the main terms, and when also including the second order interaction terms, while ignoring that the main terms are equal, and that the second order interaction terms are equal. In general, the numerical maximization of the likelihood is more challenging than with the univariate model, and the resulting estimates become less accurate when the linkage precision decreases. Consequently, a good initialization procedure is needed, which is described in the Appendix C. For comparison, we also compute the estimators proposed by Ding and Fienberg (1994), Di Consiglio and Tuoto (2015), and Račinskij et al. (2019), as well as the naive capture-recapture estimator, which ignores the linkage errors. Note that this latter estimator is computed as the ratio of the number of links by the second linkage rule over $|S_B|$.

5.5 Results

The simulation results are shown in Table 5.4. In scenario 1, where the conditional independence assumption applies, the best performance is obtained with the estimators by Račinskij et al. (2019) and the neighbor estimators, both in terms of the relative bias and the mean square error, with an advantage for the neighbor estimators when looking at this latter performance measure. Among the neighbor estimators, the

univariate model offers the best performance in terms of bias, variance and mean square error. Without surprise, the naive estimator has the worst performance, while the estimators by Ding and Fienberg (1994), and Di Consiglio and Tuoto (2015) perform better but have a large variance, because they incorporate clerical estimates of the linkage accuracy. It is notable that the log-linear mixture proposed by Račinskij et al. (2019) has a larger bias, variance and mean square error, than the estimators based on the multivariate neighbor models, with one small exception. (In Table 5.4, the log-linear mixture has a relative bias slightly smaller that of the multivariate neighbor model with interactions.) Indeed, all these estimators aim to estimate the coverage by leveraging the correlation structure of the linkage variables, which the log-linear mixture does fully by incorporating the independence of the linkage variables both in the matched pairs and in the unmatched pairs. However, the multivariate neighbor models only exploit the information about the correlation structure in the matched pairs without constraint on the unmatched ones. Yet, they yield estimators that are significantly more accurate than the log-linear mixture in terms of mean square error. This illustrates the important difference between classical log-linear mixtures and the multivariate neighbor models, when the latter incorporate a log-linear specification of the correlation structure in the matched pairs.

In scenario 2, the proposed estimators still offer the smallest mean square errors. As before, the univariate model offers the best overall performance in terms of bias, variance and mean square error. Of the two multivariate models, the one including the interactions performs better as one might expect, with a bias that is about forty times smaller and a mean square error that is about five times smaller. Without surprise, the estimator by Račinskij et al. (2019) has a worse performance than in the previous scenario, because the conditional independence assumption is violated in this scenario. However, this degradation is such that it has a worst performance than the naive estimator for each performance measure. An interesting observation is that it also performs much worse than the estimator based on the multivariate neighbor model without interactions, with a relative bias and mean square error that are bigger by more than an order of magnitude, while this latter estimator performs better than the naive estimator for each performance measure. A possible explanation is that the multivariate neighbor model implicitly accounts for the all the interactions in the distribution of the unmatched pairs, while the model by Račinskij et al. (2019) ignores these interactions. This is a further illustration of the difference between classical log-linear mixtures and multivariate neighbor models. As before, the estimators by Ding and Fienberg (1994), and Di Consiglio and Tuoto (2015) have a worse performance than the neighbor models, due to the variance of the estimated linkage accuracy. However, in terms of bias and mean square error, they perform better than the naive estimator and that from Račinskij et al. (2019). Scenario 3 differs from scenario 2 by adding a third order interaction with coefficient 1/4. However, this change has a negligible impact on the obtained results and observed trends.

In scenario 4, the first linkage rule has a small false negative rate of about 0.4% (i.e., an imperfect recall), by not linking the pairs that have no exact agreement on the surname, day of birth or month of birth. This has a direct impact on the univariate neighbor estimator, which now has the third smallest mean square error, behind the two multivariate neighbor estimators; the one with the interactions offering the best

performance. Excluding the pairs with no exact agreement further degrades the performance of the log-linear mixture estimator (compared to the scenarios 3 and 4), which still has the largest mean square error and a worst performance than the naive estimator. However, this change has a limited impact on the performance of the estimators by Ding and Fienberg (1994), and Di Consiglio and Tuoto (2015). Scenario 5 differs from scenario 4 by adding a third order interaction with coefficient 1/4. However, this change has a negligible impact on the results.

In summary, the simulation results demonstrate that the proposed estimators may be used to estimate the coverage with a small relative bias and a smaller mean square error than the alternative estimators proposed by Ding and Fienberg (1994), Di Consiglio and Tuoto (2015), and Račinskij et al. (2019), when the false negatives are negligible or the true positive probabilities are constrained by a log-linear specification.

Table 5.4
Simulation results.

Scenario	Estimator	Relative bias (%)	Variance $\times 10^{-7}$	Mean square error $\times 10^{-7}$
1	Naive	-5.522	12.90	24,711.31
	R	-0.034	824.62	817.32
	DF	1.618	2,377.74	4,475.68
	DT	1.159	2,550.13	3,613.51
	UN	-0.003	8.43	8.35
	MN with no interactions	-0.023	28.25	28.41
	MN with 2nd order interactions	-0.119	27.06	38.25
2 & 3	Naive	-4.994	15.10	20,216.56
	R	-7.784	324.65	49,403.40
	DF	1.667	3,381.80	5,598.43
	DT	0.961	3,560.11	4,272.01
	UN	-0.004	8.31	8.24
	MN with no interactions	-0.423	21.05	165.44
	MN with 2nd order interactions	-0.091	23.70	30.12
4 & 5	Naive	-4.292	15.46	14,934.57
	R	-3.497	280,414.54	287,515.73
	DF	1.760	2,255.34	4,740.96
	DT	1.159	2,550.13	3,613.51
	UN	-0.393	9.30	134.50
	MN with no interactions	-0.423	21.05	165.44
	MN with 2nd order interactions	-0.091	23.70	30.12

DF: estimator by Ding and Fienberg (1994)
DF: estimator by Di Consiglio and Tuoto (2015)
MN: estimator based on the multivariate neighbor model
Naive: Lincoln-Petersen estimator that ignores the linkage errors
R : estimator by Račinskij et al. (2019)
UN: estimator based on the univariate neighbor model

6. Conclusion

A new methodology has been described for capture-recapture estimation with linkage errors, which is based on modeling the number of links from a record, without clerical reviews, including a univariate model and a related multivariate model. With the univariate model, the coverage is estimated by linking the records with a sufficiently high recall. With the multivariate model, the coverage is estimated by constraining the interactions in the matched pairs through a log-linear specification, while allowing arbitrary interactions in the unmatched pairs; a major difference with classical log-linear mixtures. In this latter case, the records must be linked with a high precision to obtain a reliable estimate of the coverage. Simulations with public census data demonstrate the good performance of the proposed estimators, when compared to previous solutions. Future work will look at obtaining variances and confidence intervals, and at validating the log-linear specification when the multivariate model is used.

Acknowledgements

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada.

Appendix

A. Extension for undercoverage

This appendix aims to extend the results from Dasylva and Goussanou (2022) to show that

$$\frac{\text{TP}}{\text{TP} + \text{FN}} \xrightarrow{p} P(i \in S_A)^{-1}\, \overline{p},$$

$$\frac{\text{TP}}{\text{TP} + \text{FP}} \xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}},$$

$$\hat{\overline{p}} \xrightarrow{p} \overline{p},$$

$$\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} \xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}}.$$

Therefore, $P(i \in S_A)^{-1}\, \hat{\overline{p}}$ and $\hat{\overline{p}} \Big/ \left(\hat{\overline{p}} + \hat{\overline{\lambda}}\right)$ estimate the recall and precision consistently, in the sense that

$$P(i \in S_A)^{-1}\, \hat{\overline{p}} - \frac{\text{TP}}{\text{TP} + \text{FN}} \xrightarrow{p} 0,$$

$$\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} - \frac{\text{TP}}{\text{TP} + \text{FP}} \xrightarrow{p} 0.$$

The extension consists in accounting for the undercoverage in S_A. To proceed, some additional notation is needed. Call V'_j a *neighbor* of V_i, if unit j is included in S_A and V'_j is contained in the neighborhood of V_i, i.e., $\mathcal{B}_N(V_i)$. The neighbor is called matched if both records are from the same unit. Otherwise, it is called unmatched. Also, define the following additional notation. For $i, i' \in S_B$ such that $i \neq i'$, let

$$n_i^{(0)} = \sum_{j \in S_A} I\left(V'_j \in \mathcal{B}_N(V_i)\right), \tag{A.1}$$

$$\left(n_{i|U}^{(0)}, n_{i|U}\right) = \left(n_i^{(0)}, n_i\right) - \left(I\left(\{i \in S_A\} \cap \{V'_i \in \mathcal{B}_N(V_i)\}\right), I(i \in S_A)\, L_{ii}\right), \tag{A.2}$$

$$\left(n_{ii'|U}^{(0)}\right) = \sum_{t \in S_A : t \neq i, i'} I\left(V'_t \in \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V'_i)\right), \tag{A.3}$$

where $n_i^{(0)}$ is the number of neighbors of V_i, $n_{i|U}^{(0)}$ is the number of unmatched neighbors of this record, $n_{i|U}$ is the number of unmatched records, which are linked to the same record, and $n_{ii'|U}^{(0)}$ is the number of unmatched neighbors, which are common to V_i and $V_{i'}$. Corresponding to $n_i^{(0)}$ and $n_{i|U}^{(0)}$, let $S_{Ai} = \left\{t \in S_A \text{ s.t. } V'_t \in \mathcal{B}_N\left(V_i\right)\right\}$ denote the subset of units, which are included in S_A and have their record in the neighborhood, and let $S_{Ai|U} = S_{Ai} - \{i\}$ denote the subset of these units that are different from unit i. These latter units are associated with the unmatched neighbors. Finally, for random variables (or vectors) X, Y and Z, denote the independence of X and Y by $X \perp\!\!\!\perp Y$, and their conditional independence given Z by $X \perp\!\!\!\perp Y \mid Z$.

The following lemma extends Lemma 2 in Dasylva and Goussanou (2022). Note that all the proofs are found in the longer version of the paper (Dasylva, Goussanou and Nambeu, 2024).

Lemma 1. *Suppose that* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V'_i\right)\right]_{1 \leq i \leq N}$ *are iid and let* $Z_1, \ldots, Z_N$ *denote identically distributed random variables, such that they are conditionally independent given* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V'_i\right)\right]_{1 \leq i \leq N}$ *with a marginal conditional distribution of* Z_i *that is only a function of* $I(i \in S_A)$, $I(i \in S_B)$, V_i, $S_{Ai|U}$, $\left[V'_t\right]_{t \in S_{Ai|U}}$, *and* V'_i. *Then*

$$\left(I(i' \in S_A), V'_{i'}, \left[V'_t\right]_{t \in S_{Ai'|U}}\right) \perp\!\!\!\perp \left(I(i \in S_A), V'_i, \left[V'_t\right]_{t \in S_{Ai'|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V'_i, S_{Ai'|U}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V'_i \notin \mathcal{B}_N(V_{i'}), \\ V'_{i'} \notin \mathcal{B}_N(V_i) \end{array} \right., \tag{A.4}$$

$$S_{Ai'|U} \perp\!\!\!\perp \left(I(i \in S_A), V'_i, \left[V'_t\right]_{t \in S_{Ai|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V_{i'}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V'_i \notin \mathcal{B}_N(V_{i'}), \\ V'_{i'} \notin \mathcal{B}_N(V_i) \end{array} \right. . \tag{A.5}$$

Also for any fixed $(v_i, v_{i'}) \in \mathcal{V}_N^* \times \mathcal{V}_N^*$, $a_i \in \{0,1\}$, $w_i \notin \mathcal{B}_N(v_{i'})$, $s_{Ai|U} \subset \{1,\ldots,N\} \setminus \{i,i'\}$ *and* $\left[w_t\right]_{t \in s_{Ai|U}}$ *such that* $\mathcal{B}_N(v_i) \cap \mathcal{B}_N(v_{i'}) = \varnothing$ *and* $w_t \in \mathcal{B}_N(v_i)$ *for all* $t \in s_{Ai|U}$

$$P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & i \in S_B, V_i = v_i, \\ V_i' = w_i, & S_{Ai|U} = s_{Ai|U}, \\ \left[V_t'\right]_{t \in S_{Ai|U}} = & i' \in S_B, V_{i'} = v_{i'} \\ \left[w_t\right]_{t \in S_{Ai|U}} & \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ & V_i' \notin \mathcal{B}_N(V_{i'}), \\ & V_{i'}' \notin \mathcal{B}_N(V_i) \end{array}\right) = P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & \\ V_i' = w_i, & i \in S_B, V_i = v_i, \\ \left[V_t'\right]_{t \in S_{Ai|U}} = & S_{Ai|U} = s_{Ai|U}, \\ \left[w_t\right]_{t \in S_{Ai|U}} & V_i' \notin \mathcal{B}_N(v_{i'}) \end{array}\right). \tag{A.6}$$

Hence

$$E\left[Z_i Z_{i'} \,\middle|\, \begin{array}{l} i \in S_B, V_i = v_i, n_{i|U}^{(0)} = k, \\ i' \in S_B, V_{i'} = v_{i'}, n_{i'|U}^{(0)} = \ell, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), V_{i'}' \notin \mathcal{B}_N(V_i) \end{array}\right] = g(v_i, k; v_{i'})\, g(v_{i'}, \ell; v_i), \tag{A.7}$$

where

$$g(v_i, k; v_{i'}) = E\left[Z_i \,\middle|\, \begin{array}{l} i \in S_B, V_i = v_i, \\ n_{i|U}^{(0)} = k, V_i' \notin \mathcal{B}_N(v_{i'}) \end{array}\right]. \tag{A.8}$$

The above lemma leads to the following theorem, which extends Theorem 1 in Dasylva and Goussanou (2022).

Theorem 1. *Consider* V_N, $\mathcal{B}_N(.)$, S_A, S_B, *and* $\left[Z_i\right]_{1 \le i \le N}$ *identically distributed random variables such that the following conditions apply.*

(C.1) $\lim_{N \to \infty} P\left(i \in S_B\right) = \tau$ *for some positive* τ.

(C.2) For $\Lambda \ge 0$ *not depending on* N, $\sup_{v \in V_N^*} (N-1)\, \lambda_N^{(0)}\left(v\right) \le \Lambda$.

(C.3) For $c \ge 0$ *not depending on* N

$$NP\left(\mathcal{B}_N\left(V_i\right) \cap \mathcal{B}_N\left(V_{i'}\right) \ne \varnothing \mid \{i,i'\} \subset S_B\right) \le c,$$

$$NP\left(V_i' \in \mathcal{B}_N(V_{i'}) \,\middle|\, \{i,i'\} \subset S_B,\, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing\right) \le c.$$

(C.4) $Z_1, \ldots, Z_N$ *are conditionally independent given* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_i'\right)\right]_{1 \le i \le N}$ *such that the marginal conditional distribution of* Z_i *is only a function of* $I(i \in S_A)$, $I(i \in S_B)$, V_i, $S_{Ai|U}$, $\left[V_t'\right]_{t \in S_{Ai|U}}$, *and* V_i'.

(C.5) $|Z_i| \leq R_N\left(n_{i|U}^{(0)}\right)$ *where* $R_N(.)$ *is a polynomial with a finite degree not depending on* N *and nonnegative coefficients of* $O\left(\left(\log N\right)^d\right)$, *where* d *does not depend on* N *either.*

(C.6) $\lim_{N\to\infty} E\left[Z_i \,\middle|\, i \in S_B\right] = \mu$, *where* $|\mu| < \infty$.

Then

$$\frac{1}{|S_B|}\sum_{i\in S_B} Z_i \xrightarrow{p} \mu. \tag{A.9}$$

The next result shows the convergence of the recall and precision to $P(i \in S_A)^{-1}\,\overline{p}$ and $\overline{p}/(\overline{p}+\overline{\lambda})$, respectively. It extends Corollary 1 in Dasylva and Goussanou (2022) and is a direct consequence of Theorem 1.

Corollary 1. *Suppose that assumptions C.1-C.3 hold and that the linkage meets the following conditions*

(C.7) $\left[L_{1j}\right]_{1\leq j\leq N}, \ldots, \left[L_{Nj}\right]_{1\leq j\leq N}$ *are conditionally independent given* $\left[\left(I(i\in S_A), I(i\in S_B), V_i, V_i'\right)\right]_{1\leq i\leq N}$, *where the conditional distribution of* $\left[L_{ij}\right]_{1\leq j\leq N}$ *is only a function of* $I(i\in S_A)$, $I(i\in S_B)$, V_i, $S_{Ai|U}$, $\left[V_t'\right]_{t\in S_{Ai|U}}$, *and* V_i'.

(C.8)

$$\lim_{N\to\infty} E\left[\left(p_N(V_i), (N-1)\,\lambda_N(V_i)\right) \;\middle|\; i\in S_B\right] = \left(\overline{p}, \overline{\lambda}\right). \tag{A.10}$$

Then

$$\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right) \xrightarrow{p} \left(\frac{\overline{p}}{P(i\in S_A)}, \frac{\overline{p}}{\overline{p}+\overline{\lambda}}\right). \tag{A.11}$$

In particular, (A.11) holds if C.8 is replaced by the condition

$$\left(p_N(V_i), (N-1)\,\lambda_N(V_i)\right) \;\middle|\; \{i\in S_B\} \xrightarrow{d} F(.,.)$$

with $\overline{p} = \int p\,dF(p,\lambda)$ *and* $\overline{\lambda} = \int \lambda\,dF(p,\lambda)$.

Other results from Dasylva and Goussanou (2022) remain valid, based on Theorem 1 and Corollary 1, such as Lemma 2 (convergence in distribution of n_i to a mixture as in the right-hand side of (3.8)), Theorem 2 (consistency of the estimator by Blakely and Salmond (2002) in the homogeneous case) and Theorem 3 (consistency of the maximum likelihood estimator). This is easily seen by inspecting the related proofs in Dasylva and Goussanou (2022). Therefore, the estimators $\hat{\overline{p}}$ and $\hat{\overline{\lambda}}$ are consistent, and $P(i\in S_A)^{-1}\hat{\overline{p}}$ and $\hat{\overline{p}}\Big/\left(\hat{\overline{p}}+\hat{\overline{\lambda}}\right)$ are consistent estimators of the recall and precision, respectively.

B. Multivariate extension

To describe the multivariate version of the neighbor model, let Γ denote the index set of a collection of rules and let $L_{ij}^{(\gamma)}$ indicate whether the pair (i, j) is linked by rule $\gamma \in \Gamma$. (To avoid any conflict with the previously defined notation, it is assumed that the rules are indexed such that Γ does not contain 0.) The set Γ may take various forms, such as a subset of consecutive integers starting from 1, or it can be a subset of $\{0,1\}^K$ if linking the records with K linkage variables and performing an exact comparison for each variable. For $\gamma \in \Gamma$, let $n_i^{(\gamma)} = \sum_{j \in S_A} L_{ij}^{(\gamma)}$, $\mathbf{n}_i = \left[n_i^{(\gamma)} \right]_{\gamma \in \Gamma}$, and redefine $\mathcal{B}_N(v)$ to be a subset of $\mathcal{V}_N$, which satisfies the condition

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s.t. } E\left[\sum_{\gamma \in \Gamma} L_{ij}^{(\gamma)} \middle| (i, j) \in S_B \times S_A, \left(V_i, V_j'\right) = (v, v') \right] > 0 \right\}.$$

In words, $\mathcal{B}_N(v)$ is a superset of record values, which are linked by at least one rule in the collection with a positive probability, given that $i \in S_B$ and $V_i = v$. The function $\lambda_N^{(0)}(.)$ is still defined by (4.4), while $p_N(.)$ and $\lambda_N(.)$ are replaced by the vectors $\mathbf{p}_N(v) = \left[p_N^{(\gamma)}(v) \right]_{\gamma \in \Gamma}$ and $\boldsymbol{\lambda}_N(v) = \left[\lambda_N^{(\gamma)}(v) \right]_{\gamma \in \Gamma}$, with

$$p_N^{(\gamma)}(v) = E\left[I\left(i \in S_A\right) L_{ii}^{(\gamma)} \,\middle|\, i \in S_B, V_i = v \right], \tag{B.1}$$

$$\lambda_N^{(\gamma)}(v) = E\left[I\left(j \in S_A\right) L_{ij}^{(\gamma)} \,\middle|\, i \in S_B, V_i = v \right], j \neq i, \tag{B.2}$$

and $\sum_{\gamma \in \Gamma} p_N^{(\gamma)}(v) \leq 1$ for all $v \in \mathcal{V}_N^*$. In words, $p_N^{(\gamma)}(v)$ and $\lambda_N^{(\gamma)}(v)$ are the expected numbers of true positives and false positives for rule γ, given that $i \in S_B$ and $V_i = v$. The regularity condition of (3.7) is replaced by the following more general condition.

$$\left(\mathbf{p}_N(V_i),\ (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} F. \tag{B.3}$$

A case of special interest is when $\mathbf{p}_N(v)$ is of the form $\varrho\left(\boldsymbol{\beta}_N(v)\right)$, for some function $\boldsymbol{\beta}_N : \mathcal{V}^* \to \mathbb{R}^m$, and some *injective* function $\varrho : \mathbb{R}^m \to [0,1]^{|\Gamma|}$ independent of N, where $m < |\Gamma|$. In this case, (3.7) is replaced by the following condition instead.

$$\left(\boldsymbol{\beta}_N(V_i), (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} H, \tag{B.4}$$

where H does not depend on N.

The next lemma states the convergence of $\mathbf{n}_i$ to a multivariate mixture, when $N \to \infty$ under the conditions given by (3.6), and (B.3) or (B.4). The mixing distribution is given by F or H depending on whether (B.3) or (B.4) applies. In both cases, the component distributions come from $|\Gamma|$-variate families of discrete distributions, which correspond to the convolution of a multinomial distribution with a product of independent Poisson distributions. To further describe the limiting distributions, let $\mathcal{F}$ denote the family of the component distributions based on (B.3), where each member is of the form

$$\text{IMultinomial}(1, \mathbf{p}) * \text{PPoisson}(\boldsymbol{\lambda}),$$

for $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma \in \Gamma}$ and $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)} \right]_{\gamma \in \Gamma}$. When (B.4) applies, the component distributions come from the subset $\mathcal{F}_\varrho$ of $\mathcal{F}$, where $\mathbf{p} = \varrho(\boldsymbol{\beta})$ for some $\boldsymbol{\beta} \in \mathbb{R}^m$. A member of this family is a parametric distribution with parameters $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\boldsymbol{\lambda} \in (0, +\infty)^{|\Gamma|}$. As before, note that all the proofs are found in the longer version of the paper (Dasylva et al., 2024).

Lemma 2. *Suppose that (3.6) applies. If (B.3) also applies,* $\mathbf{n}_i$ *converges in distribution to the mixture of distributions from* $\mathcal{F}$*, with mixing distribution* F*. If (B.4) also applies,* $\mathbf{n}_i$ *instead converges in distribution to the mixture of distributions from* $\mathcal{F}_\varrho$*, with mixing distribution* H*.*

The next lemma extends Lemma 4 from Dasylva and Goussanou (2022). It gives sufficient conditions for the identification of finite mixtures over $\mathcal{F}$ (or $\mathcal{F}_\varrho$); a key property for proving the consistency of maximum likelihood estimators. The lemma requires a lexicographic order over $(0, \infty)^{|\Gamma|}$. To do so, order the elements of Γ based on some bijection from $\{1, \ldots, |\Gamma|\}$ into Γ, which is denoted by $\gamma(.)$ with a slight abuse of the notation. Next, denote a tuple $\left[\lambda^{(\gamma)} \right]_{\gamma \in \Gamma}$ equivalently by $\left[\lambda^{(\gamma(t))} \right]_{1 \le t \le |\Gamma|}$, and write $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ (i.e., $\boldsymbol{\lambda}$ greater than $\boldsymbol{\lambda}'$), if $\lambda^{(\gamma(1))} > \lambda'^{(\gamma(1))}$ or if there exists $t_0 = 2, \ldots, |\Gamma|$ such that $\lambda^{(\gamma(t))} = \lambda'^{(\gamma(t))}$ for $t < t_0$ and $\lambda^{(\gamma(t_0))} > \lambda'^{(\gamma(t_0))}$. Also, let $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}$ if $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ otherwise let $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}'$.

Lemma 3. *For positive integers* G *and* G'*, let* $\boldsymbol{\lambda}_1 \succ \ldots \succ \boldsymbol{\lambda}_G$*, and* $\boldsymbol{\lambda}'_1 \succ \ldots \succ \boldsymbol{\lambda}'_G$*, and denote by* h_g *and* h'_g *the members of* $\mathcal{F}$ *with the parameters* $(\mathbf{p}_g, \boldsymbol{\lambda}_g)$ *and* $(\mathbf{p}'_g, \boldsymbol{\lambda}'_g)$*, respectively, and suppose that the mixtures* $h = \sum_{g=1}^{G} \alpha_g h_g$ *and* $h' = \sum_{g=1}^{G'} \alpha'_g h'_g$ *are equal, where* α_g *and* α'_g *are positive for each* g*. Then* $G = G'$*,* $\alpha_g = \alpha'_g$ *and* $(\mathbf{p}_g, \boldsymbol{\lambda}_g) = (\mathbf{p}'_g, \boldsymbol{\lambda}'_g)$*, for* $g = 1, \ldots, G$*. Furthermore, if there exists an injective function* $\varrho: \mathbb{R}^m \to [0,1]^{|\Gamma|}$ *such that* $\mathbf{p}_g = \varrho(\boldsymbol{\beta}_g)$ *and* $\mathbf{p}'_g = \varrho(\boldsymbol{\beta}'_g)$ *for each* g*, we also have* $\boldsymbol{\beta}_g = \boldsymbol{\beta}'_g$ *for each* g*.*

The next theorem extends Theorem 3 from Dasylva and Goussanou (2022), which is about the consistency of the maximum likelihood estimators. In order to state this extension, more notation is needed. For $G \ge 1$, consider the finite mixture of G distributions from $\mathcal{F}$, where the g-th component has probability α_g and parameters $\mathbf{p}_g$ and $\boldsymbol{\lambda}_g$. Also, denote by $\boldsymbol{\theta} = \left[(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g) \right]_{1 \le g \le G}$ the associated mixture parameters and by $q(.; \boldsymbol{\theta})$ the corresponding PMF.

$$\begin{aligned} q(\mathbf{n}; \boldsymbol{\theta}) \quad = \quad & \sum_{g=1}^{G} \alpha_g \left(I(|\mathbf{n}| = 0)\left(1 - |\mathbf{p}_g|\right) e^{-|\boldsymbol{\lambda}_g|} + I(|\mathbf{n}| > 1) \left(\left(1 - |\mathbf{p}_g|\right) \prod_{\gamma \in \Gamma} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)}}}{n^{(\gamma)}!} \right.\right. \\ & \left.\left. + \sum_{\gamma \in \Gamma : n^{(\gamma)} > 0} p_g^{(\gamma)} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)} - 1}}{\left(n^{(\gamma)} - 1\right)!} \prod_{\gamma' \in \Gamma \setminus \{\gamma\}} \frac{e^{-\lambda_g^{(\gamma')}} \left(\lambda_g^{(\gamma')}\right)^{n^{(\gamma')}}}{n^{(\gamma')}!} \right) \right), \quad \mathbf{n} = \left[n^{(\gamma)} \right]_{\gamma \in \Gamma} \in \mathbb{N}^{|\Gamma|}, \end{aligned} \tag{B.5}$$

Define

$$M_N(\boldsymbol{\theta}) = \frac{1}{|S_B|}\sum_{i\in S_B}\log q(\mathbf{n}_i;\boldsymbol{\theta}), \tag{B.6}$$

and for an integer $\tau_N > 0$, define

$$\begin{aligned} M_N(\boldsymbol{\theta};\tau_N) \;=\; & \frac{1}{|S_B|}\sum_{i\in S_B}\Big(I\big(|\mathbf{n}_i|\le\tau_N\big)\log q(\mathbf{n}_i;\boldsymbol{\theta}) \\ & + I\big(|\mathbf{n}_i|\ge\tau_N+1\big)\log\Big(\sum_{\mathbf{n}\in\mathbb{N}^{|\Gamma|}:|\mathbf{n}|\ge\tau_N+1} q(\mathbf{n};\boldsymbol{\theta})\Big)\Big). \end{aligned} \tag{B.7}$$

For $\boldsymbol{\theta}_\varrho = \big[\big(\alpha_g,\boldsymbol{\beta}_g,\boldsymbol{\lambda}_g\big)\big]_{1\le g\le G}$, let $\boldsymbol{\theta}\big(\boldsymbol{\theta}_\varrho\big) = \big[\big(\alpha_g,\varrho\big(\boldsymbol{\beta}_g\big),\boldsymbol{\lambda}_g\big)\big]_{1\le g\le G}$. Also for a positive integer d, let $\mathbf{0}_d$ denote the d-tuple with all zeros.

As before, the mixture parameters may be estimated by maximizing the log-likelihood of the $\mathbf{n}_i$'s, i.e., $M_N(.)$ or $M_N(;\tau_N)$. The following theorem states that the resulting estimators are consistent under suitable conditions, which include (B.3) or (B.4). In the latter case, it is assumed that the mapping ϱ is injective.

Theorem 2. *For $G^* \ge 2$ and $\nu \in (0,1)$, let $\Theta_1,\ldots,\Theta_{G^*}$ denote compact subsets of $\mathbb{R}^{(2|\Gamma|+1)G^*-1}$ such that*

$$\begin{aligned} \Theta_G \subset \Big\{ & \Big[\big(\alpha_g,\mathbf{p}_g,\boldsymbol{\lambda}_g\big)\Big]_{1\le g\le G^*} \in \mathbb{R}^{(2|\Gamma|+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g,\mathbf{p}_g,\boldsymbol{\lambda}_g\big)\in(\nu,1]\times[0,1]^{|\Gamma|}\times[\nu,\Lambda]^{|\Gamma|} \text{ and } \big|\mathbf{p}_g\big|\le 1-\nu \text{ and} \\ & \boldsymbol{\lambda}_{g+1}\succ\boldsymbol{\lambda}_g \text{ if } g\ge G^*-G+1, \\ & \big(\alpha_g,\mathbf{p}_g,\boldsymbol{\lambda}_g\big)=(0,\mathbf{0}_{|\Gamma|},\mathbf{0}_{|\Gamma|}) \text{ if } g\le G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\},\ G=1,\ldots,G^*. \end{aligned}$$

and let $\Theta = \bigcup_{G=1}^{G^*}\Theta_G$. For an injective mapping $\varrho:\mathbb{R}^m\to[0,1]^{|\Gamma|}$, also let $\Theta_{\varrho 1},\ldots,\Theta_{\varrho G^*}$ denote compact subsets of $\mathbb{R}^{(|\Gamma|+m+1)G^*-1}$ such that

$$\begin{aligned} \Theta_{\varrho G} \subset \Big\{ & \Big[\big(\alpha_g,\boldsymbol{\beta}_g,\boldsymbol{\lambda}_g\big)\Big]_{1\le g\le G^*} \in \mathbb{R}^{(|\Gamma|+m+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g,\boldsymbol{\beta}_g,\boldsymbol{\lambda}_g\big)\in(\nu,1]\times\mathbb{R}^m\times[\nu,\Lambda]^{|\Gamma|} \text{ and } \big|\varrho\big(\boldsymbol{\beta}_g\big)\big|\le 1-\nu \text{ and} \\ & \boldsymbol{\lambda}_{g+1}\succ\boldsymbol{\lambda}_g \text{ if } g\ge G^*-G+1, \\ & \big(\alpha_g,\boldsymbol{\beta}_g,\boldsymbol{\lambda}_g\big)=(0,\mathbf{0}_m,\mathbf{0}_{|\Gamma|}) \text{ if } g\le G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\}, G=1,\ldots,G^*, \end{aligned}$$

and let $\Theta_\varrho = \bigcup_{G=1}^{G^*}\Theta_{\varrho G}$. Suppose that all the linkage rules are simple (i.e., each rule is such that the decision to link two records involves no other record), C.1-C.3 (from Theorem 1) apply, and (B.3) also applies with

$$F(\mathbf{p},\boldsymbol{\lambda}) = \sum_{g=1}^{G^*}\alpha_{0g}I\Big((\mathbf{p},\boldsymbol{\lambda}) = \big(\mathbf{p}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big),$$

$\boldsymbol{\theta}_0 = \left[\left(\alpha_{0g}, \mathbf{p}_{0g}, \boldsymbol{\lambda}_{0g}\right)\right]_{1 \le g \le G^*} \in \Theta_{G_0}$ *and* $1 \le G_0 \le G^*$, *and let* $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ *and* $\hat{\boldsymbol{\theta}}_{3N}$ *denote the estimators, which respectively maximize* $M_N(.)$ *over* Θ_{G_0}, $M_N(.)$ *over* Θ, *and* $M_N(.;\tau_N)$ *over* Θ, *where* τ_N *is a positive integer such that* $\tau_N \to \infty$ *and* $\tau_N = O(\log N)$. *Then* $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ *and* $\hat{\boldsymbol{\theta}}_{3N}$ *converge in probability to* $\boldsymbol{\theta}_0$. *Furthermore, suppose that* $\mathbf{p}_N(v)$ *is also of the form* $\varrho(\boldsymbol{\beta}_N(v))$ *for* $\boldsymbol{\beta}_N : \mathcal{V}_N^* \to \mathbb{R}^m$, *which satisfies (B.4) with*

$$H(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{g=1}^{G^*} \alpha_{0g} I\left((\boldsymbol{\beta}, \boldsymbol{\lambda}) = \left(\boldsymbol{\beta}_{0g}, \boldsymbol{\lambda}_{0g}\right)\right),$$

$\varrho\left(\boldsymbol{\beta}_{0g}\right) = \mathbf{p}_{0g}$ *for* $g = G^* - G_0 + 1, \ldots, G^*$ *and* $\boldsymbol{\theta}_{\varrho 0} = \left[\left(\alpha_{0g}, \boldsymbol{\beta}_{0g}, \boldsymbol{\lambda}_{0g}\right)\right]_{1 \le g \le G^*} \in \Theta_{\varrho G_0}$. *Then,* $\boldsymbol{\theta}_{\varrho 0}$ *is also estimated consistently by maximizing* $M_N(.)$ *over* $\Theta_{\varrho G_0}$, $M_N(.)$ *over* Θ_ϱ, *or* $M_N(.;\tau_N)$ *over* Θ_ϱ.

In the above theorem, the mapping $\varrho(.)$ must be injective. The next lemma shows that this condition is met when $\varrho(.)$ is based on a nonsaturated log-linear specification of the interactions in the matched pairs.

Lemma 4. *For positive integers* $K, H_1, \ldots, H_K$, *let* $\Gamma = \{0, \ldots, H_1\} \times \ldots \times \{0, \ldots, H_K\} - \mathbf{0}_K$ *and* $\mathbf{p} = \left[p^{(\gamma)}\right]_{\gamma \in \Gamma}$ *be of the form* $p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}$, *where* $\sum_{\gamma \in \Gamma \cup \{\mathbf{0}_K\}} r^{(\gamma)} = 1$ *and* $r^{(\gamma)}$ *has the following log-linear form with no interactions of order greater than* $d < K$.

$$r^{(\gamma)} = \exp\left(u + \sum_{t=1}^{d} \sum_{1 \le k_1 < \ldots < k_t \le K} u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})}\right), \tag{B.8}$$

where the term $u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})}$ *is set to zero if one of* $\gamma_{k_1}, \ldots, \gamma_{k_t}$ *is null, according to the dummy coding convention. Let* $\boldsymbol{\beta}$ *denote the vector that comprises* $P(i \in S_A)$ *and the parameters of* $r^{(\gamma)}$, *which are not set to zero by this convention. Then the mapping* $\varrho : \boldsymbol{\beta} \mapsto \mathbf{p}$ *is injective.*

C. Initialization procedure

This section describes the initialization procedure for fitting the multivariate neighbor model in the simulations. The model parameters include the mixing proportions (the α_g's), the parameters of the false positives distribution (the $\boldsymbol{\lambda}_g$'s) and those for the true positives distribution (the common value of the $\mathbf{p}_g$'s). With G classes, the mixing proportion α_g is set to $1/G$ for each class. The other starting values are chosen as follows.

For the false positives distribution, each $\boldsymbol{\lambda}_g$ is set to a common value that is denoted by $\hat{\boldsymbol{\lambda}} = \left[\hat{\lambda}^{(\gamma)}\right]_{\gamma \in \Gamma}$, where $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$. For $\gamma \in \Gamma$, $\hat{\lambda}^{(\gamma)}$ is set to the estimate of $\bar{\lambda}$ that is obtained by fitting the univariate neighbor model, when the records are linked according to γ. For example, when $\gamma = (1,0,1)$, this means linking a pair if it is linked by the first linkage rule, and there is exact agreement on the surname and birth month but disagreement on the birth day.

Based on $\hat{\boldsymbol{\lambda}}$, the starting values are chosen for the coverage probability $P(i \in S_A)$ and the log-linear parameters (i.e., $u_{1(1)}$, $u_{2(1)}$, $u_{3(1)}$, $u_{12(11)}$, $u_{13(11)}$ and $u_{23(11)}$). These values are found in three steps as follows. In the first step, an estimate $\hat{\mathbf{p}} = \left[\hat{p}^{(\gamma)}\right]_{\gamma \in \Gamma}$ of the vector of true positives probabilities is computed by maximizing the log-likelihood of the multivariate model with a single class, where $\hat{\boldsymbol{\lambda}}$ is plugged in, and the true positives probabilities are not constrained to have a log-linear form. This step corresponds to a convex optimization, because the log-likelihood of the multivariate model is concave with respect to the true positives probabilities, when the other parameters are given. In the second step, the starting values for the log-linear parameters of $r^{(\gamma)}$ (in (5.2)) are found by a method of moments, based on $\hat{\mathbf{p}}$ as follows. When fitting the model without interactions, let $\hat{q}_k = \sum_{\gamma \in \Gamma: \gamma_k = 1} \hat{p}^{(\gamma)}$ (for $k = 1, 2, 3$) and $\hat{q}_{k_1 k_2} = \sum_{\gamma \in \Gamma: \gamma_{k_1} = \gamma_{k_2} = 1} \hat{p}^{(\gamma)}$ (for $1 \le k_1 < k_2 \le 3$) and choose the starting values as

$$\hat{u}_{1(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_2} + \frac{\hat{q}_{13}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{2(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{3(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{13}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_2}\right)\right),$$

with $\hat{u}_{12(11)} = \hat{u}_{13(11)} = \hat{u}_{23(11)} = 0$. When including the interactions, choose the starting values as

$$\begin{aligned}
\hat{u}_{1(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{2(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{3(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{12(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{13(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{23(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right).
\end{aligned}$$

Finally, let $\hat{\mathbf{r}} = \left[\hat{r}^{(\gamma)}\right]_{\gamma \in \{0,1\}^3}$ denote the vector of probabilities that correspond to the above starting values of the log-linear parameters (i.e., $\hat{r}^{(\gamma)}$ is equal to the right-hand side of (5.2), where the starting values are plugged in) and choose the starting coverage as

$$\hat{P}(i \in S_A) = \frac{\sum_{\gamma \in \Gamma} \hat{p}^{(\gamma)}}{\sum_{\gamma \in \Gamma} \hat{r}^{(\gamma)}}.$$

References

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

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

Brown, J., Bycroft, C., Di Cecco, D., Elleouet, J., Powell, G., Račinskij, V., Smith, P.A., Tam, S.-M., Tuoto, T. and Zhang, L.-C. (2020). Exploring developments in population size estimation. *Survey Statistician*, 82, 27-39.

Chambers, R. (2009). Regression analysis of probability-linked data. In *Research Series in Official Statistics*. Government of New Zealand.

Chipperfield, J.O., and Chambers, R.L. (2015). Using the bootstrap to analyse binary data obtained via probabilistic linkage. *Journal of Official Statistics*, 31, 397-414.

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

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

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

Dasylva, A., Abeysundera, M., Akpoué, B., Haddou, M. and Saïdi, A. (2016). Measuring the quality of a probabilistic linkage through clerical reviews. Proceedings: *Symposium 2016, Growth in Statistical Information: Challenges and Benefits*, Statistics Canada.

Dasylva, A., and Goussanou, A. (2020). Estimating linkage errors under regularity conditions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687-692.

Dasylva, A., and Goussanou, A. (2021). Estimating the false negatives due to blocking in record linkage. *Survey Methodology*, 47, 2, 299-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00002-eng.pdf.

Dasylva, A., and Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*, 5, 181-216. DOI: https ://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., and Goussanou, A. (2024). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*, 7, 17-56. DOI: https ://doi.org/10.1007/s42081-023-00228-9.

Dasylva, A., Goussanou, A. and Nambeu, C.O. (2024). *Models of Linkage Error for Capture-Recapture Estimation without Clerical Reviews*. https://arxiv.org/pdf/2403.11438.pdf.

de Wolf, P.-P., van der Laan, J. and Zult, D. (2019). Connection correction methods for linkage error in capture-recapture. *Journal of Official Statitsics*, 35, 577-597.

Di Consiglio, L., and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31, 415-429.

Ding, Y., and Fienberg, S. (1994). Dual system estimation of census undercount in the presence of matching error. *Journal of the American Statistical Association*, 20, 149-158.

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

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001). On bayesian record linkage. *Research in Official Statistics*, 4, 185-198.

Haque, S., and Mengersen, K. (2022). Improved assessment of the accuracy of record linkage via an extended macsim approach. *Journal of Official Statistics*, 38, 429-451.

Haque, S., Mengersen, K. and Stern, S. (2021). Assessing the accuracy of record linkages with markov chain based monte carlo simulation approach. *Journal of Big Data*, 8, 1-26.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Lahiri, P., and Larsen, D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-227.

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

Lincoln, F. (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular*, 118, 1-4.

Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

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

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6, 5-84.

Račinskij, V., Smith, P.A. and van der Heijden, P. (2019). *Linkage Free Dual System Estimation*. https://arxiv.org/abs/1903.10894, 2019.

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

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

Steorts, R., Hall, R. and Fienberg, S.E. (2016). A bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660-1672.

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

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

US Census Bureau (2016). *File b: Surnames Occurring 100 or More Times*. https://www2.census.gov/topics/genealogy/2010surnames/names.zip. (Accessed: 2020-10-17).

US Census Bureau (2020). *Annual State Resident Population Estimates for 6 Race Groups (5 Race Alone Groups and Two or More Races) by Age, Sex, and Hispanic Origin: April 1, 2010 to July 1, 2019*. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv. (Accessed: 2020-10-17).

Winglee, M., Valliant, R. and Scheuren, F. (2005). A case study in record linkage. *Survey Methodology*, 31, 1, 3-11. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005001/article/8085-eng.pdf.

Winkler, W.E. (1993). Improved decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274-279.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 409-433
Statistics Canada, Catalogue No. 12-001-X

Investigating mode effects in interviewer variances using two representative multi-mode surveys

Wenshan Yu, Michael R. Elliott and Trivellore E. Raghunathan[1]

Abstract

As mixed-mode designs become increasingly popular, their effects on data quality have attracted much scholarly attention. Most studies focused on the bias properties of mixed-mode designs; few of them have investigated whether mixed-mode designs have heterogeneous variance structures across modes. While many characteristics of mixed-mode designs, such as varied interviewer usage, systematic differences in respondents, varying levels of social desirability bias, among others, may lead to heterogeneous variances in mode-specific point estimates of population means, this study specifically investigates whether interviewer variances remain consistent across different modes in mixed-mode studies. To address this research question, we utilize data collected from two distinct study designs. In the first design, when interviewers are responsible for either face-to-face or telephone mode, we examine whether there are mode differences in interviewer variances for 1) sensitive political questions, 2) international items, 3) and item missing indicators on international items, using the Arab Barometer wave 6 Jordan data. In the second design, we draw on Health and Retirement Study (HRS) 2016 core survey data to examine the question on three topics when interviewers are responsible for both modes. The topics cover 1) the CESD depression scale, 2) interviewer observations, and 3) the physical activity scale. To account for the lack of interpenetrated designs in both data sources, we include respondent-level covariates in our models. We find significant differences in interviewer variances on one item (twelve items in total) in the Arab Barometer study; whereas for HRS, the results are three out of eighteen. Overall, we find the magnitude of the interviewer variances larger in FTF than TEL on sensitive items. We conduct simulations to understand the power to detect mode effects in the typically modest interviewer sample sizes.

Key Words: Interviewer effects; Mixed-mode design; Mode effects; Multimode study.

1. Introduction

Interviewers play a central role in survey data collection. Depending on the mode and sampling design of data collection, they may need to list addresses to generate sampling frames, recruit respondents, ask survey questions, and record participants' responses. Therefore, from a total survey error framework, interviewers can affect survey data quality by generating or reducing coverage error, nonresponse error, measurement error, and processing error (West and Blom, 2017). Most research examining interviewers' effects focuses on measurement error (Schuman and Converse, 1971; Hanson and Marks, 1958; Ehrlich and Riesman, 1961), which can be further decomposed into a systematic part, the bias due to interviewers (when respondents alter answers either because of the presence of interviewers or their observable traits), and a random component, interviewer variance. This interviewer variance inflates the uncertainty of the estimates, sometimes to an even greater degree than the correlation induced by geographical clustering (Schnell and Kreuter, 2003). This study focuses on determining the effect of different modes of data collection – specifically telephone (TEL) versus face-to-face (FTF) – on interviewer variances in mixed-mode surveys.

1. Wenshan Yu, Survey Research Center at the Institute for Social Research, The University of Michigan. E-mail: yuwens@umich.edu; Dr. Michael R. Elliott, Survey Research Center at the Institute for Social Research and Biostatistics Department, The University of Michigan. E-mail: mrelliot@umich.edu; Dr. Trivellore E. Raghunathan, Survey Research Center at the Institute for Social Research and Biostatistics Department, The University of Michigan. E-mail: teraghu@umich.edu.

Interviewer variances were first studied in the context of face-to-face interviews (Kish, 1962). When telephone surveys became an alternative to FTF interviews, researchers evaluated interviewer variances in telephone surveys and generally found that they were less substantial than those in personal surveys (Groves and Magilavy, 1986; Tucker, 1983; Groves and Kahn, 1979). Specifically, the intraclass correlation ρ_{int}, a common measure used to assess interviewer effects and defined by the ratio of interviewer variances to the total variance, ranged from 0.005 to 0.102 in FTF surveys, whereas those computed in centralized TEL surveys ranged from 0.0018 to 0.0184 (Groves and Magilavy, 1986; Groves and Kahn, 1979). The finding is aligned with theoretical expectations, as interviewers in the centralized TEL setting are more closely monitored and supervised than field interviewers are (Schaeffer, Dykema and Maynard, 2010). Since then, the research domain has received little scholarly attention. However, as mixed-mode designs become increasingly used, the subject of study calls for more research. There is a lack of first-hand evidence as the prior findings are mostly based on different surveys that employ one mode (FTF or TEL). Besides, mixed-mode surveys naturally provide an opportunity where the survey context and the questionnaires are highly comparable (if not the same) when comparing interviewer variances in both modes. Furthermore, depending on whether interviewers are responsible for both modes in mixed-mode surveys, interviewers can potentially carry their influence from one mode to another. These factors can lead to different results in comparing interviewer variances between modes.

Investigating mode effects in interviewer variances is also useful to facilitate mixed-mode designs and serve as an indicator of data quality. First, quantifying mode-specific interviewer variance can help researchers to determine and choose the mode with low interviewer variance in a multimode design. The current state-of-the-art mixed-mode inference strategy focuses on the bias property of modes (Elliott, Zaslavsky, Goldstein, Lehrman, Hambarsoomians, Beckett and Giordano, 2009; Kolenikov and Kennedy, 2014), but little was done to incorporate the potential heterogeneous variance structure (Suzer-Gurtekin, Heeringa and Valliant, 2013). Part of the reason is that little literature sheds light on the variance properties of mixed-mode designs (Vannieuwenhuyze, 2015), especially what goes into the variances. Second, identifying the questions associated with large interviewer variance mode effects can inform how interviewer variance is generated and thus might be reduced. For example, researchers show that attitudinal, sensitive, ambiguous, complex, and open-ended questions are generally more vulnerable to interviewer effects (Schaeffer, Dykema and Maynard, 2010), as those questions introduce more opportunities for the interviewer to help the respondents (West and Blom, 2017). If sensitive questions only present a large interviewer effect in FTF but not in TEL, that may suggest the questions bring a burden to field interviewers. To address that, survey organizations can provide additional training to standardize how to ask the question or use other approaches [such as audio computer-assisted self-interviewing [ACASI] or the item count technique (Holbrook and Krosnick, 2010)] to collect information for sensitive items. Third, in mixed-mode designs where interviewers are responsible for both modes, we can potentially find specific interviewers that have a large effect on responses in both modes or only in one mode, which provide the basis for real-time intervention and interviewer training at a more granular level.

In this paper, we consider two representative multi-mode studies: 1) the Arab Barometer Study (ABS) Wave 6 Jordan experiment and 2) the Health and Retirement Study (HRS) 2016. Drawing on both data sources, we consider mode effects in interviewer variances for interviewers in different countries, for different target populations, and for a variety of outcome variables. Additionally, the use of the two studies offers distinct perspectives for examining our research question. The ABS interviewer design is commonly used in surveys where different modes are managed by separate data collection agencies, resulting in different interviewers across modes. On the other hand, the HRS interviewer design, where the same interviewers are utilized in both modes, facilitates a more precise estimation of the differences in interviewer variances solely due to modes, by eliminating the portion of interviewer variances that result from using different interviewers across modes.

The remainder of this paper is organized as follows. In Section 2, we describe the study design and analytical strategy, and present the results using our first data source – ABS. Section 3 introduces the second data source – HRS, along with the corresponding analytical approach and the results pertaining to interviewer variance associated with the HRS data. In Section 4, we conduct a simulation study to illustrate the power to detect mode effects in interviewer variances using both the ABS and the HRS setup. Finally, in Section 5, we discuss the implications of our study.

2. The Arab Barometer Study

2.1 Study description

The ABS is the largest repository of public opinion data in the Middle East and North Africa (MENA) region. In wave 6, it embedded a mode experiment in Jordan between March and April 2021, where participants were randomly assigned to either a personal interview or a TEL recontact interview. Center for Strategic Studies in Jordan conducted the field work using the 2015 Population and Housing Census as the sampling frame. They implemented an area probability sample stratified on governorate and urban-rural cleavages. Separate interviewers were used in the FTF and TEL interviews. The TEL-assigned households were initially recruited via FTF for a short 5-minute survey, and the majority of the survey items were asked approximately a week later in a telephone follow-up. In the FTF mode, 31 interviewers collected data from 1,193 respondents, while 13 interviewers interviewed 1,212 participants via phone.

We focus on three types of outcome variables (Y): 1) sensitive political questions (6 items), 2) less sensitive international questions (3 items), and 3) whether reported do not know or refused to answer international relationship questions (3 items). Except for the item missing indicators, the other outcome variables were initially measured by four ordinal categories; we collapsed them into binary outcomes by setting the cutoff point in the middle. The original and collapsed categories are available in Appendix A of (Yu, Elliott and Raghunathan, 2024).

Outcome variables (Y) can be subject to two types of mode effects: 1) mode effects that lead to a shift in the means of outcome variables (referred to as mode effects in means) and 2) mode effects in interviewer

variances. We consider, in total, q interviewers collect information in only one of two modes (FTF and TEL) from n sample units from a finite population. Interviewers also collect respondent-level covariates (X) that are predictive of the outcome variables (Y). The covariates (X) are assumed to be independent of any mode effects. We consider covariates (X) including respondents' age, gender, marital status, household size, and regions in this paper.

2.2 Analytical strategy

First, to illustrate the descriptive statistics of interviewer variation in the collected responses, we compute the between-interviewer standard deviation (SD) and the average within-interviewer SD. Specifically, we calculate the average proportions for each variable and interviewer $(\bar{y}_{(m)j})$. In the ABS setup, where interviewers are nested within each mode, these statistics are inherently mode-specific; therefore, we enclose m in parentheses to emphasize this point. We then calculate the SD of these average proportions across interviewers, termed the between-interviewer SD. The within-interviewer SD (v_j^m) is derived from the responses collected by each interviewer. The average within-interviewer SD is computed as the mean of the within-interviewer SDs across all interviewers for each mode. We show the formula to compute the relevant statistics in (2.1), where i indexes respondents, j indexes interviewers, m indexes modes, $n_{(m)j}$ reflects the number of interviews conducted by interviewer j using mode m, n_m represents the number of respondents in mode m, n_j^m indicates the number of interviewers using mode m, and $y_{i(m)j}$ indicates the responses provided by respondent i interviewed by interviewer j using mode m. For survey data collection agencies, a small SD between interviewers and a large average within-interviewer SD are desirable, as this may indicate an interviewer assignment that is close to random and minimal effects from interviewers on the collected responses. We report the statistics for both the covariates and the outcomes of interest. The statistics for the covariates can suggest interviewer selection effects, thereby highlighting the importance of considering the covariates in the final analytical model. The statistics for the outcome variables may provide initial evidence of the presence of interviewer effects and justify further investigation.

$$
\begin{aligned}
\text{Average proportion per interviewer } \bar{y}_{(m)j} &= \frac{\sum_i^{n_{(m)j}} y_{i(m)j}}{n_{(m)j}} \\
\text{Average proportion per mode } \bar{y}_m &= \frac{\sum_i^{n_m} y_{i(m)j}}{n_m} \\
\text{Between-interviewer SD} &= \sqrt{\frac{\sum_j^{n_j^m} (\bar{y}_{(m)j} - \bar{y}_m)^2}{n_j^m}} \\
\text{Within-interviewer SD } v_j^m &= \sqrt{\frac{\sum_i^{n_{(m)j}} (\bar{y}_{i(m)j} - \bar{y}_{(m)j})^2}{n_{(m)j}}}
\end{aligned}
\tag{2.1}
$$

$$\text{Average within-interviewer SD} = \frac{\sum_j^{n_j^m} v_j^m}{n_j^m}.$$

To test whether interviewer variances are equal across modes, since all the outcome variables are binary, we fit the following probit model to each of the variables, where m indexes modes (f for FTF and t for TEL), M and $J_{j,\, j=1,\ldots,q-1}$ are dummy variables (length of n) to indicate modes ($M=1$ for the FTF mode and $M=0$ for the TEL mode) and interviewers:

$$\begin{aligned}
Y^*_{ij(m)} &= \beta_0 + \beta_1 M_i + b_{j(m)} + \epsilon_{ij(m)}, \\
Y_{ij(m)} &= 1 \text{ if } Y^*_{ij(m)} > 0 \text{ and } Y_{ij(m)} = 0 \text{ if } Y^*_{ij(m)} \le 0, \\
b_{j(m)} &\sim N(0, \sigma_m^2), \\
\epsilon_{ij(m)} &\sim N(0,1), \\
\sigma_f, \sigma_t &\sim \text{ half} - T(3,1) \text{ (for Bayesian modeling)}, \\
\boldsymbol{\gamma}, \beta_0, \beta_1 &\sim N(0, 10^6) \text{ (for Bayesian modeling)}.
\end{aligned} \tag{2.2}$$

In model (2.2), the interviewer random effects are represented as $b_{j(m)}$ as interviewers are nested within the modes. Our research question, "Are interviewer variances equal between modes in a randomized mixed-mode design?" is addressed by evaluating if $\alpha = \log(\sigma_f) - \log(\sigma_t)$ is equal to zero for each variable in model (2.2). To determine this, we examine if the 95% confidence or HPD credible intervals of α include zero. If the intervals do not include zero for some variables, it suggests that the interviewer variances are not equal between modes for those variables.

By fitting (2.2), we can also obtain estimates of mode effects (β_1) for each variable by computing and testing if the quantity differs from 0. Note that the estimates may include some mode selection effects; despite the random mode assignment, differential nonresponse can happen across the modes (West, Kreuter and Jaenichen, 2013).

Suppose evidence suggests that $\alpha \neq 0$, we then consider whether the mode-specific interviewer variance is spurious due to the lack of interpenetrated designs by adding respondent-level covariates (x_{si}, where s denotes covariate s) to model (2.2):

$$\begin{aligned}
Y^*_{ij(m)} &= \beta_0 + \beta_1 M_i + b_{j(m)} + \sum_s^S \gamma_s x_{si} + \epsilon_{ij(m)}, \\
Y_{ij(m)} &= 1 \text{ if } Y^*_{ij(m)} > 0 \text{ and } Y_{ij(m)} = 0 \text{ if } Y^*_{ij(m)} \le 0, \\
b_{j(m)} &\sim N(0, \sigma_m^2), \\
\epsilon_{ij(m)} &\sim N(0,1), \\
\sigma_f, \sigma_t &\sim \text{ half} - T(3,1) \text{ (for Bayesian modeling)}, \\
\boldsymbol{\gamma}, \beta_0, \beta_1 &\sim N(0, 10^6) \text{ (for Bayesian modeling)}.
\end{aligned} \tag{2.3}$$

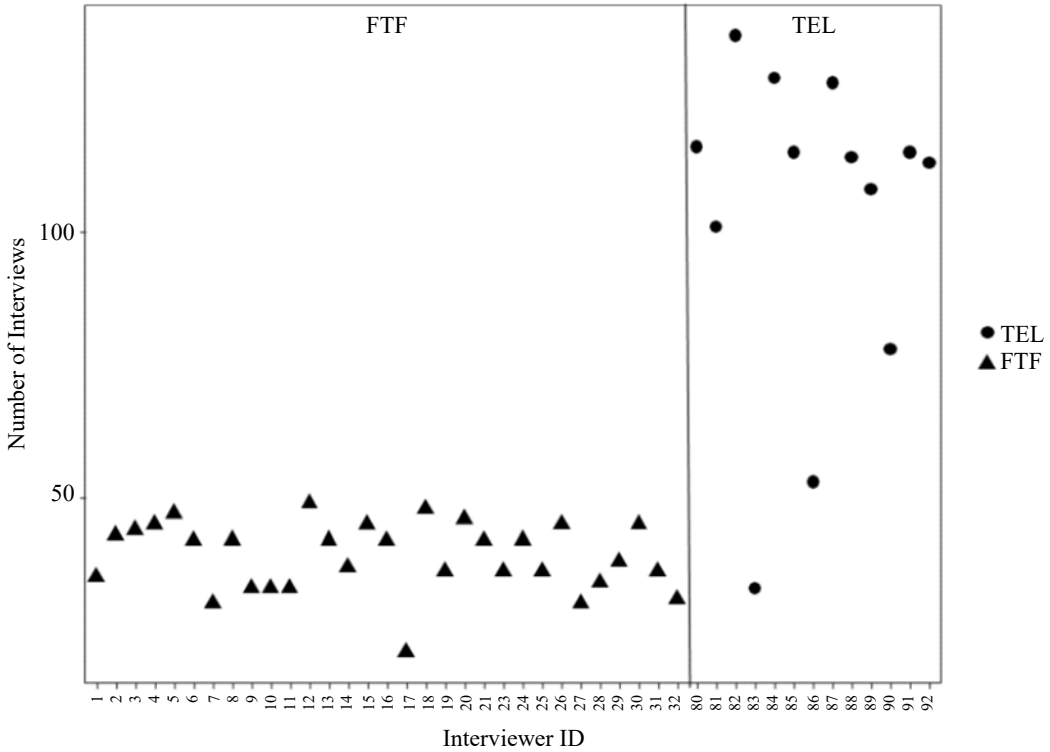
We implement the models using both likelihood (Proc Nlmixed) and Bayesian approaches (Proc MCMC) in the SAS programming language. In the likelihood approach, we take log transformation on σ_f^2 and σ_t^2 to stabilize the variance of the parameters and improve the coverage property. We compute the variance of the estimated α using the delta method, given by $\text{var}(\alpha) = \frac{1}{4}\text{var}(\log(\sigma_f^2)) + \frac{1}{4}\text{var}(\log(\sigma_t^2))$ (see the derivations in the Appendix A), then use a normal distribution to estimate the 95% confidence interval. In the Bayesian approach, we use one chain with 200,000-300,000 draws, depending on the autocorrelation and effective sample size, and select every 100th value as the thinning rate. For the ease of illustration, we only report the results of the model with covariates added and estimated using Bayesian modeling model (2.3) in the later section.

2.3 Results

2.3.1 Descriptive statistics

We assume interviewers are interchangeable in this paper. To partly evaluate this assumption, we present the interviewer workloads in the FTF and TEL modes in the ABS in Figure 2.1. In Figure 2.1, we note that in the FTF mode, each interviewer conducts a similar number of interviews. In contrast, both the mean and the variation in the number of interviews per interviewer are larger in the TEL mode.

Figure 2.1 Interviewer workloads per mode in the Arab Barometer Study.



We report mode-specific sample means, between-interviewer SDs, and average within-interviewer SDs in Table 2.1. From Table 2.1. First, we observe that for sensitive political questions, the average proportions reported via TEL are generally higher than those reported in FTF interviews, suggesting that TEL may be associated with more positive reporting. Second, between-interviewer SDs in FTF are generally larger than those in TEL for most outcomes, while the average within-interviewer SDs are larger in TEL than in FTF for sensitive political questions and missing indicators. This provides some initial evidence that interviewers seem to have a larger effect in FTF than in TEL. We provide the distribution of the outcome variables per interviewer in Appendix C of (Yu, Elliott and Raghunathan, 2024).

Table 2.1
Distribution of outcome variables in the Arab Barometer Study across interviewers by modes.

Questions	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average within interviewer SD (FTF)	Average within interviewer SD (TEL)
Sensitive political questions						
1. Freedom of the media	0.403	0.588	0.191	0.117	0.455	0.480
2. trust in government	0.356	0.533	0.165	0.122	0.455	0.487
3. trust in courts	0.594	0.770	0.139	0.123	0.477	0.398
4. satisfied with healthcare	0.491	0.592	0.155	0.071	0.482	0.489
5. performance on inflation	0.140	0.243	0.146	0.142	0.291	0.406
6. performance during COVID-19	0.402	0.576	0.171	0.161	0.464	0.470
International Questions						
7. favorable of the United States	0.394	0.415	0.187	0.189	0.467	0.459
8. favorable of Germany	0.488	0.560	0.224	0.186	0.464	0.464
9. favorable of China	0.468	0.507	0.207	0.203	0.470	0.463
Whether missing on international questions (constructed)						
10. missing on favorable of the United States	0.253	0.297	0.235	0.158	0.341	0.425
11. missing on favorable of Germany	0.320	0.381	0.247	0.199	0.384	0.442
12. missing on favorable of China	0.283	0.329	0.252	0.180	0.359	0.431

Note: FTF = face-to-face; TEL = telephone; SD = standard deviation.

We show unweighted sample characteristics in the FTF and the TEL modes in Table 2.2. Under the randomized mixed-mode design, the Jordan sample is roughly balanced on key demographic and socioeconomic variables (age, gender, education, marital status, household size, and region) across modes. However, there are slightly more males (0.55 vs 0.50) respondents in the TEL mode relative to the FTF mode, possibly due to differential nonresponse. We note that for these covariates, the between-interviewer SD in FTF is usually much larger than that in TEL, suggesting potentially larger selection effects in FTF, since we assume the covariates are not susceptible to measurement error.

Table 2.2
Distribution of sample characteristics of the Arab Barometer Study across interviewers by modes.

Respondent Variables	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average within interviewer SD (FTF)	Average within interviewer SD (TEL)
Age 18-24	0.166	0.164	0.085	0.039	0.361	0.369
Age 25-34	0.226	0.203	0.072	0.038	0.415	0.402
Age 35-44	0.227	0.215	0.088	0.052	0.412	0.408
Age 45-54	0.199	0.219	0.069	0.031	0.394	0.414
Age 55+	0.183	0.198	0.071	0.032	0.381	0.399
Male	0.497	0.549	0.291	0.041	0.369	0.499
Less than secondary education	0.345	0.337	0.125	0.106	0.463	0.463
Secondary education	0.365	0.357	0.098	0.082	0.477	0.474
Higher than secondary education	0.290	0.307	0.101	0.051	0.445	0.461
Unmarried	0.238	0.264	0.106	0.063	0.412	0.438
Married	0.693	0.684	0.082	0.062	0.459	0.463
Divorced, widows, separated	0.069	0.053	0.044	0.024	0.230	0.219
Household size: Less than 3	0.208	0.222	0.083	0.040	0.399	0.416
Household size: 4-5	0.345	0.349	0.081	0.061	0.475	0.475
Household size: 6-7	0.281	0.288	0.079	0.072	0.447	0.449
Household size: 8+	0.165	0.141	0.089	0.056	0.353	0.330
Region: Central	0.523	0.509	0.154	0.255	0.482	0.429
Region: North	0.261	0.282	0.101	0.188	0.424	0.388
Region: South	0.216	0.209	0.175	0.119	0.333	0.367

Note: FTF = face-to-face; TEL = telephone; SD = standard deviation.

2.3.2 Mode effects in means and interviewer variances

This section reports the modeling results that incorporate respondent information model (2.3) using Bayesian estimation in Table 2.3. With respect to the mode effects in means, we observe negative estimates for all sensitive items. For example, the probability of an unmarried male participant aged 18-24, with higher than secondary education, living in a household with fewer than three individuals, and residing in the North region of Jordan, reporting that media freedom is guaranteed to a great or medium extent, decreases by 17.9% when interviewed via FTF compared to TEL interviews. The 17.9% is calculated using $\phi(\beta_0 + \beta_1 + \sum_s^S \gamma_s x_{si})\,\beta_1$, where ϕ is the pdf of a standard normal distribution and S is the number of covariates (x). The estimates of γ_s are not provided in the paper but can be provided upon request. The negative mode effects in means suggest that respondents expressed lower opinions of the government when answering FTF interviews, which could be more honest responses given Jordan's authoritarian regime. Table 2.3 also indicates that missing rates for international questions are lower in FTF interviews compared to TEL interviews (though this is not statistically significant at the 0.05 level). We did not incorporate sample weights in the analysis as our focus of inference is repeated sampling under the same survey design.

Next, we turn our attention to the interviewer variances. Firstly, the magnitude of interviewer variances is generally large in the ABS. For sensitive political questions, the interviewer variances range from 0.03 to 0.393 (Table 2.3). Previous literature examining interviewer effects usually reported interviewer intraclass correlation (ρ_{int}) to reflect the proportion of variance due to interviewers. To compute mode-specific $\rho_{m,\text{int}}$, we can use the formula $\rho_{m,\text{int}} = \frac{\text{var}_{m,\text{int}}}{1+\text{var}_{m,\text{int}}}$, since the residual variance in the probit model is 1. Consequently, the previously mentioned results correspond to ρ_{int} ranging from 0.029 to 0.282. As a reference, based on the literature, a value of ρ_{int} below 0.01 is considered small, while a value higher than 0.12 is regarded as large (West and Olson, 2010). In Table 2.3, we observe that $\rho_{f,\text{int}}$ and $\rho_{t,\text{int}}$ can vary substantially for the

same outcome. For example, for satisfaction with healthcare, $\rho_{f,\text{int}}$ is 0.125, while $\rho_{t,\text{int}}$ is 0.029. It is important to consider these differences when using the $\rho_{m,\text{int}}$ values to calculate the effective sample sizes associated with a specific data collection mode.

For one sensitive item, performance in the healthcare system, we observe marginally significant difference in interviewer variances in Table 2.3 using Bayesian estimation. The results are significant when using likelihood estimation, as shown in Appendix D of (Yu, Elliott and Raghunathan, 2024). In this item, the estimates of interviewer variances are considerably larger in the FTF mode. For 5 out of 6 sensitive items, FTF interviewer variances are somewhat larger than TEL interviewer variances. The differences are not statistically significant, possibly due to the limited power determined by the small number of interviewers in this study. The larger interviewer variances in FTF are consistent with theoretical expectations, as interviewers may exhibit greater heterogeneity in administering sensitive questions and establishing rapport with respondents during in-person interviews.

Table 2.3
Interviewer variances per mode for selected items in the Arab Barometer Study adjusting for covariates using Bayesian estimation.

Questions	σ_f^2	σ_t^2	$\rho_{f,\text{int}}$	$\rho_{t,\text{int}}$	α	β_1
Sensitive political questions						
1. Freedom of the media	0.252 [0.122, 0.428]	0.135 [0.036, 0.284]	0.201 [0.109, 0.3]	0.119 [0.035, 0.221]	0.355 [-0.223, 0.898]	**-0.526 [-0.795, -0.222]**
2. trust in government	0.188 [0.083, 0.322]	0.127 [0.029, 0.275]	0.158 [0.077, 0.244]	0.113 [0.028, 0.216]	0.239 [-0.382, 0.838]	**-0.504 [-0.768, -0.238]**
3. trust in courts	0.113 [0.038, 0.201]	0.214 [0.05, 0.445]	0.102 [0.037, 0.167]	0.176 [0.048, 0.308]	-0.29 [-0.94, 0.318]	**-0.555 [-0.881, -0.273]**
4. satisfied with healthcare	0.143 [0.051, 0.251]	0.03 [0, 0.075]	0.125 [0.049, 0.201]	0.029 [0, 0.07]	0.906 [-0.054, 1.758]	**-0.278 [-0.475, -0.085]**
5. performance on inflation	0.393 [0.153, 0.672]	0.204 [0.051, 0.435]	0.282 [0.133, 0.402]	0.169 [0.049, 0.303]	0.361 [-0.275, 0.927]	**-0.523 [-0.861, -0.153]**
6. performance during COVID-19	0.202 [0.084, 0.34]	0.224 [0.07, 0.443]	0.168 [0.077, 0.254]	0.183 [0.065, 0.307]	-0.026 [-0.602, 0.508]	**-0.51 [-0.841, -0.205]**
International Questions						
7. favorable of the United States	0.198 [0.074, 0.34]	0.362 [0.104, 0.719]	0.165 [0.069, 0.254]	0.266 [0.094, 0.418]	-0.278 [-0.841, 0.282]	-0.057 [-0.45, 0.318]
8. favorable of Germany	0.292 [0.12, 0.514]	0.33 [0.092, 0.663]	0.226 [0.107, 0.339]	0.248 [0.084, 0.399]	-0.037 [-0.603, 0.548]	-0.147 [-0.551, 0.236]
9. favorable of China	0.205 [0.083, 0.361]	0.378 [0.116, 0.787]	0.17 [0.077, 0.265]	0.274 [0.104, 0.44]	-0.282 [-0.869, 0.245]	-0.15 [-0.549, 0.19]
Whether missing on international questions (constructed)						
10. missing on favorable of the United States	0.995 [0.48, 1.71]	0.343 [0.104, 0.668]	0.499 [0.324, 0.631]	0.255 [0.094, 0.4]	**0.557 [0.014, 1.121]**	-0.298 [-0.805, 0.172]
11. missing on favorable of Germany	0.844 [0.404, 1.324]	0.464 [0.16, 0.857]	0.458 [0.288, 0.57]	0.317 [0.138, 0.461]	0.324 [-0.169, 0.839]	-0.287 (0.24) [-0.765, 0.149]
12. missing on favorable of China	0.936 [0.434, 1.552]	0.452 [0.118, 0.933]	0.483 [0.303, 0.608]	0.311 [0.106, 0.483]	0.398 [-0.134, 0.949]	-0.244 [-0.73, 0.229]

Notes: Significant results are marked in bold. β_1 refers to the mode effect estimates in means. α_1 refers to the mode effect estimates in interviewer variances. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,\text{int}}$ and $\rho_{t,\text{int}}$ are interviewer intraclass correlation in FTF and TEL, respectively.

Counterintuitively, for substantive responses to nonsensitive international attitude questions (items 7-9), the interviewer variance estimates are generally larger in TEL compared to FTF (not significantly). The interviewer variances of whether reporting don't know or refusing to answer the nonsensitive international questions are larger in FTF than in TEL (significant on the first item). This finding may be because interviewers assigned to FTF mode tried to persuade respondents to give substantive answers, and whether the persuasion happens or is successful can differ by interviewers.

3. Health and retirement study 2016

3.1 Study description

The HRS is a longitudinal panel study that surveys people over age 50 (and their spouses) in the United States. It is conducted biennially, started in 1992, and has studied more than 43,000 people (Fisher and Ryan, 2018). The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The HRS sample was drawn using a multistage, national area-clustered probability sample frame (Heeringa and Connor, 1995). Since 2006, The HRS has initiated the rotation of enhanced FTF and TEL across waves at the household level, except when the household includes participants aged 80 or older, who alternate between regular and enhanced FTF, or newly recruited participants, who are assigned to enhanced FTF in their first wave. In this study, we are interested in analyzing the HRS 2016 data, when the Late Baby Boomers (LBB) cohort was added to replenish the HRS sample. Although not every interviewer collects data in both modes, under the HRS design, interviewers are responsible for data collection in both FTF and TEL modes. The HRS 2016 was fielded from April 2016 to April 2018, with a sample size of 20,912 [response rate: 82.8%, (HRS, 2023)]. In our analytical sample, we excluded respondents who were missing data on mode indicators, interviewer IDs, and covariates, resulting in a sample size of 20,868.

We consider four types of outcome variables in the HRS study, including 1) nine items of the Center for Epidemiologic Studies Depression Scale (CESD), 2) six items of interviewer observations, and 3) a three-item physical activity scale. The question wordings, the original response categories and categories used in the study can be found in Appendix E of (Yu, Elliott and Raghunathan, 2024). We consider eight respondent-level covariates (X), including age, sex, race/ethnicity, interview language, education, whether respondents are coupled and working. All participants are included in our sample, unless they are missing data in either the outcome or predictor variables. Missing rates for predictor variables are minor, and those for outcome variables are less than 0.05.

3.2 Analytical strategy

Similar to the descriptive statistics reported in the ABS, we report the between-interviewer SD and the average within-interviewer SD to gain an intuitive understanding of the interviewer effects in the HRS.

Next, we fit multilevel models to each of the outcome variables using the same notation as in model (2.2). Unlike the ABS, interviewers are not nested in model hence a single interviewer can interview in both modes, and thus interviewer effects can be correlated across modes. Therefore we posit a bivariate normal model for the interviewer effects:

$$
\begin{aligned}
Y^*_{ijm} &= \beta_0 + \beta_1 M_i + b_{jm} + \sum_s^S \gamma_s x_{si} + \epsilon_{ijm}, \\
-\ Y_{ijm} &= 1 \text{ if } Y^*_{ijm} > 0 \text{ and } Y_{ijm} = 0 \text{ if } Y^*_{ijm} \le 0, \\
\begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix} \right), \\
\epsilon_{ijm} &\sim N(0,1), \\
\sigma_f, \sigma_t &\sim \text{ half} - T(3,1) \text{ (for Bayesian modeling)}, \\
\rho &\sim U(-1,1) \text{ (for Bayesian modeling)}, \\
\boldsymbol{\gamma}, \beta_0, \beta_1 &\sim N(0, 10^6) \text{ (for Bayesian modeling)}.
\end{aligned}
\tag{3.1}
$$

Similarly, we use $\alpha = \log(\sigma_f) - \log(\sigma_t)$ as a metric to answer our research question. To test if α is equal to zero for each variable, we assess if the 95% credible intervals or confidence intervals include zero. Additionally, to control for interviewer selection effects, we include respondent-level covariates as fixed effects in the model.

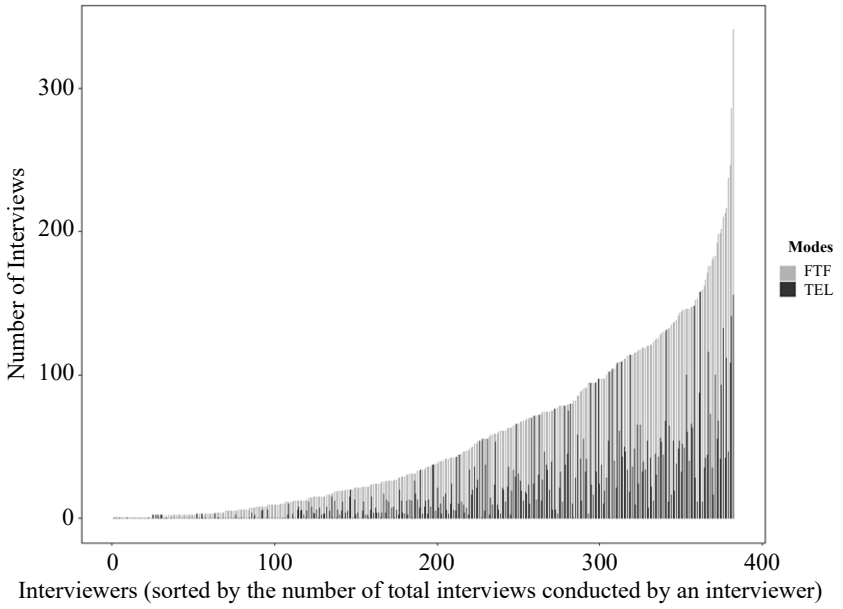
We apply the Fisher Z transformation $\left(z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right)$ when constructing the 95% confidence interval for ρ in the likelihood approach. We calculate the variance of α using the delta method, given by $\text{var}(\alpha) = \frac{1}{4}\text{var}(\log(\sigma_f^2)) + \frac{1}{4}\text{var}(\log(\sigma_t^2)) - \frac{1}{2}\text{cov}(\log(\sigma_f^2), \log(\sigma_t^2))$, which is slightly different from the ABS (see the derivations in Appendix A).

3.3 Results

3.3.1 Descriptive statistics

First, we illustrate the interviewer load in Figure 3.1. In HRS 2016, 382 interviewers were employed for data collection. The number of interviews conducted in FTF and TEL is very different across interviewers. Eighty-two (21.5%) interviewers exclusively conducted telephone interviews, while thirty-seven (9.7%) solely conducted in-person interviews. The remaining 263 (68.9%) interviewers conducted both types of interviews. All interviews are included in the analysis, although the estimation of the covariances between the FTF and TEL effects within interviewer is limited to the subsample of interviewers who conducted both types of interviews.

Figure 3.1 Interviewer workloads per mode in the Health and Retirement Study.



Second, we present unweighted sample characteristics for both FTF and TEL modes in Table 3.1. Compared to TEL respondents, a higher proportion of FTF respondents were under 60 or over 80 years old, belonged to minority groups, were not in a relationship, had not completed high school, and were employed. This unbalanced sample distribution underscores the importance of including demographic and socioeconomic status variables in the analytical model when analyzing interviewer effects. Comparing the statistics from the HRS to those from the ABS, we note that in the HRS, the between-interviewer SDs are generally higher and the average within-interviewer SDs are generally lower. This suggests that interviewer selection effects are potentially a larger threat when analyzing interviewer variance in the HRS. This is consistent with our expectations, as randomized mode assignment is applied in the ABS but not in the HRS.

Table 3.1
Distribution of sample characteristics in the Health and Retirement Study 2016 across interviewers by modes.

Respondent characteristics	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average within interviewer SD (FTF)	Average within interviewer SD (TEL)
Age: less than 60	0.449	0.305	0.359	0.315	0.294	0.363
Age: 60-69	0.188	0.343	0.156	0.239	0.260	0.398
Age: 70-79	0.181	0.287	0.150	0.255	0.240	0.342
Age: 80+	0.182	0.066	0.185	0.151	0.232	0.163
Currently Working	0.368	0.329	0.280	0.265	0.426	0.427
Male	0.414	0.415	0.177	0.198	0.503	0.503
Spanish-speaking Hispanic	0.091	0.085	0.194	0.212	0.087	0.072
English-speaking Hispanic	0.077	0.074	0.158	0.146	0.198	0.189
Black	0.219	0.200	0.274	0.246	0.315	0.344
White	0.613	0.641	0.315	0.300	0.376	0.397
Coupled	0.601	0.632	0.260	0.275	0.431	0.428
Education:less than 12 years	0.203	0.188	0.190	0.225	0.348	0.324
Education:12 years	0.303	0.290	0.188	0.218	0.420	0.416
Education:13-15 years	0.259	0.268	0.196	0.200	0.406	0.421
Education:16 years +	0.249	0.262	0.214	0.220	0.391	0.374

Note: FTF = face-to-face; TEL = telephone; SD = standard deviation.

Next, we present the descriptive statistics of the HRS, including mode-specific sample means, between-interviewer SDs, and average within-interviewer SDs in Table 3.2. First, for the CESD scale, the prevalence rates are generally higher in FTF interviews than in TEL interviews, suggesting that FTF may be associated with more honest reporting. Second, the magnitude of the between-interviewer SDs appears larger in the interviewer observation and physical activity items compared to the CESD items, indicating potentially different levels of interviewer effects in different outcomes.

Table 3.2
Distribution of outcome variables in the Health and Retirement Study 2016 across interviewers by modes.

Questions	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average within interviewer SD (FTF)	Average within interviewer SD (TEL)
CESD questions						
1. you felt depressed.	0.156	0.117	0.177	0.148	0.306	0.268
2. you felt that everything you did was an effort.	0.336	0.252	0.230	0.215	0.431	0.389
3. your sleep was restless.	0.352	0.301	0.222	0.219	0.442	0.428
4. you were happy (REVERSED CODE).	0.174	0.142	0.190	0.173	0.325	0.295
5. you felt lonely.	0.207	0.152	0.191	0.158	0.365	0.321
6. you enjoyed life (REVERSED CODE).	0.113	0.077	0.151	0.115	0.261	0.214
7. you felt sad.	0.246	0.192	0.218	0.184	0.381	0.353
8. you could not get going.	0.211	0.173	0.178	0.176	0.375	0.332
9. Depressed (≥ 4 symptoms)	0.182	0.117	0.185	0.138	0.338	0.275
Interviewer observations						
10. attentive to the questions	0.799	0.797	0.210	0.230	0.334	0.320
11. understanding of the questions	0.463	0.474	0.272	0.304	0.437	0.405
12. cooperation	0.716	0.660	0.259	0.284	0.377	0.396
13. difficulty remembering things	0.539	0.588	0.288	0.316	0.422	0.383
14. difficulty hearing you	0.803	0.741	0.202	0.254	0.325	0.359
15. quality of this interview	0.591	0.623	0.323	0.326	0.378	0.366
Physical activity						
16. vigorous sports or activities	0.352	0.347	0.218	0.221	0.443	0.454
17. moderately energetic sports or activities	0.673	0.651	0.216	0.234	0.426	0.441
18. mildly energetic sports or activities	0.806	0.771	0.168	0.207	0.362	0.379

Note: FTF = face-to-face; TEL = telephone; SD = standard deviation; CESD = Center for Epidemiological Studies Depression.

3.3.2 Mode effects in means and interviewer variances

Last, we discuss the modeling results presented in Table 3.3 using Bayesian estimation. Positive mode effects in means are found in four of the nine depression items. These items are felt depressed, everything was an effort, sleep was restless, and an overall indicator for depression. For example, for a female under 60 years old, who is an English-speaking Hispanic, not in a relationship, not currently employed, and with less than a high school education, participating in a FTF interview increases the probability of being classified as depressive by 8.01%, compared to a TEL interview. Similarly, we compute 8.01% using $\phi(\beta_0 + \beta_1 + \sum_s^S \gamma_s x_{si})\,\beta_1$, where ϕ is the pdf of a standard normal distribution and S is the number of covariates (x). Since depressive symptoms constitute sensitive information, and admitting to them might cause embarrassment for respondents, we believe that a higher level of reported depressive symptoms is closer to the truth. For the interviewer observation items, positive mode effects in means are present in three out of six items. In the FTF mode, interviewers rated respondents as more cooperative, with better hearing and overall quality of the interview, compared to the TEL mode (Table 3.3). Lastly, in the physical activity

items, respondents tend to report engaging in mildly energetic sports more often when responding via FTF, compared to TEL.

We observe smaller interviewer variances in the substantive responses in the HRS (Table 3.3) compared to the ABS. For depression items, the interviewer variances in FTF and TEL range from 0.002 to 0.032, corresponding to ICCs between 0.002 and 0.032. In the physical activity items, the interviewer variances range from 0.007 (ICC: 0.007) to 0.031 (ICC: 0.030). When comparing the magnitude of interviewer variances across variables, we notice larger interviewer variances for the interviewer observation items (ranging from 0.271 [ICC: 0.213] to 0.881 [ICC: 0.468]).

In terms of mode effects in interviewer variances, we find significant differences for three out of the eighteen questions examined in the HRS study, specifically one in the depression scale and two in the interviewer observation questions (Table 3.3). When asking participants if they felt sad, the results reveal that FTF is associated with larger interviewer variances. Additionally, interviewer variance in the FTF mode is marginally larger than in the TEL mode for the item everything was an effort. Generally, for the depression items, the interviewer variances in the FTF mode are larger than those in the TEL mode for seven out of nine items, though not always significantly. This outcome aligns with the Arab Barometer findings and may be due to interviewers approaching sensitive items differently in FTF compared to the TEL mode.

In assessing whether respondents have any difficulty remembering and hearing things, the results suggest that TEL interviewer variances are larger than FTF interviewer variances. This finding may be attributed to interviewers having fewer cues to evaluate interview quality in TEL, as opposed to FTF, where interviewers can rely on respondents' facial expressions or body language to infer participants' ability to hear questions. This might lead to responses being primarily determined by interviewers' subjective judgments and thus causing larger variances. Regarding the physical activity items, there is no evidence to reject the null hypothesis that interviewer variances are equal between modes.

It is not surprising to find higher correlations ($\rho > 0.8$) between the random interviewer effects across modes for interviewer observation variables, which interviewers directly answer. In contrast, for the other two scales (CESD and physical activity scales), the effects of interviewers on responses are mediated through respondents, resulting in a smaller and less stable correlation between the FTF and TEL modes.

Although we focus on reporting the Bayesian results, we provide the inferences from both the likelihood and the Bayesian procedures in Appendix B. We note that, in general, the estimates from the two procedures are similar, except when estimating the correlation (ρ). The correlations are associated with wide intervals in the CESD scales and the physical activity items. Moreover, the point estimates of the correlation are sometimes quite different between the two procedures, especially for the two types of items mentioned above. On two items, happy and felt sad, the correlation cannot be estimated using the likelihood approach. This might be due to the small interviewer variances in the scale, making the estimation of the covariance numerically challenging and thus unstable. Additionally, this might be attributed to the unbalanced interviewer burden between modes. Approximately 30% of interviewers only conduct interviews in one mode, and 51% of interviewers carry out fewer than five interviews in either FTF or TEL. This imbalance may result in insufficient information for estimating ρ.

Table 3.3
Interviewer variances per mode for selected items in Health and Retirement Study adjusting for covariates using Bayesian estimation.

Questions	σ_f^2	σ_t^2	$\rho_{f,\text{int}}$	$\rho_{t,\text{int}}$	α	β_1	ρ
CESD questions							
felt depressed	0.011	0.013	0.011	0.013	0.044	**0.056**	0.07
	[0, 0.022]	[0, 0.03]	[0.000, 0.022]	[0.000, 0.029]	[-1.148, 1.533]	[0.005, 0.114]	[-0.551, 0.874]
everything was an effort	0.025	0.007	0.024	0.007	0.746	**0.118**	-0.128
	[0.013, 0.037]	[0.001, 0.016]	[0.013, 0.036]	[0.001, 0.016]	[-0.002, 1.496]	[0.071, 0.175]	[-0.56, 0.254]
restless sleep	0.002	0.005	0.002	0.005	-0.486	**0.053**	0.337
	[0, 0.007]	[0, 0.012]	[0.000, 0.007]	[0.000, 0.012]	[-1.89, 0.925]	[0.011, 0.095]	[-0.162, 0.849]
happy	0.011	0.011	0.011	0.011	0.128	0.032	**-0.518**
	[0.003, 0.021]	[0, 0.022]	[0.003, 0.021]	[0.000, 0.022]	[-0.889, 1.333]]	[-0.024, 0.083]	[-0.989, -0.006]
lonely	0.006	0.006	0.006	0.006	0.178	0.048	0.055
	[0, 0.014]	[0, 0.016]	[0.000, 0.014]	[0.000, 0.016]	[-1.455, 1.846]]	[-0.005, 0.099]	[-0.108, 0.218]
enjoyed life	0.01	0.007	0.010	0.007	0.551	0.061	**0.56**
	[0.001, 0.021]	[0, 0.025]	[0.001, 0.021]	[0.000, 0.024]	[-1.096, 2.148]]	[-0.005, 0.134]	[0.223, 0.921]
felt sad	0.032	0.003	0.031	0.003	**1.694**	0.046	**0.296**
	[0.018, 0.048]	[0, 0.009]	[0.018, 0.046]	[0.000, 0.009]	[0.463, 3.775]	[-0.01, 0.097]	[0.037, 0.577]
could not get going	0.02	0.02	0.020	0.020	-0.051	0.051	0.274
	[0.007, 0.029]	[0.006, 0.035]	[0.007, 0.028]	[0.006, 0.034]	[-0.732, 0.515]	[-0.01, 0.109]	[-0.346, 0.797]
overall indicator	0.016	0.012	0.016	0.012	0.093	**0.15**	0.244
	[0.002, 0.024]	[0.001, 0.027]	[0.002, 0.023]	[0.001, 0.026]	[-0.901, 1.075]	[0.102, 0.207]	[-0.18, 0.573]
Interviewer Observations							
attentive	0.298	0.351	0.230	0.260	-0.081	0.018	**0.878**
	[0.233, 0.356]	[0.262, 0.431]	[0.189, 0.263]	[0.208, 0.301]	[-0.197, 0.038]	[-0.049, 0.088]	[0.803, 0.955]
understanding	0.413	0.465	0.292	0.317	-0.058	0	**0.91**
	[0.341, 0.493]	[0.366, 0.56]	[0.254, 0.330]	[0.268, 0.359]	[-0.149, 0.043]	[-0.064, 0.061]	[0.861, 0.958]
cooperation	0.459	0.41	0.315	0.291	0.057	**0.178**	**0.931**
	[0.378, 0.556]	[0.321, 0.51]	[0.274, 0.357]	[0.243, 0.338]	[-0.039, 0.138]	[0.108, 0.236]	[0.881, 0.971]
remembering	0.483	0.605	0.326	0.377	**-0.112**	-0.062	**0.931**
	[0.392, 0.574]	[0.489, 0.721]	[0.282, 0.365]	[0.328, 0.419]	[-0.205, -0.028]]	[-0.124, 0.002]	[0.885, 0.972]
hearing	0.271	0.375	0.213	0.273	**-0.161**	**0.151**	**0.87**
	[0.212, 0.335]	[0.274, 0.462]	[0.175, 0.251]	[0.215, 0.316]	[-0.284, -0.037]	[0.084, 0.229]	[0.795, 0.947]]
Overall quality	0.881	0.788	0.468	0.441	0.057	**0.086**	**0.94**
	[0.749, 1.04]	[0.641, 0.949]	[0.428, 0.510]	[0.391, 0.487]	[-0.032, 0.14]	[0.014, 0.158]	[0.913, 0.983]
Physical activity							
vigorous sports	0.017	0.007	0.017	0.007	0.523	-0.037	0.36
	[0.007, 0.026]	[0, 0.015]	[0.007, 0.025]	[0.000, 0.015]	[-0.209, 1.45]	[-0.081, 0.014]	[-0.446, 0.827]
moderate sport	0.015	0.019	0.015	0.019	-0.086	0.031	0.233
	[0.006, 0.024]	[0.004, 0.033]	[0.006, 0.023]	[0.004, 0.032]	[-0.655, 0.464]	[-0.019, 0.078]	[-0.351, 0.698]
mild sport	0.02	0.031	0.020	0.030	-0.355	**0.134**	0.144
	[0.002, 0.03]	[0.014, 0.052]	[0.002, 0.029]	[0.014, 0.049]	[-1.097, 0.324]	[0.073, 0.184]	[-0.264, 0.962]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,\text{int}}$ is the interviewer intraclass correlation associated with the FTF mode. $\rho_{t,\text{int}}$ is the interviewer intraclass correlation associated with the TEL mode. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects. CESD = Center for Epidemiological Studies Depression.

To address the numerical challenges and evaluate whether the estimation of other parameters (e.g., σ_f^2, σ_t^2, and α) is sensitive to ρ, we set ρ to 0 and to the posterior mean obtained with the Bayesian procedure, and rerun model (3.1) for the CESD items. We find that the estimates of the interviewer variances remain nearly unchanged when specifying ρ to different values or estimating ρ. The results can be found in Appendix G of (Yu, Elliott and Raghunathan, 2024). Thus, we conclude that there is little sensitivity in the inferences provided by the likelihood estimation to ρ.

4. Simulation study

To understand the repeated sampling properties of our proposed method, including the power to detect mode effects in the typically modest interviewer sample sizes available, we conducted simulation studies using the ABS and the HRS setup.

4.1 Arab Barometer Study

This simulation study is designed such that the number of respondents ($n = 2{,}521$) and interviewers (13 in the TEL mode and 31 in the FTF mode) are the same as the ABS, as well as how respondents are matched to interviewers. We consider four scenarios, 1) no difference scenario where the FTF interviewer variance is equal to the TEL interviewer variance ($\sigma_f^2 = \sigma_t^2 = 0.14,\ \alpha_0 = -0.98$ and $\alpha = 0$), 2) small differences where $\sigma_f^2 = 0.20,\ \sigma_t^2 = 0.14,\ \alpha_0 = -0.98$ and $\alpha = 0.18$, 3) medium differences where $\sigma_f^2 = 0.24,\ \sigma_t^2 = 0.14,\ \alpha_0 = -0.98$ and $\alpha = 0.27$, and 4) large differences where $\sigma_f^2 = 0.50,\ \sigma_t^2 = 0.14,\ \alpha_0 = -0.98$ and $\alpha = 0.64$. We consider the true data generation model as follows:

$$\begin{aligned} \eta_i &= \Phi(\beta_0 + \beta_1 M_{ij} + b_{j(m)}), \\ b_{j(m)} &\sim N(0, \sigma_m^2), \\ y_i &\sim \text{Bernoulli}(\eta_i), \end{aligned}$$

where i indexes respondents, j indexes interviewers, m indicates modes (f or t), $\Phi()$ is the cumulative distribution function of the standard normal distribution, and M is a $n \times 1$ vector of the mode that each participant used to participate in the survey.

We fit the same analytical model (2.2) to the simulated data, implemented separately using Proc Nlmixed and Proc MCMC in the SAS programming language. The simulation is repeated $K = 200$ times, where for each iteration, the point estimates, standard errors, and 95% confidence intervals or credible intervals of β_1, σ_f^2, σ_t^2, and α are computed and saved. Based on these statistics, we report the bias, coverage rate, SE ratio, and power in each scenario for the parameters.

$$\text{Bias}(\hat{\delta}) = \frac{1}{K}\sum_k^K \hat{\delta}_k - \delta,$$

$$\text{Coverage Rate}(\hat{\delta}) = \frac{1}{K}\sum_k^K I(\hat{\delta}_{k,\text{lw}} < \delta \ \&\ \hat{\delta}_{k,\text{up}} > \delta),$$

$$\text{SE Ratio}(\hat{\delta}) = \frac{1}{K}\sum_k^K \sqrt{\hat{\text{var}}(\hat{\delta}_k)} \Big/ \sqrt{\frac{1}{K-1}\sum_k^K (\hat{\delta}_k - \bar{\hat{\delta}}_k)^2},$$

$$\text{Power}(\hat{\delta}) = 1 - \frac{1}{K}\sum_k^K I(\hat{\delta}_{k,\text{lw}} < 0 \ \&\ \hat{\delta}_{k,\text{up}} > 0) \ \text{ when } \delta \neq 0,$$

where δ refers to the parameters that we are interested in estimating (i.e., σ_f^2, σ_t^2, β_1, and α), $\hat{\delta}_k$ is the estimated point estimate of δ obtained in iteration k, $\hat{\delta}_{k,\text{lw}}$ and $\hat{\delta}_{k,\text{up}}$ is the lower bound and upper bound of the estimated parameter.

Table 4.1 displays the simulation results using the ABS setup. When $\alpha = 0$, the power reported in Table 4.1 represents the Type 1 error rate. We observe that the power to reject the null hypothesis stating that interviewer variances are equal $(\alpha = 0)$ is limited across the scenarios. However, as the differences grow larger (0.18-0.64), the power does increase from 0.075 to 0.520 in the Bayesian procedure and from 0.110 to 0.633 in the frequentist approach. There are some differences in the power provided by the likelihood and Bayesian approaches. This is because the likelihood procedures do not offer nominal coverage rates in Scenarios 1 to 3; as a result, the power obtained from the likelihood and Bayesian procedures is based on different significance levels. The small power of α is primarily due to the very limited number of interviewers in both FTF and TEL modes. Conversely, the power of rejecting the null hypothesis that there are no mode effects in means (β_1) when the alternative hypothesis is true is considerably higher (around 0.90). However, as α becomes larger and the interviewer variances increase simultaneously, we observe a declining power of β_1, due to the decline in effective sample size from the increased ICC.

Table 4.1
Simulation study using the Arab Barometer Study setup.

Parameters	Likelihood results				Bayesian results			
	Bias	Coverage rate	SE ratio	Power	Bias	Coverage rate	SE ratio	Power
Scenario 1: No differences								
$\sigma_f^2 = 0.14$	-0.002	0.950	1.000	N/A	0.017	0.940	1.059	N/A
$\sigma_t^2 = 0.14$	-0.001	0.955	1.014	N/A	0.049	0.975	1.346	N/A
$\beta_1 = 0.5$	-0.003	0.965	1.023	0.935	0.006	0.955	1.121	0.930
$\alpha = 0$	0.028	0.930	0.888	0.070	-0.033	0.985	1.107	0.015
Scenario 2: Small differences								
$\sigma_f^2 = 0.20$	-0.012	0.960	0.948	N/A	0.028	0.975	1.105	N/A
$\sigma_t^2 = 0.14$	-0.007	0.935	0.974	N/A	0.059	0.955	1.161	N/A
$\beta_1 = 0.5$	-0.002	0.940	0.926	0.950	-0.001	0.950	1.078	0.900
$\alpha = 0.18$	0.042	0.920	0.928	0.110	-0.020	0.950	0.955	0.075
Scenario 3: Medium differences								
$\sigma_f^2 = 0.24$	-0.002	0.920	0.947	N/A	0.039	0.920	0.980	N/A
$\sigma_t^2 = 0.14$	-0.013	0.955	1.009	N/A	0.061	0.980	1.311	N/A
$\beta_1 = 0.5$	0.004	0.935	0.940	0.920	-0.010	0.960	1.184	0.860
$\alpha = 0.27$	0.079	0.905	0.922	0.230	-0.042	0.960	1.075	0.085
Scenario 4: Large differences								
$\sigma_f^2 = 0.50$	-0.007	0.970	1.058	N/A	0.078	0.950	1.093	N/A
$\sigma_t^2 = 0.14$	-0.009	0.960	1.055	N/A	0.054	0.935	1.231	N/A
$\beta_1 = 0.5$	0.022	0.935	0.965	0.824	-0.016	0.980	1.097	0.690
$\alpha = 0.64$	0.079	0.945	0.906	0.633	0.012	0.955	0.882	0.520

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances.

4.2 Health and Retirement Study

In the simulation study using the HRS setup, we consider the following data generation model using the same notations as in the ABS simulation study. We use b_{jf} to represent random interviewer effects in the FTF mode and b_{jt} to represent random interviewer effects in the TEL mode:

$$\begin{aligned} \eta_i &= \Phi(\beta_0 + \beta_1 M_{ij} + b_{jf} M_{ij} + b_{jt}(1 - M_{ij})), \\ \begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix} \right), \\ y_i &\sim \text{Bernoulli}(\eta_i). \end{aligned}$$

We consider four scenarios: 1) $\sigma_f^2 = \sigma_t^2 = 0.03$, $\alpha_0 = -1.75$ and $\alpha = 0$; 2) $\sigma_f^2 = 0.05$, $\sigma_t^2 = 0.03$, $\alpha_0 = -1.75$, and $\alpha = 0.26$; 3) $\sigma_f^2 = 0.06$, $\sigma_t^2 = 0.03$, $\alpha_0 = -1.75$, and $\alpha = 0.35$; 4) $\sigma_f^2 = 0.09$, $\sigma_t^2 = 0.03$, $\alpha_0 = -1.75$ and $\alpha = 0.55$. Across all scenarios, $\beta_1 = 0.5$ and $\rho = 0.5$. We report bias, coverage rate, SE ratio, and power for these parameters and the logarithmic differences of interviewer variances between FTF and TEL (α) in Table 4.2.

Table 4.2
Simulation study using the Health and Retirement Study setup.

Parameters	Likelihood results				Bayesian results			
	Bias	Coverage rate	SE ratio	Power	Bias	Coverage rate	SE ratio	Power
Scenario 1: No differences								
$\sigma_f^2 = 0.03$	-0.000	0.980	1.085	N/A	0.003	0.965	1.704	N/A
$\sigma_t^2 = 0.03$	-0.001	0.975	1.049	N/A	0.002	0.935	1.469	N/A
$\beta_1 = 0.5$	-0.002	0.940	1.029	1.000	-0.000	0.960	1.128	1.000
$\rho = 0.5$	0.012	0.965	1.009	0.470	-0.020	0.925	1.061	0.690
$\alpha = 0$	0.022	0.965	1.019	0.035	0.047	0.965	0.928	0.035
Scenario 2: Small differences								
$\sigma_f^2 = 0.05$	0.000	0.940	0.999	N/A	0.001	0.955	1.507	N/A
$\sigma_t^2 = 0.03$	-0.000	0.975	1.125	N/A	0.002	0.945	1.249	N/A
$\beta_1 = 0.5$	0.003	0.960	0.996	1.000	0.001	0.950	0.983	1.000
$\rho = 0.5$	0.020	0.980	1.084	0.695	-0.021	0.925	1.032	0.755
$\alpha = 0.26$	0.018	0.940	0.978	0.270	0.008	0.940	0.934	0.295
Scenario 3: Medium differences								
$\sigma_f^2 = 0.06$	-0.001	0.945	0.999	N/A	0.001	0.950	1.268	N/A
$\sigma_t^2 = 0.03$	-0.001	0.975	1.045	N/A	0.002	0.940	1.103	N/A
$\beta_1 = 0.5$	-0.001	0.920	0.993	1.000	0.001	0.965	1.007	1.000
$\rho = 0.5$	0.011	0.970	1.030	0.665	-0.009	0.930	1.014	0.815
$\alpha = 0.35$	0.024	0.910	0.919	0.510	0.008	0.945	0.949	0.530
Scenario 4: Large differences								
$\sigma_f^2 = 0.09$	0.000	0.930	0.983	N/A	0.002	0.950	1.201	N/A
$\sigma_t^2 = 0.03$	-0.001	0.955	1.054	N/A	-0.001	0.915	1.089	N/A
$\beta_1 = 0.5$	0.004	0.950	1.009	1.000	-0.002	0.955	1.031	1.000
$\rho = 0.5$	0.009	0.985	1.085	0.750	0.004	0.970	1.121	0.860
$\alpha = 0.55$	0.029	0.915	0.977	0.935	0.070	0.950	0.955	0.990

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects.

Table 4.2 illustrates that as α rises from 0 to 0.55, the power correspondingly increases from 0.035 to 0.990 using the Bayesian procedure, and from 0.035 to 0.935 employing the likelihood approach. The findings suggest that When α is large enough, we can achieve a reasonably high power using the HRS setup. Upon comparing Table 4.1 and Table 4.2, we observe that the power to reject the null hypothesis asserting equal interviewer variances, when the alternative hypothesis holds true, surpasses that in the ABS simulation. This outcome aligns with expectations, given the larger number of interviewers involved in the HRS. In addition, we note that the likelihood approach may not always reach the 95% nominal coverage rates (in Scenarios 3 and 4), thus the power computed using the likelihood and the Bayesian procedures are based on different significance levels.

5. Discussion

This paper explores the presence of mode effects in interviewer variances across multiple items in two national surveys. In the ABS, we find statistical evidence for differing interviewer effects between the FTF and TEL modes in one (marginally) out of six sensitive items and one out of three item missing indicators. Besides, for sensitive items and missing indicators in the ABS, interviewer variances from the FTF mode are generally larger than those from the TEL mode. Meanwhile, we should interpret the ABS results with caution. Due to the small number of interviewers used in the study, null findings cannot be translated into small or no effects, somewhat hampering the strength of the evidence. Utilizing the 2016 HRS data, we observe significant mode effects in interviewer variances on two depression items (one marginally) and two interviewer observation item. For sensitive depression items, a similar pattern emerges, with larger interviewer variances in FTF than in TEL. These findings indicate that sensitive questions and item missing items are crucial challenges when stabilizing interviewer variances between modes. Besides, the magnitude of interviewer variances is much larger on interviewer observation items than substantive responses. In addition, evidence suggests that TEL interviewer variances are larger than FTF interviewer variances on these items. This could be because these questions involve more subjective evaluations and may offer greater opportunities to reduce interviewer variances by standardizing interviewer protocols for such items, especially in the TEL mode.

Simulation studies suggest that it is possible to achieve reasonable power with the ABS or HRS setup if there are substantial mode effects in interviewer variances. However, with small mode effects, the power is limited, especially in the ABS setup. The observation of significant mode effects in interviewer variances in both the ABS and HRS data highlights the importance of considering the role of modes on interviewer effects, particularly when addressing sensitive topics and item nonresponse. Given the typically limited number of interviewers employed in most surveys, a null finding may not necessarily indicate equal interviewer variance. However, it is still useful for survey agencies to consider such investigation as a positive finding is valid and should capture the attention of researchers. Moreover, in the presence of multiple underpowered studies that employ few interviewers, a meta-analysis can be conducted to combine the inferences made from these studies and better explore the mode effects in interviewer variances.

The literature has extensively documented whether modes affect measurement errors at the respondent level (Tourangeau and Smith, 1996; Kreuter, Presser and Tourangeau, 2008). However, few studies have investigated whether and how modes influence interviewer-related measurement errors, particularly following the widespread adoption of mixed-mode designs. This paper addresses this gap by analyzing two national surveys with distinct mixed-mode design features, such as the number of interviewers and whether the interviewers are nested under modes. When interviewers are nested under modes, it is hard to determine if the observed differences are attributable to modes or interviewers. The current modeling approach presumes that all systematic differences between responses collected in TEL and FTF are a consequence of modes, not interviewers. If survey organizations possess information on interviewer characteristics, they can evaluate this assumption by comparing the characteristics of interviewers between modes. Such an analysis would help disentangle the effects of modes from those of interviewers, providing valuable insights for survey data quality.

For designs that allow interviewers to collect data in both modes, the models presented in this paper enable the estimation of individual interviewer effects in each mode. This is useful for detecting interviewers with a substantial impact on responses in one or both modes. Utilizing these estimated interviewer effects, we can further identify if specific interviewers consistently exhibit large effects across variables, potentially signaling the need for intervention by interviewer supervisors. If particular variables are associated with significant interviewer variances in a certain mode, this may warrant improved interviewer training for those items. For instance, based on this study's findings, a more standardized interview protocol could be considered for sensitive items and when respondents answer don't know to questions in FTF mode. As such, we recommend that survey agencies incorporate these analyses into their routine data quality assessments. Future research could investigate whether interviewer characteristics can explain the differential interviewer effects observed across modes, potentially shedding light on the underlying mechanisms at play.

When determining which mode to use for generating population estimates in mixed-mode studies, it is desirable to have smaller bias and lower interviewer variances, which might result in smaller mean squared error. However, in reality, the mode with smaller bias and lower interviewer variance may not always be the same, as shown in this paper. For instance, FTF interviews may be linked with less bias but larger interviewer variance. How to balance the trade-offs between bias and variance in a formal method will be a topic for future research. This study showcases two survey examples to evaluate mode effects both in means and interviewer variances. If such analyses are routinely adopted by researchers who design and implement mixed-mode studies, more evidence can be accumulated about whether and how interviewers could have performed differently in different modes of data collection. This can become the basis for developing future mixed-mode protocols. When reporting the results of the analysis, we recommend that survey agencies explain how their interviewers are assigned to or self-select different modes and clarify whether the observed mode effects in interviewer variances are consistent with their expectations.

In this paper, we observe some discrepancies between the results obtained from the maximum likelihood procedure and the Bayesian procedure implemented in the SAS programming language. When interviewer

variances are small, fitting the analytical model with correlated interviewer random effects across modes using the likelihood approach can be challenging. In this situation, the Bayesian approach can be particularly useful, as employing proper and informative priors helps ensure that we draw inferences from proper posterior distributions.

This study has three main limitations. First, like other similar studies (West, Ong, Conrad, Schober, Larsen and Hupp, 2022; Groves and Magilavy, 1986), it faces the issue of limited statistical power, as demonstrated in the simulation study. Second, we consider dichotomized outcomes in this study due to computational reasons; however, this may not be an optimal approach for studying interviewer variance, as collapsing categories may reduce variances. Future studies can explore this research question using different types of outcomes and larger sample sizes. Last, both surveys lack randomization in the interviewer assignment scheme. Ideally, when estimating interviewer variances, interpenetrated designs should be used to ensure that the variability is solely due to the interviewer measurement process, rather than differences among respondents. As a workaround for the absence of randomization, we included respondent characteristics in the analysis model. However, interviewer variances might still be overestimated due to unobserved covariates not accounted for in the models.

Acknowledgements

This project was supported by the Daniel Katz Dissertation Fellowship in Psychology and Survey Methodology at the University of Michigan Institute for Social Research. The authors thank the investigators, the staff, and the participants of the Arab Barometer Study and the Health and Retirement study for their valuable contributions. We appreciate Dr. Brady West for his invaluable assistance with this study. We acknowledge that we utilized the AI language model, ChatGPT, developed by OpenAI, for assistance with grammar correction and refinement of the wording in this paper.

Appendix A

Derivations of the variance of α using Delta Method

$$
\begin{aligned}
\text{var}(\alpha) &= \text{var}\left(\log(\sigma_f) - \log(\sigma_t)\right) \\
&= \text{var}\left(\log(\sigma_f)\right) + \text{var}\left(\log(\sigma_t)\right) - 2\text{cov}\left(\log(\sigma_f), \log(\sigma_t)\right) \\
&= \frac{1}{4}\text{var}\left(2\log(\sigma_f)\right) + \frac{1}{4}\text{var}\left(2\log(\sigma_t)\right) - 2\text{cov}\left(\log(\sigma_f), \log(\sigma_t)\right) \\
&= \frac{1}{4}\text{var}\left(\log(\sigma_f^2)\right) + \frac{1}{4}\text{var}\left(\log(\sigma_t^2)\right) - \frac{1}{4} \times 2\text{cov}\left(\log(\sigma_f^2), \log(\sigma_t^2)\right) \\
&= \frac{1}{4}\text{var}\left(\log(\sigma_f^2)\right) + \frac{1}{4}\text{var}\left(\log(\sigma_t^2)\right) - \frac{1}{2}\text{cov}\left(\log(\sigma_f^2), \log(\sigma_t^2)\right).
\end{aligned}
$$

We express $\text{var}(\alpha)$ as a function of $\text{var}(\log(\sigma_f^2))$, $\text{var}(\log(\sigma_t^2))$, and $\text{cov}(\log(\sigma_f^2), \log(\sigma_t^2))$, as we apply a log transformation to σ_t^2 and σ_f^2 to stabilize their variances. The covariance between $\log(\sigma_f^2)$ and $\log(\sigma_t^2)$ can be assumed to be 0 when the random interviewer effects of FTF and TEL are not correlated, as is the case in the ABS. In contrast, in the HRS, when the random interviewer effects are correlated across modes, the covariance between the two estimates should be considered when calculating $\text{var}(\alpha)$.

Appendix B

Full results on interviewer variances in the Health and Retirement Study

Table B.1
Interviewer variances per mode for selected items in Health and Retirement Study adjusting for covariates.

Questions	Likelihood					Bayesian				
	σ_f^2	σ_t^2	α	β_1	ρ	σ_f^2	σ_t^2	α	β_1	ρ
CESD questions										
Felt depressed	0.011	0.014	-0.095	0.056	0.222	0.011	0.013	0.044	**0.056**	0.07
	(0.006)	(0.008)	(0.389)	(0.029)	(0.448)	(0.006)	(0.009)	(0.643)	(0.029)	(0.391)
	[0.004,	[0.004,	[-0.857,	[-0.001,	[-0.603,	[0,	[0,	[-1.148,	[0.005,	[-0.551,
	0.031]	0.045]	0.667]	0.113]	0.818]	0.022]	0.03]	1.533]	0.114]	0.874]
Everything was an effort	0.022	0.004	0.893	**0.116**	-0.099	0.025	0.007	0.746	**0.118**	-0.128
	(0.006)	(0.005)	(0.682)	(0.025)	(0.618)	(0.014)	(0.005)	(0.38)	(0.029)	(0.264)
	[0.013,	[0,	[-0.445,	[0.066,	[-0.867,	[0.013,	[0.001,	[-0.002,	[0.071,	[-0.56,
	0.037]	0.052]	2.23]	0.165]	0.809]	0.037]	0.016]	1.496]	0.175]	0.254]
Restless sleep	0.003	0.004	-0.13	**0.057**	-0.698	0.002	0.005	-0.486	**0.053**	0.337
	(0.003)	(0.004)	(0.719)	(0.022)	(1.018)	(0.002)	(0.004)	(0.754)	(0.021)	(0.312)
	[0,	[0,	[-1.54,	[0.013,	[-1,	[0,	[0,	[-1.89,	[0.011,	[-0.162,
	0.02]	0.032]	1.279]	0.1]	0.995]	0.007]	0.012]	0.925]	0.095]	0.849]
happy	0.007	0.010	-0.253	0.033	NA	0.011	0.011	0.128	0.032	-0.518
	(0.015)	(0.014)	(1.519)	(0.032)	(NA)	(0.005)	(0.007)	(0.539)	(0.027)	(0.314)
	[0,	[0,	[-2.348,	[-0.019,	[NA,	[0.003,	[0,	[-0.889,	[-0.024,	[-0.989,
	0.019]	0.023]	2.915]	0.085]	NA]	0.021]	0.022]	1.333]	0.083]	-0.006]
lonely	0.006	0.004	0.223	0.046	-0.208	0.006	0.006	0.178	0.048	0.055
	(0.004)	(0.006)	(0.973)	(0.026)	(1.053)	(0.004)	(0.005)	(0.878)	(0.028)	(0.084)
	[0.001,	[0,	[-1.685,	[-0.004,	[-0.983,	[0,	[0,	[-1.455,	[-0.005,	[-0.108,
	0.025]	0.12]	2.131]	0.097]	0.96]	0.014]	0.016]	1.846]	0.099]	0.218]
Enjoyed life	0.009	0.011	-0.124	**0.07**	-0.823	0.01	0.007	0.551	0.061	**0.56**
	(0.006)	(0.009)	(0.528)	(0.033)	(0.718)	(0.006)	(0.009)	(0.944)	(0.036)	(0.187)
	[0.002,	[0.002,	[-1.16,	[0.006,	[-1,	[0.001,	[0,	[-1.096,	[-0.005,	[0.223,
	0.037]	0.052]	0.911]	0.134]	0.997]	0.021]	0.025]	2.148]	0.134]	0.921]
Felt sad	0.03	0	2.475	0.047	NA	0.032	0.003	**1.694**	0.046	**0.296**
	(0.007)	(0.001)	(2.213)	(0.025)	(NA)	(0.008)	(0.003)	(0.951)	(0.027)	(0.137)
	[0.018,	[0,	[-1.863,	[-0.003,	[NA,	[0.018,	[0,	[0.463,	[-0.01,	[0.037,
	0.048]	1.263]	6.813]	0.097]	NA]	0.048]	0.009]	3.775]	0.097]	0.577]
Could not get going	0.016	0.019	-0.098	0.05	0.314	0.02	0.02	-0.051	0.051	0.274
	(0.005)	(0.008)	(0.274)	(0.027)	(0.318)	(0.032)	(0.008)	(0.339)	(0.033)	(0.307)
	[0.008,	[0.008,	[-0.635,	[-0.003,	[-0.351,	[0.007,	[0.006,	[-0.732,	[-0.01,	[-0.346,
	0.03]	0.044]	0.439]	0.103]	0.768]	0.029]	0.035]	0.515]	0.109]	0.797]
Overall indicator	0.012	0.012	0.006	**0.152**	-0.264	0.016	0.012	0.093	**0.15**	0.244
	(0.005)	(0.007)	(0.37)	(0.028)	(0.428)	(0.028)	(0.008)	(0.487)	(0.027)	(0.211)
	[0.005,	[0.004,	[-0.718,	[0.096,	[-0.825,	[0.002,	[0.001,	[-0.901,	[0.102,	[-0.18,
	0.029]	0.04]	0.731]	0.207]	0.558]	0.024]	0.027]	1.075]	0.207]	0.573]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects. We use N/A to mask estimates that cannot be estimated due to numerical difficulties. CESD = Center for Epidemiological Studies Depression.

Table B.1(continued)
Interviewer variances per mode for selected items in Health and Retirement Study adjusting for covariates.

Questions	Likelihood					Bayesian				
	σ_f^2	σ_t^2	α	β_1	ρ	σ_f^2	σ_t^2	α	β_1	ρ
Interviewer Observations										
attentive	0.29	0.342	-0.082	0.013	**0.893**	0.298	0.351	-0.081	0.018	**0.878**
	(0.032)	(0.043)	(0.063)	(0.035)	(0.036)	(0.032)	(0.044)	(0.062)	(0.035)	(0.039)
	[0.233,	[0.268,	[-0.205,	[-0.056,	[0.795,	[0.233,	[0.262,	[-0.197,	[-0.049,	[0.803,
	0.361]	0.438]	0.041]	0.081]	0.946]	0.356]	0.431]	0.038]	0.088]	0.955]
understanding	0.408	0.461	-0.062	-0.003	**0.921**	0.413	0.465	-0.058	0	**0.91**
	(0.04)	(0.05)	(0.049)	(0.032)	(0.023)	(0.039)	(0.051)	(0.049)	(0.032)	(0.041)
	[0.336,	[0.373,	[-0.158,	[-0.066,	[0.86,	[0.341,	[0.366,	[-0.149,	[-0.064,	[0.861,
	0.494]	0.571]	0.033]	0.059]	0.956]	0.493]	0.56]	0.043]	0.061]	0.958]
cooperation	0.45	0.404	0.053	0.174	**0.941**	0.459	0.41	0.057	0.178	**0.931**
	(0.043)	(0.044)	(0.047)	(0.031)	(0.021)	(0.047)	(0.048)	(0.047)	(0.032)	(0.025)
	[0.373,	[0.327,	[-0.039,	[0.113,	[0.884,	[0.378,	[0.321,	[-0.039,	[0.108,	[0.881,
	0.542]	0.5]	0.145]	0.234]	0.971]	0.556]	0.51]	0.138]	0.236]	0.971]
remembering	0.482	0.593	**-0.103**	**-0.065**	**0.941**	0.483	0.605	**-0.112**	**-0.062**	**0.931**
	(0.047)	(0.065)	(0.047)	(0.032)	(0.019)	(0.047)	(0.059)	(0.047)	(0.033)	(0.029)
	[0.398,	[0.478,	[-0.195,	[-0.128,	[0.89,	[0.392,	[0.489,	[-0.205,	[-0.124,	[0.885,
	0.584]	0.735]	-0.011]	-0.001]	0.969]	0.574]	0.721]	-0.028]	0.002]	0.972]
hearing	0.27	0.372	**-0.161**	**0.152**	**0.888**	0.271	0.375	**-0.161**	**0.151**	**0.87**
	(0.03)	(0.046)	(0.063)	(0.034)	(0.035)	(0.032)	(0.048)	(0.065)	(0.038)	(0.064)
	[0.217,	[0.291,	[-0.285,	[0.084,	[0.796,	[0.212,	[0.274,	[-0.284,	[0.084,	[0.795,
	0.336]	0.476]	-0.038]	0.219]	0.94]	0.335]	0.462]	-0.037]	0.229]	0.947]
Overall quality	0.879	0.782	0.058	**0.09**	**0.96**	0.881	0.788	0.057	**0.086**	**0.94**
	(0.08)	(0.079)	(0.04)	(0.034)	(0.014)	(0.077)	(0.08)	(0.046)	(0.038)	(0.088)
	[0.736,	[0.642,	[-0.019,	[0.023,	[0.923,	[0.749,	[0.641,	[-0.032,	[0.014,	[0.913,
	1.051]	0.953]	0.136]	0.156]	0.98]	1.04]	0.949]	0.14]	0.158]	0.983]
Physical activity										
Vigorous sports	0.015	0.008	0.298	-0.036	0.642	0.017	0.007	0.523	-0.037	0.36
	(0.004)	(0.006)	(0.381)	(0.022)	(0.41)	(0.011)	(0.004)	(0.406)	(0.026)	(0.372)
	[0.008,	[0.002,	[-0.45,	[-0.079,	[-0.541,	[0.007,	[0,	[-0.209,	[-0.081,	[-0.446,
	0.027]	0.032]	1.045]	0.007]	0.972]	0.026]	0.015]	1.45]	0.014]	0.827]
Moderately energetic sports	0.013	0.015	-0.043	0.028	0.478	0.015	0.019	-0.086	0.031	0.233
	(0.004)	(0.006)	(0.274)	(0.023)	(0.326)	(0.008)	(0.008)	(0.28)	(0.025)	(0.248)
	[0.007,	[0.006,	[-0.581,	[-0.017,	[-0.297,	[0.006,	[0.004,	[-0.655,	[-0.019,	[-0.351,
	0.026]	0.035]	0.494]	0.073]	0.873]	0.024]	0.033]	0.464]	0.078]	0.698]
Mildly energetic sports	0.015	0.03	-0.353	**0.135**	0.107	0.02	0.031	-0.355	**0.134**	0.144
	(0.005)	(0.009)	(0.238)	(0.028)	(0.278)	(0.042)	(0.01)	(0.361)	(0.028)	(0.29)
	[0.007,	[0.016,	[-0.819,	[0.08,	[-0.416,	[0.002,	[0.014,	[-1.097,	[0.073,	[-0.264,
	0.03]	0.056]	0.114]	0.19]	0.577]	0.03]	0.052]	0.324]	0.184]	0.962]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects. We use N/A to mask estimates that cannot be estimated due to numerical difficulties. CESD = Center for Epidemiological Studies Depression.

References

Ehrlich, J.S., and Riesman, D. (1961). Age and authority in the interview. *Public Opinion Quarterly*, 39-56.

Elliott, M.N., Zaslavsky, A.M. Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M.K. and Giordano, L. (2009). Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. *Health Services Research*, 44, 501-518.

Fisher, G.G., and Ryan, L.H. (2018). Overview of the health and retirement study and introduction to the special issue. *Work, Aging and Retirement*, 4, 1-9.

Groves, R.M., and Kahn, R.L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*.

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

Hanson, R.H., and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

Heeringa, S.G., and Connor, J.H. (1995). Technical description of the health and retirement survey sample design. *Ann Arbor: University of Michigan*.

Holbrook, A.L., and Krosnick, J.A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, 74, 37-67.

Health and Retirement Study (HRS), Staff (2023). HRS core interview sample sizes and response rates. Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. Available online.

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

Kolenikov, S., and Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology*, 2, 126-158.

Kreuter, F., Presser, S. and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and WEB surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847-865.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). Interviewers and interviewing. *Handbook of Survey Research*, 2, 437-471.

Schnell, R., and Kreuter, F. (2003). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 3, 389-410.

Schuman, H., and Converse, J.M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35, 44-68.

Suzer-Gurtekin, Z.T., Heeringa, S.G. and Valliant, R. (2013). Investigating the bias of alternative statistical inference methods in mixed-mode surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3393-3407.

Tourangeau, R., and Smith, T.W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60, 275-304.

Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.

Vannieuwenhuyze, J.T.A. (2015). Mode effects on variances, covariances, standard deviations, and correlations. *Journal of Survey Statistics and Methodology*, 3, 3, 296-316.

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

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

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

West, B.T., Ong, A.R., Conrad, F.G., Schober, M.F., Larsen, K.M. and Hupp, A.L. (2022). Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology*, 10, 317-336.

Yu, W., Elliott, M.R. and Raghunathan, T.E. (2024). Investigating mode effects in interviewer variances using two representative multi-mode surveys. *arXiv preprint arXiv:2408.11874*.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 435-463
Statistics Canada, Catalogue No. 12-001-X

Robust adaptive survey design for time changes in mixed-mode response propensities

Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek and Barry Schouten[1]

Abstract

Adaptive survey designs (ASDs) tailor recruitment protocols to population subgroups that are relevant to a survey. In recent years, effective ASD optimization has been the topic of research and several applications. However, the performance of an optimized ASD over time is sensitive to time changes in response propensities. How adaptation strategies can adjust to such variation over time is not yet fully understood. In this paper, we propose a robust optimization approach in the context of sequential mixed-mode surveys employing Bayesian analysis. The approach is formulated as a mathematical programming problem that explicitly accounts for uncertainty due to time change. ASD decisions can then be made by considering time-dependent variation in conditional mode response propensities and between-mode correlations in response propensities. The approach is demonstrated using a case study: the 2014-2017 Dutch Health Survey. We evaluate the sensitivity of ASD performance to 1) the budget level and 2) the length of applicable historic time-series data. We find there is only a moderate dependence on the budget level and the dependence on historic data is moderated by the amount of seasonality during the year.

Key Words: Adaptive survey designs; Bayesian approach; Optimization; Response propensity model; Time series analysis.

1. Introduction

Adaptive survey designs (ASDs, Wagner, 2008 and Schouten, Peytchev and Wagner, 2017) have gradually become a viable choice in contemporary surveys; a single survey protocol is no longer offered to all individuals or subgroups but it is tailored to efficiently attain individual's responses based on known population characteristics and characteristics observed during fieldwork. This shift was accelerated by persistent declines in response rates, limited budgets, a larger variety of data sources, the emergence of all kinds of mobile devices, and the gradual migration to mixed-mode surveys. These developments imply more urgency and more options in design.

A key element in ASD is the optimization strategy, i.e., the set of decision rules. Such strategies rely on input on response propensities and other survey design parameters. The main approaches to optimization include case prioritization (Peytchev, Riley, Rosen, Murphy and Lindblad, 2010; Wagner, 2013 and Wagner and Hubbard, 2013), trial and error, and mathematical and statistical optimization (van Berkel, van der Doef and Schouten, 2020; Calinescu, Bhulai and Schouten, 2013 and Schouten, Calinescu and Luiten, 2013); see Schouten et al. (2017) for the advantages and disadvantages of each approach. However, in these contributions the inaccuracy in response propensities estimated from historic data was most often ignored. In mathematical programming, objectives can be parameterized as functions of response propensities acting as one of the main inputs to optimization. Error would be introduced in making decisions when true response propensities change over time and these changes are not accounted for in estimation. As a result, inaccuracy

1. Shiya Wu, Utrecht University, Department of Methodology and Statistics; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods; Mirjam Moerbeek, Utrecht University, Department of Methodology and Statistics; Barry Schouten, Statistics Netherlands, Department of Statistical Methods and Utrecht University, Department of Methodology and Statistics. E-mail: jg.schouten@cbs.nl.

renders any ASD as suboptimal, or even worse, makes it ineffective. Placing ASD optimization in a Bayesian context is natural to address this issue, yet the relevant survey methodology research is still in its infancy. Recently, Ma (2021) developed methodology to efficiently optimize a stratification by holding out for accurate estimates of response propensities in a Bayesian manner, given the most recent historic data.

Time changes in response propensities and inaccurate estimates from historic data endanger the robustness of ASD optimization; see Schouten et al. (2017) and Chun, Heeringa and Schouten (2018) for more discussion. Recently, ASD research started to focus on developing response propensity models and improving prediction accuracy; Schouten, Mushkudiani, Shlomo, Durrant, Lundquist and Wagner (2018) pioneered Bayesian updating methods to combat this bias by statistically leveraging accumulated survey data and historic data generated from past implementations of the same survey. Being the most informative, prior beliefs gathered from past survey data can enhance current data for prediction purposes. Clearly, translating external data sources to prior beliefs is a requisite for the development of response propensity models. To do so, using a literature review (West, Wagner, Coffey and Elliott, 2023) and eliciting expert knowledge (Coffey, West, Wagner and Elliott, 2020 and Wu, Schouten, Meijers and Schouten, 2022) are recent approaches to source prior information. Survey researchers treat the matter of historic data timeliness incompletely and consider response propensities at different survey phases overall, whereas some facts, such as consistently reduced response rates over years, indicate that accurate estimates of response propensities are dependent on time, and response propensities in sequential designs are likely to correlate.

The most closely related work by Wu, Boonstra, Moerbeek and Schouten (2023) explored deconstructing time changes in response propensities at multiple levels to study the influence of the length of applicable historic survey data on response propensity prediction accuracy. There, only the Computer-Assisted Web Interviewing (CAWI) data collection phase in the Dutch Health Survey (GEZO) was considered, and not the Computer-Assisted Personal Interviewing (CAPI) phase that applies to the CAWI non-respondents. In the present study, we generalize the model development for response propensity prediction to multiple phases of data collection, and in particular to the case of both CAWI and CAPI mode data collection phases of GEZO. This allows, e.g., to evaluate conditional prediction accuracy of CAPI response propensities given the CAWI response realization in a certain period. As in Wu et al. (2023), we adopt a Bayesian approach that allows full uncertainty quantification of response propensities and derived quantities. The second extension of this paper relative to Wu et al. (2023) is the analysis of ASD performance under various external constraints, taking into account the uncertainty of response propensity predictions.

Taken together, this paper aims to make two contributions to sequential mixed-mode (MM) designs: predicting each survey mode response propensity as accurately as possible and making adaptive decisions in as optimally as possible. To fulfil this ambition by leveraging historic time-series data in the evaluation, we raise three research questions:

- How can time-series models be constructed to improve response propensity prediction accuracy in a sequential mixed-mode design?
- How sensitive is ASD performance to the specified budget level?
- How does ASD performance depend on the length of applicable historic data?

In response to the first question, we extend the binomial multilevel time-series models for a single mode proposed by Wu et al. (2023) to multinomial multilevel time-series models for multiple modes, and illustrate by means of an application to the GEZO survey, for which we use data from the period 2014-2017. This extension also considers the incorporation of between-mode correlation parameters in modeling the response propensities of CAWI and CAPI.

Concerning the second and third questions, a strategy that accommodates uncertainty about input to optimization when optimizing probabilistic allocations is in great demand. The survey design performance is monitored by evaluating representation of relevant background characteristics linked through administrative data. The representation is operationalized through the coefficient of variation (CV) of response propensities (Schouten, Cobben and Bethlehem, 2009). We benchmark the ASD performance against the performance of CAWI-only and nonadaptive designs to ensure that the determined allocations can improve the ASD performance. To determine the sensitivity of ASD performance, we conduct two experiments. In the first experiment, we select one quarter of the year and gradually decrease the budget level. This experiment enables us to explore the sensitivity of the ASD performance to budget constraints. In the second experiment, the budget level is fixed, and the time series window of historic data moves forward to the next new data collection quarter. The historic data is first used to set a prior. Next, with incoming new quarters, the prior is repeatedly updated to a posterior that serves as prior to the upcoming quarter. We evaluate how the prior changes and if and how this affects allocations of sample units.

The outline of this paper is as follows: we begin by constructing the time-series model for sequential mixed-mode designs in Section 2. In Section 3, we describe the optimization problem. We introduce the case study in Section 4 and address the research questions. In the last section, we discuss the advantages and disadvantages of our method and conclude with some thoughts on future research.

2. Methods

In this section, a multivariate time series model is developed for response propensities in sequential mixed-mode designs. We extend the multi-level time series model suggested by Wu et al. (2023) by introducing conditional response propensities of follow-up modes.

2.1 Modeling response propensities in sequential mixed-mode designs

The time series model of Wu et al. (2023) generates precise estimates of response propensities for survey designs with a sole mode, or for the first mode of mixed-mode surveys during fieldwork. Here, the objective evolves into making reliable predictions for each mode of mixed-mode surveys to broaden the model's appeal. Notably, discrete-valued time series data, including the size of a sample and the number of respondents to each mode, are considered from a multinomial distribution, while Wu et al. (2023) considered a binomial distribution for data.

The response propensity (RP) is the theoretical propensity of a sampled subject being a responder in a specific interview mode given a set of known characteristics. These characteristics may include paradata collected in a particular phase or mode of the survey. This subject can be either an individual or a well-defined group of persons. Of interest, a group is made by cross-classifying several auxiliary variables that are regarded as strong predictors of survey variables. Within a group, units have homogeneous demographic attributes, such as age. Such a set of groups can vary with time or design change (see Schouten et al., 2017 for more discussion on stratification), but here we assume stratification is constant and specified before fitting the model.

To model mode-level RPs, in this section, we suppress the subscripts indicating a specific group in the propensity parameter and indicating a specific time point, but the next section must specify this subscript to decompose a time series to some fixed or random effects at the stratum, time, and/or mode level.

Assume that a mixed-mode survey is provided with $M-1$ modes of data collection. We add an Mth "mode" corresponding to nonresponse, i.e. the category that represents no response to the $M-1$ modes. Let a random sample of size n be known before data collection starts, and let r_j denote the observed number of respondents in the jth mode, where $j \in \{1, \ldots, M\}$. Consider a multinomial distribution in M modes with response propensity ρ_j for the jth mode, where $\rho_j \in [0,1]$. ρ_M is the nonresponse propensity; however, it is no longer explicitly modeled later.

Vector $\mathbf{r} = (r_1, \ldots, r_M)$ follows a multinomial distribution with sample size n and response propensity $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)$, i.e., the joint distribution of $\mathbf{r}$ is a multivariate generalization of a binomial distribution,

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) = \frac{n!}{\prod_{j=1}^{M} r_j!} \prod_{j=1}^{M} \rho_j^{r_j}. \tag{2.1}$$

Linderman, Johnson and Adams (2015) used a stick-breaking transformation to reformulate the multinomial distribution as a product of binomial distributions, where the constructed parameters are dependent. This offered a chance to rewrite the m-dimensional (2.1) recursively in terms of $M-1$ binomials. In the stick-breaking representation, the propensity vector $\boldsymbol{\rho}$ serves as a stick that is recursively split into two pieces to create binomial variable $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_{M-1})$. To provide a derivation, let the jth mode response variable r_j follow a binomial density with parameters n_j and $\tilde{\rho}_j$, i.e., $\text{bin}(r_j \mid n_j, \tilde{\rho}_j)$, where n_j and $\tilde{\rho}_j$ represent the remaining size of the sample and the fraction of the remaining probability approached by the jth mode,

$$n_j = n - \sum_{k<j} r_k, \tag{2.2}$$

$$\tilde{\rho}_j = \frac{\rho_j}{1 - \sum_{k<j} \rho_k}, \tag{2.3}$$

where $j \in \{2, \ldots, M\}$. When $j = 1$, parameter $n = n_1 = \sum_{j \in \{1, \ldots, M\}} r_j$ and parameter $\tilde{\rho}_j = \rho_1$. Using (2.3), we have

$$\tilde{\rho}_j^{r_j}\,(1-\tilde{\rho}_j)^{n_j-r_j} = \rho_j^{r_j}\;\frac{1}{\left(1-\sum_{k<j}\rho_k\right)^{n_j}}\left(1-\sum_{k\le j}\rho_k\right)^{n_j-r_j}. \tag{2.4}$$

Note that r_j sums to n over j and $n_j = n_{j-1} - r_{j-1}$ for any $j \in \{2, \ldots, M\}$. This means that pairs of terms $\left(1-\sum_{k<j}\rho_k\right)^{n_j}$ will cancel in the product of (2.4) over j leading to the same format as the multinomial exponential term in (2.1),

$$\prod_{j=1}^{M-1}\tilde{\rho}_j^{r_j}\,(1-\tilde{\rho}_j)^{n_j-r_j} = \prod_{j=1}^{M}\rho_j^{r_j}. \tag{2.5}$$

The normalization constants follow the same reasoning. Combined with exponential terms in (2.5), (2.1) can be rewritten as

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) = \prod_{j=1}^{M-1}\text{bin}(r_j \mid n_j, \tilde{\rho}_j). \tag{2.6}$$

We use the stick-breaking representation of the multinomial model for practical reasons: for this representation, there is a simple and efficient Gibbs sampler for the multinomial (multilevel) model. As explained in Linderman et al. (2015), it uses the same Polya-Gamma data augmentation method (Polson, Scott and Windle, 2013) that was used for the binomial models in Wu et al. (2023), and the stick-breaking representation allows sampling the model coefficients for all $M-1$ modeled categories in a block, thereby improving the convergence of the Gibbs sampler. This representation also has a drawback: the definition of $\tilde{\rho}_j$ makes interpretation of the underlying model coefficients more difficult, particularly the interpretation of correlation coefficients in the models detailed in Section 2.2.

Section 2.2 employs a structural time-series model to decompose an observed time series into some underlying time-related components.

2.2 Multinomial multilevel time series model

To measure the dependence of response propensities among the modes, we extend the models of Wu et al. (2023) by introducing a new hierarchical parameter indicative of correlation coefficients. Such dependence spread over time-series components of interest is similar to those adopted by Wu et al. (2023); it is suggested to revisit that paper for more details on each component's definition and for technical details.

To describe each model component at the most detailed level, let the dependent propensity parameter vector of the sequential modes be associated with a specific ASD stratum and time point, i.e., $\tilde{\boldsymbol{\rho}}_{g,t} = \{\tilde{\rho}_{g,t,j} \mid j \in \{1, \ldots, M-1\}\}$, where the jth entry denotes the propensity parameter of the jth mode in ASD stratum g at time t, as defined in (2.3). The numbers of ASD strata, time points and survey modes are G, T, and $M-1$, respectively. Note that in this paper, we use the term ASD stratum to indicate a population group that may receive a different treatment. This is not to be confused with sample strata that receive different inclusion probabilities. We omit the reference to ASD in ASD stratum in most of the following.

In this paper, we assume that the choice of ASD strata has already been made and variable selection itself is not part of the modeling strategy. Strata will be based on auxiliary information that is predictive of the survey variables of interest and/or survey nonresponse. However, when the amount of auxiliary information is large, a variable selection strategy should be integrated into the model fitting and ASD optimization strategy. We return to this issue in the discussion.

We let $\theta_{g,t,j} = \text{logit}(\tilde{\rho}_{g,t,j})$. A logit link function is used to convert the constrained scale of a probability to the unconstrained scale of a linear predictor, i.e. a linear combination of model components. We denote the linear predictor by $\theta_{g,t,j} = \text{logit}(\tilde{\rho}_{g,t,j})$. Therefore, the multinomial likelihood function in (2.1) can be rewritten by substituting the inverse transformation for $\tilde{\rho}_{g,t,j}$,

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) \propto \prod_{j=1}^{M-1} \left(\frac{e^{\theta_{g,t,j}}}{1+e^{\theta_{g,t,j}}} \right)^{r_j}. \tag{2.7}$$

The multilevel models considered for modeling the linear predictor $\theta_{g,t,j}$ take the general form of additive decomposition, which refers to a function of the sum of time-series components. Thus,

$$\theta_{g,t,j} = \beta_j + \beta'_{xj} x_g + \delta'_s s_t + u_{t,j} + v_{g,j} + z_{g,t,j} + e_{g,t,j}, \tag{2.8}$$

where the first three and the last four terms are modeled as fixed effects and random effects, respectively.

The first regression fixed effects β_j are mode-specific intercepts, measuring the main effect on $\theta_{g,t,j}$. The second fixed effects β_{xj} are mode-specific regression coefficients associated with p-vector covariate x_g, corresponding to specific demographic characteristics associated with stratum g. The third fixed effects δ_s are season-specific regression coefficients associated with the q-vector season indicator variable s_t. See the definition of strata and seasons in Appendix A. Currently, all strata share common seasonal and mode effects. In a broader sense, these fixed effects can be stratum-specific. In the present application throughout this paper, x_g and s_t are binary vectors corresponding to categorical variables, but also ordinal or numerical and even time-varying variables can be taken into account, if needed.

Each random effect term in (2.8) implicitly allows for correlation between survey modes. Refer to Wu et al. (2023) for a description of the random effect components. As stressed, these terms are now crossed with the mode, i.e., separate variance parameters for each mode and correlation parameters among the modes are introduced. The global time trend $\mathbf{u}$, random intercept for strata $\mathbf{v}_g$, and stratum-specific trend $\mathbf{z}_g$ conform to this rule. White noise random effects $e_{g,t,j}$ are also crossed with mode, but we use a single common variance parameter for all modes and no correlation.

We adopt a Bayesian approach to estimate the model in (2.8) and to obtain reliable predictions of the response propensities at the mode, stratum and time levels. As noted, the priors are the same for coefficients corresponding to different modes. For notational convenience, we suppress subscripts g, t and j in each model component term. Fixed effects β and δ are assigned weakly informative priors normally distributed with zero mean and diagonal variance matrix, where the standard error takes a relatively large value of 10.

Each random effects vector is assumed to follow a multivariate normal prior with mean $\mathbf{0}$ and covariance matrix defined as the Kronecker product of two covariance matrices $\mathbf{A}$ and $\mathbf{V}$. Here, $\mathbf{V}$ is a fully parameterized covariance matrix, which is assigned a scaled-inverse Wishart prior (Gelman and Hill, 2007; O'Malley and Zaslavsky, 2008), and $\mathbf{A}$ is a fixed matrix, which may be a simple diagonal matrix for unstructured effects or a structured matrix corresponding to random walks over time. More technical details about the prior specification and the estimation strategy, including the Gibbs sampler and the required full conditional distributions, can be found in Boonstra and van den Brakel (2019) and Wu et al. (2023).

A more parsimonious model can be obtained by omitting the mode-oriented interaction and replacing the fully parameterized covariance matrix $\mathbf{V}$ by a diagonal matrix, if there is little interest in the between-mode effects on propensity predictions. This model is called the no-correlation model and it provides a base model against which to evaluate the performance of the full correlation model.

Model estimation is carried out using R package *mcmcsae* (Boonstra, 2022). Convergence of the Markov Chain Monte Carlo (MCMC) simulation results is assessed using trace plots and the potential scale reduction factor or R-hat diagnostic (Gelman and Rubin, 1992). These diagnostics show that the Gibbs sampler is converging fast, both for models including and excluding correlations between modes. To a large extent this is due to sampling all fixed and random effects in a single block, which is possible by virtue of the stick-breaking representation of the multinomial distribution in combination with Polya-Gamma data augmentation, as mentioned previously.

2.3 Extensions to more general complex sampling designs

The model laid out in Section 2.2 supports stratified random sampling survey designs, but no other complex sampling features. For the GEZO application that we focus on this is justifiable. Even though the GEZO uses a two-stage design with municipalities being the first-stage sampling units, the resulting clustering effects are very minor because most municipalities are selected. Furthermore, first and second stage inclusion probabilities are such that the overall inclusion probabilities are equal for all persons. The stratification used in this paper is chosen such that the response probabilities are reasonably homogeneous within strata.

We now briefly describe how our method can be extended to support surveys with more general complex sampling designs. If sampling probabilities are unequal, one may in some cases still define a stratification such that the sampling probabilities are (approximately) equal within strata. If this is not possible, an analysis based on a unit-level model would be more appropriate. The multinomial model at the stratum-level then becomes a categorical or Multinoulli model at the person-level, i.e. the special case of the multinomial distribution with $n_i = 1$ for each person i. Note that the derivations regarding the stick-breaking representation in Section 2.2 are still valid in this case. Essentially, stratum subscripts g in (2.8) would become person subscripts i.

Unit-level covariates that explain variation in both sampling and response probabilities should be included in the model to mitigate bias as far as possible. In particular, the inclusion probabilities themselves or the underlying variables that they depend on are valuable covariates. It can also help to model the dependence on inclusion probabilities in a flexible way, see e.g. Chen, Elliott and Little (2010). A stratification can still be defined specifically for the purpose of ASD such that the response probabilities are relatively homogeneous within strata. Such a stratification does not need to coincide with a possible stratification used for sampling. The latter can be handled in a unit-level model by including the stratum indicators in the model. Similarly, for a clustered design the unit-level model can account for cluster effects by including cluster indicators. However, since typically only a subsample of clusters is included in the sample, and the number of observations per observed cluster may be small, the corresponding cluster coefficients should be modelled as random effects (Scott and Smith, 1969). Finally, we note that the described model extensions to support more complex sampling designs can be handled using the same Gibbs sampler framework that we use for the GEZO application. In particular, the same data augmentation approach can be used for the categorical/Multinoulli family, as it is a special case of the multinomial distribution. The only difference is that one may need to incorporate more covariates with corresponding fixed effects, as well as additional random effects for clusters, possibly at multiple stages. Such unit-level models can be fit in the same way e.g. using R package *mcmcsae*, although computation times will increase due to the larger unit-level data size and model size.

3. Optimizing mode allocation under the Bayesian multilevel time series model

This section explores an allocation problem accounting for such uncertainty to grasp the timeliness and implementation of ASDs. Section 3.1 outlines the main ingredients for the construction and operation of this problem in a Bayesian framework. A strategy is proposed in Section 3.2 to assess the gain of adaptive allocations against nonadaptive allocations concerning nonresponse bias risk reduction by monitoring a measure of bias risk.

3.1 Main ingredients

Generally, mathematical optimization involves the selection of the "best available" values of some objective function relative to a number of constraints by choosing input values from an allowed set. Establishing optimization models entails three major elements: decision variables to optimize the goal, objectives to be minimized or maximized, and constraints on the decision variables. Because of optimization on the Bayesian setting, we emphasize that all mentioned statistical parameters are considered to be random variables with values that change over time. Consequently, objective functions and constraint functions are

also random variables. In the following, the main ingredients are first introduced for a non-Bayesian setting and are then developed for the Bayesian setting.

Decision variables are symbolic representations of an intervention decided by the decision maker. They represent unknown parts of an objective function that can be manipulated and may take on any possible value within an allowed set if specified. In this paper, an intervention is supposed to allocate interview modes to strata when preceding modes fail to obtain their data. Therefore, decision variables refer to allocation probabilities that indicate how likely nonrespondents are to be approached via a follow-up mode. Allocation probability $s_{g,t} \in [0,1]$ makes a decision on the size of follow-up candidates in stratum g at time point t, where $s_{g,t} = 0$ implies that data collection for stratum g stops, and $s_{g,t} = 1$ means that all stratum g nonrespondents in the preceding modes are allocated to the upcoming mode.

The objective function defines the criterion to evaluate candidate values of the decision variables. Apart from the decision variables, it depends only on observed and estimated quantities. Our optimization goal is to minimize the expected risk of nonresponse bias via optimal allocation. Since nonresponse cannot be observed directly, this paper considers a proxy indicator of nonresponse bias that is a function of response propensities. We employ the CV of response propensities (Schouten et al., 2009). The true population CV bounds the absolute standardized bias of respondent means. However, the CV estimated on a specified set of auxiliary variables will observe only a piece of that overall bias. While there are alternative indicators, see Moore, Durrant and Smith (2018) and Nishimura, Wagner and Elliott (2016), it is, most of all, the available auxiliary information that plays a decisive role. In multi-purpose surveys, availability may be less of an issue as one may focus on general representation and any improvement will be useful. In surveys with only a few key statistics, availability of relevant auxiliary variables is key.

The CV is the weighted standard deviation divided by the weighted response rate

$$\mathrm{CV}(s,t) = \frac{\sqrt{\sum_g d_{g,t}\left(\rho_{g,t} - \bar{\rho}_t\right)^2}}{\bar{\rho}_t}. \tag{3.1}$$

Weight $d_{g,t}$ is the sample proportion of the stratum g size at time t against the overall size at time t, that is, $d_{g,t} = n_{g,t} \big/ \sum_g n_{g,t}$. This notation implicitly assumes that the sampling design leads to equal inclusion weights, but if not, the design weights should also be incorporated. This addition is straightforward but makes the notation intractable. Mixed response propensity $\rho_{g,t}$ denotes the overall propensity over modes, which is the sum of the marginal response propensity of the starting mode and the joint response propensities of mode $j \geq 2$ supposing that stratum g did not respond to the last $j-1$ modes, (Here, we implicitly assume that all nonrespondents in a mode are eligible for follow-up. In practice, some types of nonresponse, such as due to physical or mental illness, may not be eligible.)

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j \in \{2,\ldots,m-1\}} \rho_{g,t,j} \prod_{i \leq j-1} (1 - \rho_{g,t,k}). \tag{3.2}$$

The individual (conditional) propensities $\rho_{g,t,j}$ for any mode j are estimated by multinomial models in Section 2. (3.2) assumes that all nonrespondents to the preceding modes will be recruited by mode j for a nonadaptive survey; however, this can be modified to an adaptive survey by reducing joint propensity to decision variable $s_{g,t,j} \in [0,1]$, so the updated equation becomes

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j \in \{2,..,m-1\}} s_{g,t,j}\, \rho_{g,t,j} \prod_{i \le j-1} (1 - \rho_{g,t,k}). \tag{3.3}$$

Clearly, (3.3) is equivalent to (3.2) when all $s_{g,t,j} = 1$.

The denominators of (3.1), called the weighted response rates over strata, indicate the estimated level of unknown propensities, which are defined as the weighted sum of mixed propensities of (3.2) and (3.3). We call CV nonadaptive when all $s_{g,t,j} = 1$ and adaptive when at least one $s_{g,t,j} \neq 1$.

We use one single indicator motivated from multi-purpose surveys, i.e. having a large and diverse range of target survey variables. However, in surveys with one or a few target survey variables, more focused indicators may be employed. A good example is indicator $H1$ of Särndal and Lundström (2010). Doing so, would change the use of historic survey data and also include associations to the target survey variable(s).

Constraints are functional inequalities or equations that represent logical restrictions on what values of decision variables are allowed. For example, constraints might ensure a thorough search of feasible solutions from a finite solution space. In the survey design context, a constraint can be a limit placed either on the survey quality, such as solutions making the overall response rate greater than 0.5, or on the survey cost, such as the overall cost of interviewers reaching nonrespondents being lower than a specified amount. In this paper, we focus on cost constraints regarding the workload of approaching nonresponse candidates by means of a follow-up mode. In service level agreements with survey sponsors a maximal chance of a budget overrun is often specified, say at 10%. This budget overrun proportion is denoted by α. If the budget is strictly limited, then $\alpha = 0$. Let the specified budget level be h. Now the cost constraint is

$$p\left(\sum_g s_{g,t}\,(n_{g,t} - r_{g,t}) \ge \sum_g h(n_{g,t} - r_{g,t})\right) \le \alpha, \tag{3.4}$$

where p is the probability of the adaptive workload exceeding the non-adaptive workload. When the values of $s_{g,t}$ satisfy constraint (3.4), the corresponding solution of the decision variable is called acceptable; otherwise, the solution will contradict the rule. It is natural to specify lower and upper bounds on decision variable $\mathbf{s} = \{s_{g,t} \mid \forall g, t\}$, which are referred to as box constraints,

$$0 \le \mathbf{s} \le 1. \tag{3.5}$$

Therefore, the optimization problem in a sample allocation application for the non-Bayesian setting is formulated to detect a vector $\mathbf{s}$ that minimizes objective (3.1) subject to constraints (3.4) and (3.5) given parameters (n, r). As stated above, (3.1) and the workload in (3.4) are random variables in the Bayesian approach, so we take expectations of the posterior distributions.

Given that no explicit expression exists for the posterior distributions, we estimate them empirically. We use the Gibbs sampler replicates of the posterior distributions and per replicate compute the CV and the workloads. We obtain

$$\hat{E}\left(\mathrm{CV}(s,t)\right) = \frac{1}{K}\sum_k \frac{\sqrt{\sum_g d_{g,t}\left(\rho_{g,t}^{(k)} - \bar{\rho}_t^{(k)}\right)^2}}{\bar{\rho}_t^{(k)}}, \tag{3.6}$$

where $\hat{E}\left(\mathrm{CV}(s,t)\right)$ is the estimated posterior expectation at time t, $\rho_{g,t}^{(k)}$ is the kth iterated estimate from the posterior predictive function of $\rho_{g,t}$, and subscript k runs over MCMC draws. The probability of a budget overrun is estimated empirically by the frequency that the number of times the adaptive design required budget exceeds the specified budget

$$\frac{\sum_k \mathbf{1}_{\sum_g s_{g,t}\left(n_{g,t}-r_{g,t}^{(k)}\right) \geq \sum_g h\left(n_{g,t}-r_{g,t}^{(k)}\right)}}{K} \leq \alpha. \tag{3.7}$$

$\mathbf{1}_{\sum_g s_{g,t}(n_{g,t}-r_{g,t}^{(k)}) \geq \sum_g h(n_{g,t}-r_{g,t}^{(k)})}$ is an indicator function that takes a value of one when the inequality in its subscript is met for the kth iteration and is zero otherwise. Therefore, Bayesian optimization aims to minimize objective (3.6) subject to constraints (3.5) and (3.7).

Benchmark: In the Bayesian optimization problem, we set a benchmark to evaluate ASD performance from two viewpoints: improving quality and saving money. Specifically, promoting sample representativeness by recruitment can improve data collection quality, while distributing cost-intensive resources to where they are most needed can save money. This goal can be achieved, for example, by switching from a single mode to mixed but optimally reallocated modes or switching from full mixed modes to partial mixed modes. By letting decision variables $\mathbf{s}=0$ or $\mathbf{s}=1$, the optimization problem proposed above can settle those reallocations. To do so, the performance of the single-mode design and the full mixed-mode design are standards of and compared with the ASD performance.

3.2 Static ASD optimization

In analogy to adaptive treatment regimes, we call ASDs *static* when they are based solely on information available in registry and frame data before the start of data collection, and *dynamic* when they are based (also) on paradata (data collected during data collection). *Dynamic* ASDs reflect the dynamic nature of the optimization since optimization is performed at each data collection phase, i.e., after each mode is completed.

For *dynamic* ASDs in the current context, decisions on assigning interviewers to strata are made dependent on intermediate survey results from the preceding modes. The correlation between the response propensities for the non-interviewer mode and the response propensities for the interviewer mode leads to an intermediate update of the prior distributions for the latter. Theoretically, the evaluation can identify the priorities of refusers in strata to be interviewed and inform the interviewer workload. In reality, there may

be insufficient time to compute reallocated interviewers' workload in time because of geographical clustering. Additionally, reallocation requires complex logistics in case management; we leave this point to the discussion section.

Therefore, in this paper, we focus on the static ASD setting, i.e. the decision to perform a follow-up is set at the start of data collection. In this paper, we restrict ourselves to one such intervention, but the methodology would allow for multiple interventions, say after various numbers of calls or visits.

In Section 3.3, we construct the strategy to account for uncertainty in making decisions and to specify the optimization routine to determine the optimal allocations for the Bayesian optimization problem in Section 3.1.

3.3 The optimization strategy

To solve the formulated optimization problem in Section 3.1, we propose a two-step strategy at time t,

1. *Construct the posterior distribution of the response numbers* $r_{g,t}$. Let historic time series data sets up to time $t-1$ be data used for model training, and let data sets at time t be test data for prediction. All model coefficients and hyperparameters in (2.8) can be estimated by the size of a sample $\mathbf{n}_{1:t-1} = \{\mathbf{n}_{g,1:t-1} \mid \forall g\}$ and the response numbers in all modes $\mathbf{r}_{1:t-1} = \{\mathbf{r}_{g,1:t-1,j} \mid \forall g, j\}$. Under the estimated model, predictions can be obtained on dependent propensities $\tilde{\boldsymbol{\rho}}_t = \{\tilde{\rho}_{g,t,j} \mid \forall g, j\}$ and $\mathbf{r}_t = \{\mathbf{r}_{g,tj} \mid \forall g, j\}$, given data $\mathbf{n}_t = \{n_{g,t} \mid \forall g\}$.
2. *Determine optimal allocations*. Specify budget level h and overrun level α. Set multiple starting vectors of stratum allocations $\mathbf{s}$, each vector viewed as an initial state and each having a finite number of well-defined successive states. For any stratum g, assume K iterations of estimates of $\mathbf{r}_{g,t} = \{r_{g,t,m} \mid \forall j\}$ and $\tilde{\boldsymbol{\rho}}_{g,t} = \left\{\tilde{\rho}_{g,t,j} \mid \forall j\right\}$ generated from the posteriors in 1. These posterior estimates and given parameters h and α are separately substituted into (3.6) and (3.7) to compute the posterior expectation $\hat{E}\left(\mathrm{CV}(s,t)\right)$ and the posterior probability of workload excess. To detect the optima, starting from each initial state, such a computation proceeds through its successive states, produces output, and eventually terminates at the final state. Discard constraint-violated states and their output, and preserve constraint-met states and their output. Within these results, sum the minimum of $\hat{E}\left(\mathrm{CV}(s,t)\right)$ and its corresponding allocations optima.

Solving this mathematical program is a computationally intensive task. Therefore, the methods in step 1 are implemented in R using the *mcmcsae* package (Boonstra, 2022), while the methods in step 2 are implemented in R using the *auglag* (Augmented Lagrangian Minimization Algorithm) function of the *Alabama* package (Varadhan, 2022) for constrained nonlinear optimization.

3.4 Performance evaluation

This section introduces an evaluation criterion to assess the prediction accuracy. The criterion can shed light on the gain in nonresponse risk reduction from different models or from different survey designs. This

gain is quantified by the root mean square error (RMSE) of the posterior distribution of a parameter τ, e.g., response propensity or CV, relative to the "true", with the latter estimated via observations.

We consider performance over rolling windows of three months. This choice is motivated by the three-month fieldwork duration of the application in this paper, but can be changed to any length. In time window $q = \{t, t+1, t+2 \mid \forall t\}$, the RMSE of the gth stratum is then defined as

$$\text{RMSE}(\tau, q) = \sqrt{B^2(\tau, q) + \text{SD}^2(\tau q)}, \tag{3.8}$$

where the first term is called the bias term, represented as the quadratic difference between the posterior mean of parameter τ and the observed τ,

$$B^2(q) = \sum_g d_{g,q} \left(E_{\pi_q} \text{CV}(s_q, q) - \widehat{\text{CV}}(s_q, q) \right)^2 \tag{3.9}$$

and the second term is the posterior variance of CV, which is a quadratic form of the standard deviation (SD),

$$\text{SD}^2(q) = \sum_g d_{g,q} \, \text{Var}_{\pi_q} \text{CV}(s_q, q). \tag{3.10}$$

Weight $d_{g,q} = n_{g,q} \Big/ \sum_g n_{g,q}$ is the ratio of the stratum g size to the sample size in window q. The posterior distributions π_q of CV and allocation $\mathbf{s}_q$ are derived from the computing strategy in Section 3.3.

These criteria depend strongly on sample size sampling variation, especially for surveys with small sample sizes. Empirical data subject to sampling variation are used to evaluate the performance. While surveys with large sample sizes provide rich information and thus their performance can be evaluated precisely, for small surveys, a contradiction to time change becomes acute, i.e., they take longer to make a precise evaluation. Noisy criteria performance makes it more difficult to draw a sound conclusion about putting the adaptation into practice.

4. The Dutch Health Survey case study

This section explores and exploits the application of multinomial time-series models in Section 2 and the optimization approach in a Bayesian framework in Section 3 to the Dutch Health Survey (GEZO for short). Section 4.1 briefly introduces the background of GEZO. We illustrate how time changes in sequential propensities can be modeled, how the performance of optimal allocations depends on the budget level, and how optimal decisions depend on the length of applicable the historic data separately in Sections 4.2-4.4.

4.1 The Dutch Health Survey

The GEZO survey is conducted annually by Statistics Netherlands, providing a thorough overview of developments in medical contacts, lifestyle, health, and preventative behavior of the Dutch population,

including all individuals living in private households. A self-weighting two-stage sampling design is employed, which first draws a sample from municipalities and next from persons who live in the selected municipalities. The survey changed to a mixed-mode design after 2014. The observation method involves online and face-to-face interviews. First, CAWI is used to request the participation of sample units from the population. Next, nonrespondents are recruited to participate in a CAPI. As of 2018, however, adaptation is implemented to stabilize the interviewers' workload. Only a portion of CAWI nonrespondents is reapproached for CAPI to reduce survey costs and improve the representativeness. Higher response rates in CAWI sample units lead to a smaller chance they are reapproached.

In this paper, we focus on a time series of data collected from 2014 to 2017, involving 48 months. Note that data collected early in 2017 were "abnormal" because of technical issues with the web server, resulting in an interruption in data collection. This comes with practical reasons: Statistics Netherlands has implemented static ASDs since 2018, and the adaptation may waste the potential value of historic data used to improve prediction accuracy (Wu et al., 2023). Additionally, sample units are stratified into 13 disjoint strata by two auxiliary variables from the administrative frame or registers: age and ethnicity. See the stratification in Appendix A. Note that this stratification is fixed throughout this paper, and our time-series strata are different from the ASD strata (Van Berkel et al., 2020). The set of available auxiliary variables prior to the start of fieldwork is much larger. It includes several demographics such as gender, country of birth, household composition, socio-economics such as registered personal and household income, educational level, type of occupation, dwelling-related and area-related characteristics such as type of dwelling, house value, urbanization. Research into efficient and effective selection of strata is an important next step.

4.2 How can a time-series model be constructed under a sequential mixed-mode design?

This section elaborates the approach to build multilevel time-series models. We observe two potentially influential decisions. The first decision is to include a seasonal component. Adding a season will likely improve accuracy, but comes at a cost. Including a season implies that a longer series of historic data is needed; we must observe at least two years of data to evaluate a seasonal component, but preferably more. The second decision is the inclusion of explicit associations/correlations between parameters in the model. Again accuracy is likely higher, but more data are needed. There are three levels at which we can further differentiate the decisions: stratum, time and mode. We can thus make seasonal component parameters and correlations dependent on mode and stratum, and even time. We explore four scenarios: seasonal component present or absent times correlation present or absent.

Our objective is to find the most favorable of these four scenarios. Since there are many possible models, in analogy to Wu et al. (2023), we adopt a stepwise strategy to go from simple models to more advanced model. In evaluating performance, we consider two information criteria, Deviance Information Criterion (DIC, see Spiegelhalter, Best, Carlin and van der Linde, 2002) and Watanabe-Akaike Information Criterion

(WAIC, see Watanabe, 2010 and 2013). By providing a reasonable trade-off between model fit and model complexity, these measures are expected to select appropriate models for the prediction task of interest. They can be interpreted as approximations to leave-one-out cross-validation measures, and are relatively easy and cheap to compute from the MCMC simulation output (Vehtari, Gelman and Gabry, 2017). In building the models, we can choose from various random effects: white noise, a global time trend, random intercepts for each stratum and stratum-specific time trends. In addition, we can do this with or without season and have mode-independent or mode-dependent model parameters. We employ the following steps:

1. Baseline model: Set up two baseline models, one without seasonal components and one with seasonal components. The two models have fixed mode effects and fixed effects of mode-specific auxiliary variables β. The difference between the two models is whether seasonality δ is included.
2. Basic trend: Add a single random effect, i.e., {global time trend u_t}, {random intercepts for strata v_g}, or {white noise e_{gt}}, to the models in 1. Each random effect is correlated with modes or made independent of modes. Examine whether the no-season or season-inclusive model in 1 is enhanced by one of the three random effects.
3. Some stratum-dependence: Add a combination of two random effects, i.e., {global time trend u_t, random intercepts for strata v_g} or {random intercepts for strata v_g, stratum-specific time trend z_{gt}}, to the models in 1. Each random effect is correlated with modes or independent of modes. Examine whether the updated models outperform the models in 2.
4. Moderate stratum-dependence: Add a combination of three random effects, i.e., $\{u_t, v_g, e_{gt}\}$, or $\{u_t, v_g, z_{gt}\}$, to the models in 1. Each random effect is correlated with modes or independent of modes. Examine whether the updated models outperform the models in 3.
5. Full stratum-dependence: Add all random effects to the models in 1. Each random effect is correlated with modes or independent of modes. Examine whether the complete combination makes the model performance best.

Table 4.1 presents the results of the five steps. As seen in each row of Table 4.1, the models with and without seasonality are evenly matched at fitting and predicting. The information criterion (IC) results of the two baseline models (Model 1) show that the with-season model performs slightly better than the no-season model. This advantage continues with the addition of some random effects (see Models 1, 3 and 6), as the inclusion of seasonality δ yields lower ICs. On the other hand, the with-season models have slightly worse performance as can be seen in Models 2, 4, 7, 8, and 9. Notably, the results of Model 5 show mixed results. The DIC results favor modeling seasonal effects in accurate propensity predictions, but WAIC cannot conform to this.

Concerning the balance between model complexity and model fitness, the mode-independent models perform barely as well as the mode-correlated models, even though they slightly outperform (Model 5 for the no-season model and M6).

Table 4.1
ICs and effective number when evaluating the model fit and complexity.

Model	Fixed effect	Random effect	DIC		p_{DIC}		WAIC		p_{WAIC}	
			IND	COR	IND	COR	IND	COR	IND	COR
1	β	-	6,808		18		6,817		28	
	β, δ		6,800		25		6,813		38	
2	β	u_t	6,504	6,504	67	67	6,509	6,509	72	72
	β, δ		6,506	6,505	70	70	6,510	6,510	74	74
3	β	v_g	6,775	6,774	26	26	6,786	6,785	36	36
	β, δ		6,768	6,767	32	32	6,781	6,780	46	45
4	β	e_{gt}	6,602	6,601	322	321	6,479	6,479	200	200
	β, δ		6,609	6,608	322	321	6,488	6,487	201	200
5	β	u_t, v_g	6,488	6,490	147	146	6,472	6,475	131	130
	β, δ		6,474	6,472	77	77	6,477	6,476	80	81
6	β, δ	v_g, z_{gt}	6,494	6,495	143	144	6,481	6,482	130	129
	β, δ		6,488	6,490	147	146	6,472	6,475	131	130
7	β	u_t, v_g, e_{gt}	6,449	6,448	195	194	6,398	6,396	143	142
	β, δ		6,453	6,452	191	191	6,404	6,402	142	142
8	β	u_t, v_g, z_{gt}	6,396	6,395	109	108	6,381	6,380	94	94
	β, δ		6,398	6,397	112	111	6,382	6,381	96	95
9	β	u_t, v_g, z_{gt}, e_{gt}	6,398	6,398	137	137	6,371	6,371	110	110
	β, δ		6,400	6,397	132	134	6,375	6,372	108	108

Note: The time series of 2014 to 2017 is fit to a mode-independent model ("IND") and a mode-correlated model ("COR"). Each model is simplified using the fixed-effect components only and then accommodating the correlation over time or between modes by taking several random effects into account.
DIC = Deviance Information Criterion;
WAIC = Widely Applicable Information Criterion.

With random effects considered, the mixed models become better because they cause a decrease in ICs, in contrast to the models including fixed effects only (Model 1). Comparing Model 2-4 to Model 1 entails either the no-season or with-season model is improved by introducing a single random effect, where global time trend u_t induces the greatest decrease in ICs, followed by white noise e_{gt} and a random intercept for strata. Such improvement persists in ICs when applying the combinations of two random effects, as indicated by the comparisons of Model 5 to Model 2 and Model 3, and Model 6 to Model 3. Apparently, Model 5 has the most significant decrease in ICs thus far. Models 7 and 8 show that the models can be enhanced further with the addition of white noise e_{gt} and stratum-specific time trend z_{gt} to Model 5, and Model 8 makes ICs decrease more than does Model 7. Including white noise e_{gt} is of value to improved performance, as it adds little in lowering the WAIC of Model 9 despite the scarce contribution made to DIC.

As Model 9 shows, the mode-correlated and mode-independent models (COR and IND columns, respectively) perform similarly in terms of ICs when seasonality is overlooked. however, for the with-season models, modeling correlations (the COR column) come first in IC scores relative to the IND column. However, it is difficult to conclude that the with-season model has an absolute advantage over the no-season model in terms of model fitness and complexity. To identify whether seasonal effects play a vital role in adaptive allocations, we consider both the no-season and with-season models (marked in red) in Section 4.4.

4.3 How sensitive is ASD performance to the specified budget level?

This research question is concerned with how, given a budget level, we adapt allocations for CAWI nonresponses across strata to lower the risk of nonresponse the most. It also raises the question of whether such a reduction can be sustained across different budget levels.

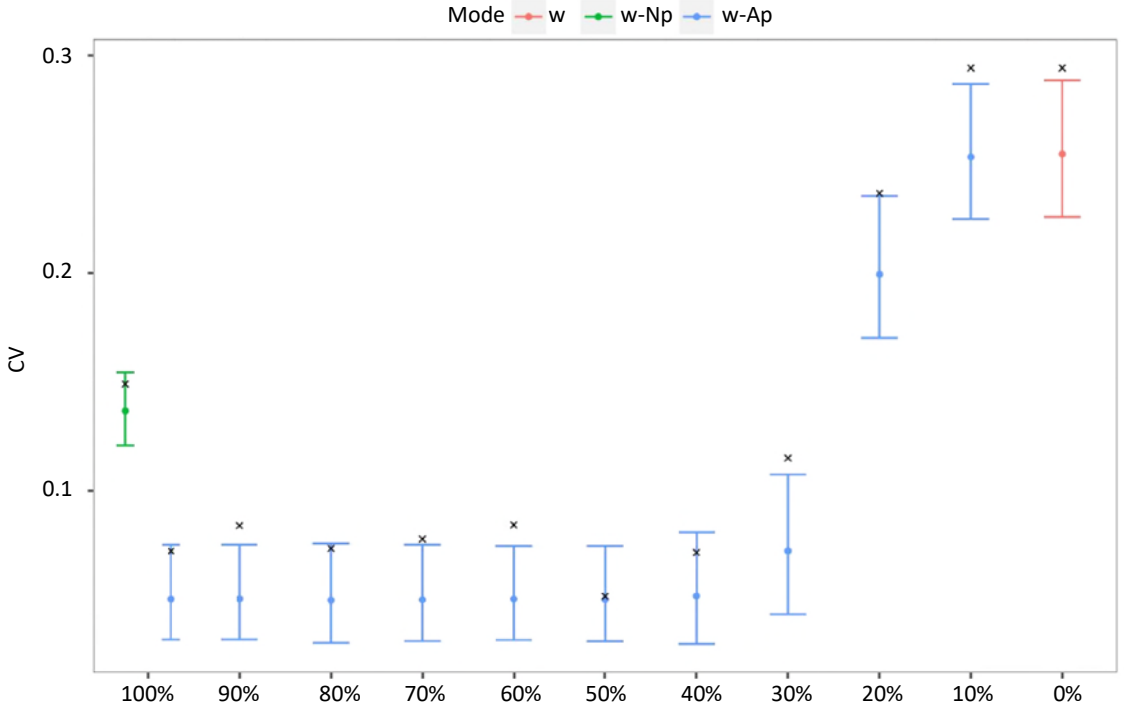
We answer this question by first minimizing (3.6) subject to (3.5) and (3.7) for the next data collection quarter when the budget level is specific, then by comparing the optimum (3.5) to the realized CV under the same budget level, and finally by comparing the optimum (3.5) under different budget levels. We focus on the next quarter because in the static case, the number of CAWI respondents is unknown until data are collected and because the sufficient sample of a quarter can ensure the prediction precision. Referring to the optimization strategy in Section 3.3, the evaluation procedure in quarter q is

1. Let budget level h begin at 100% and then successively decrease in steps of 10%, i.e., $h \in \{1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$.
2. Identify the forthcoming quarter q and set data in q as the test data set.
3. Set time-series data up to quarter $q-1$ as the training data set for estimating the selected models. The models are made of the components in Model 9 that are viewed as the "best" representation.
4. Use the sample size in q to simulate CAWI responses. For each stratum and each month within q, 3,000 draws are generated from posterior predictive distributions.
5. Based on the simulated model in step 3, individual posterior predictions of CAWI and conditional CAPI are generated separately 3,000 times for each stratum and each month in q.
6. Substitute the specified budget level h and the CAWI responses simulated in step 4 into cost constraint (3.7).
7. Compute mixed propensities by substituting level h and individual predictions in step 5 into (3.3).
8. Initialize three starting solutions of allocations probabilities, $s \in \{0, 0.5, 1\}$, each of which applies to 13 strata simultaneously.
9. Start from each initial point in step 8 to find the optimal solutions for each stratum by solver *auglag* based on steps 6 and 7.
10. Link the identified solutions to the actual sample for computing posterior CV predictions and CV realizations.
11. Conduct comparison by repeating steps 2-10 for each budget level in step 1.

$s=0$ indicates no CAPI follow-up, $s=0.5$ means half of CAWI nonresponses are assigned to CAPI, and $s=1$ represents full CAPI follow-up. To distinguish different mode strategies and ease notation, the CAWI-only, nonadaptive, and adaptive designs are denoted w, w-Ap, and w-Np throughout.

In Figure 4.1 the posterior CV predictions of 2017 Q1 are summarized for each budget level; see Table B.1 in Appendix B for the bias-adjusted CV results. We benchmark the w-Ap performance as a function of budget level h against the performance of w and w-Np. For brevity, CVs for the CAWI-only, nonadaptive, and adaptive are simplified to CV(w), CV(w-Np) and CV(w-Ap).

Figure 4.1 Comparison of coefficients of variation (CVs) of model-based response propensity predictions to bias-adjusted CV observations in 2017 Q1.



Note: The CV estimates are made separately for CAWI-only ("w"), nonadaptive ("w-Np") and adaptive ("w-Ap"). The posterior CV predictions are summarized by the 95% credible region, while the observations are marked by scatter points ("x").

Comparison of CV(w-Np) to CV(w) indicates that recruiting CAWI-nonresponses via CAPI can decrease nonresponse, as the 95% credible region (CI) of posterior CV(w-Np) is much narrower than that of CV(w), and the 97.5% quantile of posterior CV(w-Np) is far below the 2.5% quantile of CV(w). When $h = 100\%$, a further decrease in the overall variation can be achieved by the optimized allocations of the adaptive survey. Posterior predictions and observations of CV(w-Ap) deviate from 0.1 and move toward 0 relative to the 2.5% quantile of CV(w-Np), yet the broader CI for the adaptive approach indicates that prediction accuracy is compromised moderately. Because CIs scarcely alter when the budget is cut from 90% to 50%, the uncertainty reduction associated with CV(w-Ap) is unlikely to increase by more than 100%. This implies that in the interval of levels 100% to 50%, the low budget performs as well on the estimated nonresponse risk as does the high budget. The upper limits of CIs appear to approach and even run beyond the observed CVs; for instance, at the 50% level, the observation overlaps with the posterior mean.

The nonresponse risk rises with continued shrinkage of the budget since the estimates of CV(w-Ap) increase and point to an increased risk of nonresponse bias. For budget levels smaller than 50%, the allocation scheme identified puts more uncertainty on the posterior estimates of overall variation. In addition, the lower limits move toward and even far beyond 0.1 when the level is 20% or 10%, for which the solver ends up at a false local "optimum" due to the violated convergence criteria. For the 10% level, the allocation scheme is especially of less interest and loses its edge, as shown by exactly the same

CV(w-Ap) as CV(w). To determine which budget level is preferred most, we adopt a criterion, i.e., the relative cost defined as the overall cost of the adaptive size for CAPI relative to the nonadaptive size constrained by the budget level. See Table B.3 for the results of the relative cost under different levels in Appendix B.

Optimized reallocations make adaptation performance consistent across relatively large budget levels (100%-50%). Additionally, adaptation, although it loses precision slightly, wins nevertheless at the smaller estimated nonresponse risk compared to the w and w-Np designs (red and green error bars). Up to a 40% budget level, performance reverses and moves in the opposite direction, implying that the nonresponse risk grows sharply. This overall behavior is expected to be relatively robust to reasonable variations in model choice, as both budget extremes leave no freedom in ASD allocation to compensate for differences in response propensities.

4.4 How does the performance of adaptive designs depend on the available historic data?

This question is a matter of examining how the accumulating historic time series influences the adaptive design performance, that is, the nonresponse risk measured by CV and the bias-variance balance measured by RMSE. To answer this question, we explore the performance of w, w-Np, and w-Ap designs at the calendar quarter level. Additionally, we benchmark the adaptive performance against the performance of w and w-Np.

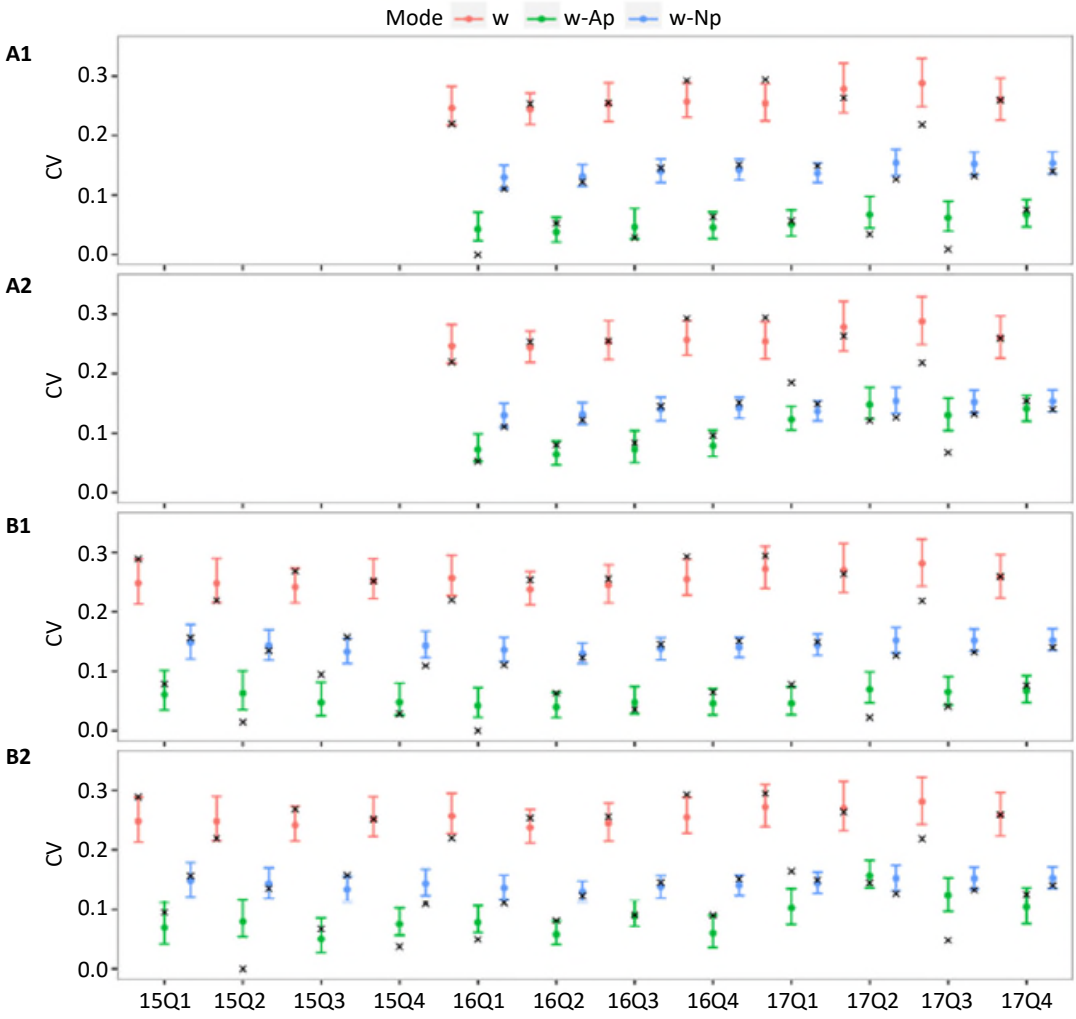
We compare and evaluate the models with/without the inclusion of seasonality and budget levels of 50% and 30%. Section 4.2 hints that seasonality is an ignorable factor since the models with and without this score have similar fitness and complexity. Section 4.3 implies that in a specific time window, budget level 50% promotes ASD performance most cost-effectively, and the ASD loses its absolute advantage for smaller values. The performance's sensitivity to the time-series length is less clear if the models consider seasonality and/or the budget level is less than 50%, so it is premature to skip them in the analysis. By crossing the two conditions, comparisons can be made simultaneously in the four scenarios of the models: (1) with the inclusion of seasonality and level 50%, (2) with the inclusion of seasonality and level 30%, (3) without the inclusion of seasonality and level 50%, and (4) without the inclusion of seasonality and level 30%.

To explore the sensitivity to the historic time-series length, the analysis is performed on a rolling basis by adding one month at a time. Recall that the initial historic time-series length should be at least one year for the models without the inclusion of seasonality (scenarios 3 and 4) but at least two years for the models with the inclusion of seasonality (scenarios 1 and 2). For each, the training process ends in 2017 Q3 because one quarter should be left for prediction.

In Figure 4.2, the uncertainty about the estimated CVs that is assessed by the 95% credible region, and the posterior means together, are compared to the CV observations over quarters and between different designs. In the w and w-Np designs, the observed CVs fall within the intervals or are very close to the

intervals' limits in most quarters, with the exception of CV(w) at 17 Q3. The ASD results in panels A1, A2, and B2 support this finding. Additionally, observations far outside of the CIs appear in 15 Q2 of Panel B2 and 17 Q1 of Panels A2 and B2. The exception implies that in corresponding quarters, it is less convinced of the evaluated adaptive performance duplicating the performance in practice.

Figure 4.2 Under a given budget level, the posterior coefficient of variation (CV) for the adaptive, nonadaptive, and CAWI-only against the observations over quarters.



Note: 95% credible regions with posterior expectations are summarized for CAWI-only (w), nonadaptive (w-p), and adaptive (w-Ap). Observations are denoted by the black cross "x" points. Panels "A" are plotted for with-season models, and panels "B" are plotted for no-season models. Panels "A1" and "B1" correspond to budget level 50%, while panels "A2" and "B2" correspond to budget level 30%. The quarter on the x-axis denotes the present quarter for prediction purposes.

As mentioned before in Section 4.3, the performance for adaptive designs under level 50% is consistently superior to the performance under level 30% across quarters, as shown by comparing A1 to A2 or B1 to B2. At 50%, the estimated CV(w-Ap) is more precise because of the narrower credible regions, which can be seen implicitly. Moreover, we can observe the absolute advantage of adaptive designs in outperforming

nonadaptive designs since the upper limits of CV(w-Ap) deviate substantially from the lower limits of CV(w-Np), but at 30%, they are competitive, for example, 17Q1 and 17Q2 (see panels A2 and B2).

As historic data accumulate, it is conjectured that the models can be optimized further, the CV prediction accuracy can show a consistent increase, and the resulting performance can be improved. Clearly, this is the case in panel B1 until 16Q4, but from that point onwards, there appears to be no room for improvement in performance, and the posterior estimates of CV (w-Ap) stabilizes at approximately 0.1. The results for panels A1 and B1 are similar, so that it must be concluded that modeling seasonality contributes little to prediction accuracy.

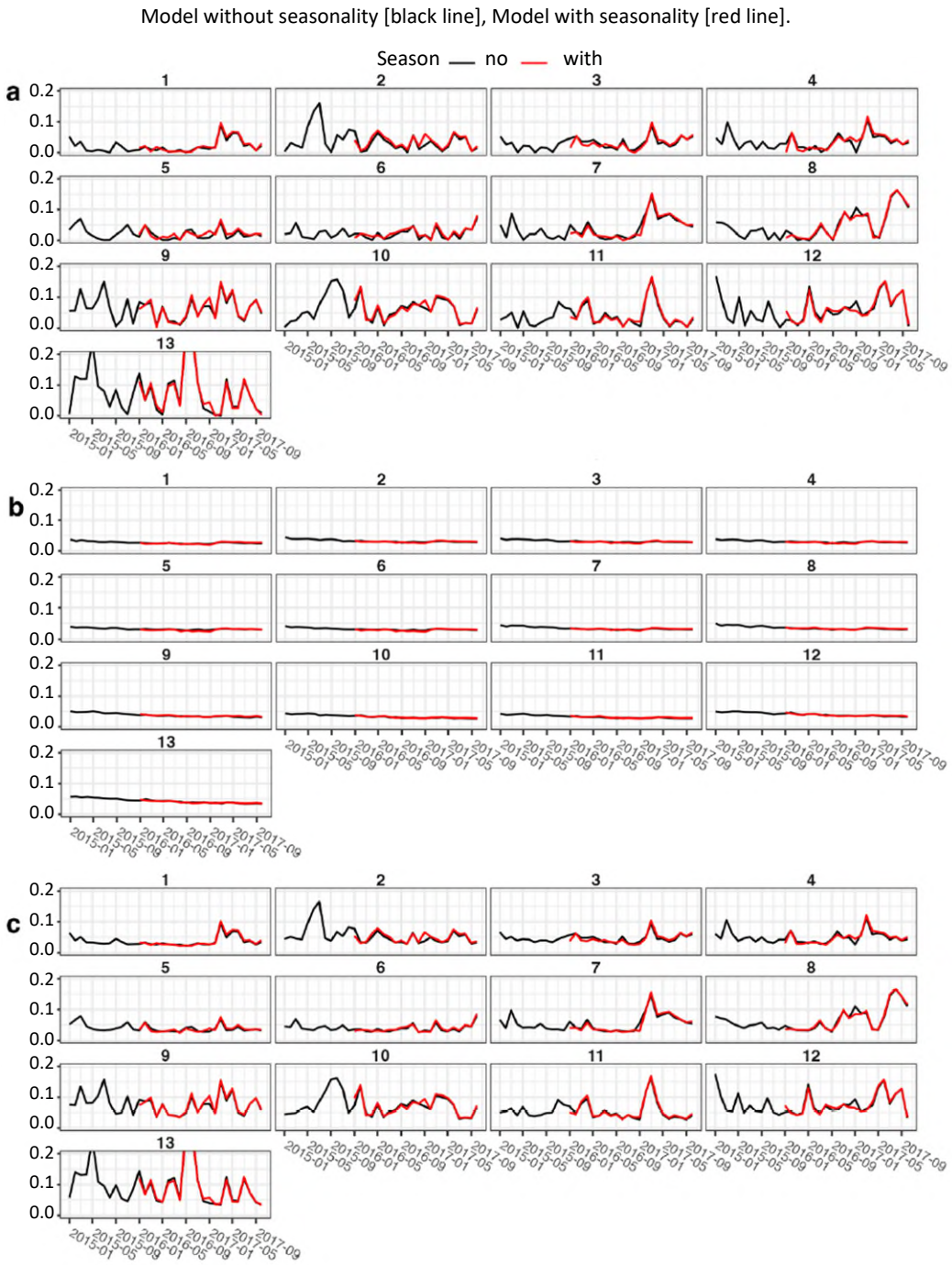
Jumping to conclusions on the ASD performance's robustness is dogmatic for two reasons. First, some strata may benefit more than others. Their individual CV estimates may be less biased and vary toward the seasonality-inclusive option despite little difference in the overall variation. Second, the sample sizes in some strata are quite small. In early data collection phases, they may have volatile behavior in the error-variation balance. The historic time-series length determined based on those results is not a guarantee of robustness.

Therefore, we evaluate individual strata performance measured according to the criteria in Section 3.4. As above, we apply the sliding time window approach moving forward on a time series. To illustrate, this is used in nonadaptive designs. With an application to ASDs, allocations must be reoptimized for the upcoming time window using the strategy in Section 3.3.

The time window slides as the width is increased to include the next upcoming new time period. In quarter q , we can evaluate the prediction performance for each stratum by substituting individual posterior response propensity estimates and individual realizations into (3.8)-(3.10), that is, $\mathrm{RMSE}(g,q)$, $B(g,q)$ and $\mathrm{SD}(g,q)$. Since this analysis is iterated on a rolling basis, a sufficiently long time series allows for thorough comprehension of how each stratum prediction performance changes with time.

Figure 4.3 shows that when comparing the no-season models (red curves) and with-season model (black curves), the introduction of seasonality is unlikely to be a trigger for an effective reduction in bias and variance. This is solidly true for almost all quarters, with the exception of the quarter involving months 2017-01 to 2017-03 in some strata (such as stratum 8) for the bias and RMSE estimates. As observed in panels a and b, the estimated variation in response propensity decreases smoothly overall, in sharp contrast to the estimated level of response propensity having volatile behavior. The volatility differs by strata. The estimated bias results of some strata (strata 1-8) fluctuate approximately 0.05 across quarters until the quarter starting in 2017-01. After that point, they experience a transient increase caused by the technical issue at that time (Wu et al., 2023 for more discussion and a possible remedy). When the training data are extended to include "normal" data, the biases can quickly decrease to 0.05. Note that stratum 8 acts in the opposite manner. In contrast, strata 9-13, which have relatively small sample sizes, obtain relatively more biased response propensity expectation estimates in most quarters.

Figure 4.3 One-step forward moving averages of estimated bias (panel a), standard deviation (panel b), and root mean square error (panel c) of the stratum level response propensity.



Note: The "black" curve represents to the no-season model, while the "red" curve refers to the with-season model. Both models include the correlations between CAWI and CAPI regarding propensity predictions. The horizontal axis represents the time point at which an ASD decision is made.

Ultimately, the analysis results suggest that when modeling a short time series, the seasonal effects, when they are assumed to be the same for different modes, can be less important to the improvement of ASD

performance. With more data available for training, the ASD performance can be consistently improved until a time point, implying that a stopping rule of data collection may be implemented and an effort-based strategy for strata of small sample sizes may be adopted.

5. Discussion

Given the survey budget, ASD seeks the optimal match between respondent behavior and design features, i.e., a set of decision rules, which can be determined through optimization approaches (see Schouten et al., 2017 for pros and cons of various approaches). Serving as the main input for ASD optimization, accurate estimates of survey design parameters, such as response propensities, are required for reliable strategies. Strictly speaking, inaccuracy jeopardizes ASD performance and design due to the suboptimal and ineffective decisions made in the optimization approach. Adverse impacts are apparent when response propensities change gradually over time.

In this paper, we discussed a methodology to evaluate the impact of temporal factors (e.g., seasonality) on the accuracy of sequential response propensity predictions in a mixed-mode survey with replication, and investigated the manner and timeliness of applying the optimal allocation schemes to population strata. We introduced a Bayesian multinomial time-series model for sequential response propensities and an optimization model for ASDs. The propensity model had a general form that described multiple time-related and strata-related factors, and accounted for the dependence of the current mode's response propensities on the preceding modes' response behaviors. The optimization model, on the other hand, enabled the inclusion of uncertainty in the follow-up workload, and described the way to allocate reviewers to each stratum for the largest decrease in nonresponse risk. Most cross-sectional mixed-mode or unique-mode surveys conducted over many years can fit into this framework. Furthermore, we constructed an analysis for the GEZO survey to examine the highest performance of the propensity model. Owing to diverse model compositions, information criteria measuring the fitness and complexity of the propensity model were adopted to compare the performance of different models. We were thus able to meet the first objective of this paper to select and construct the "favorite" time-series model (Model 9 having lowest information criteria scores) that contributed most to prediction accuracy for a sequential mixed-mode survey.

The second and third objectives were to examine the sensitivity of ASD performance to, respectively, the specified budget level and the amount of historic data that were included. In the evaluation, ASD performance had to be reoptimized when the budget level and/or the length of applicable historic data time series were updated. Then, we benchmarked ASD performance against CAWI-only and nonadaptive design performance. This analysis is essentially a comparison of the reduction in the nonresponse risk if a fraction of CAWI nonresponses (with no follow-up at all and full follow-up as special cases) is assigned to interviewers. To make this comparable for a range of scenarios, we utilized the properties of the posterior distributions, that is, the credible region and expectation. The evaluation examined, in a specific time window, the improvement in performance under different budget levels. Additionally, the evaluation

examined, for a specific budget level, the improvement over rolling time windows. The evaluation showed that ASD performance was quite solid for budget levels greater than 50%, but was inferior to budget levels less than 50%. Additionally, the evaluation showed that without taking seasonality into account, ASD performance was obviously enhanced in the early stage of accumulating data. After that, this trend slowed and even stopped despite increasing, performing almost the same as the with-season model, and consequently hinted at seasonality being of little use to further improvement in prediction accuracy and ASD performance.

Our study has limitations that ask for further research and replication in other mixed-mode surveys.

Ignoring mode-specific seasonal effects was our first simplification. This eased the complexity of the model specification but led to seemingly offset seasonal effects on propensity predictions. However, we believe that one can conveniently accommodate seasonal effects specific to each model to the adjusted model if seasonality is supposed to be a strong predictor for propensity predictions.

To ensure prediction accuracy and reliability, we considered only two candidate data collection strategies (web only and web followed by face-to-face) as the second simplification. The number of CAPI visits to sample units may be further tailored, and the optimization may include the actual number of visits. Response propensities after each visit can be modeled and estimated simultaneously, and the predictions of a follow-up mode could be assumed correlated only with its nearest predecessor. Such an application is easy, but it entails careful checking of the predictions' reliability.

The third limitation, was that our propensity model was sensitive to structural design changes (see Wu et al., 2023 for more discussion) As a result, in the case study of this paper, we noted a temporary misspecification in the prior distributions of response propensities. In an effort to characterize unexpected change, it is of importance to pinpoint the parameters (and strata) that have been affected. Robustness can then be improved by extra hierarchical model parameters. We leave this extension to further research.

Our last limitation was that in our multi-level time series models we assumed that ASD strata were specified beforehand. We did not integrate a variable selection step into the model fitting and optimization. When the number of auxiliary variables is large, there is a trade-off between the learning time of the Bayesian analysis of response propensities and the utility of the optimization, This is an important topic for further research.

The inclusion of paradata and other time-varying covariates in the multi-level time series models would be another relevant extension. This would allow optimization in a dynamic setting, i.e. during fieldwork. In a face-to-face interviewer setting, such a dynamic approach is not operationally straightforward. Interviewer workloads become known only at a point in time close to fieldwork and at set time points, often monthly. To a lesser extent, this would be true for telephone follow-up. A practical solution, applied by Statistics Netherlands, is to add a random subsampling of nonrespondents. The subsampling probabilities depend on pre-specified and fixed workloads in interviewer regions. Going even a step further and allowing a dynamic design to intervene during face-to-face follow-up may be considered as well. In order to leave some freedom to interviewers, this may be implemented as one or more pre-specified tie points where sample units are

stopped or not. Abstracting further to general (sequential) data collection phases, it would be worthwhile to extend the methodology of this paper. Such an extension has three statistical challenges. The first is the inclusion of new incoming auxiliary data, e.g. paradata, in the model and optimization. If and how to include will be a trade-off between efficacy and efficiency. The second is that much more emphasis needs to be put on the right specification of covariances of response propensities of different data collection phases. Misspecifications proliferate in subsequent phases. The third is that optimization must be performed during data collection, demanding for some parsimony in ambition.

Appendix A

Table A.1
Auxiliary variables form 13 strata and season is considered as an influential factor to predict response propensities.

Stratum	Age (years)	Ethnic
1	0 – 17	Western
2	18 – 24	Western
3	25 – 34	Western
4	35 – 54	Western
5	55 – 64	Western
6	65 – 74	Western
7	75+	Western
8	0 – 17	Non-western
9	18 – 24	Non-western
10	25 – 34	Non-western
11	35 – 54	Non-western
12	55 – 64	Non-western
13	65+	Non-western

Appendix B

Table B.1
The bias-adjusted (adj) and unadjusted (unadj) CV observations and the standard errors (se) under bias adjustment in 2017 Q1 under different budget levels.

	W			w-Np			w-Ap		
	unadj	adj	se	unadj	adj	se	unadj	adj	se
0%*	0.305	0.294	0.023	-	-	-	-	-	-
100%	-	-	-	0.156	0.149	0.013	0.102	0.072	0.021
90%	-	-	-	-	-	-	0.093	0.084	0.021
80%	-	-	-	-	-	-	0.104	0.074	0.020
70%	-	-	-	-	-	-	0.092	0.078	0.021
60%	-	-	-	-	-	-	0.093	0.084	0.021
50%	-	-	-	-	-	-	0.106	0.052	0.020
40%	-	-	-	-	-	-	0.114	0.072	0.021
30%	-	-	-	-	-	-	0.136	0.115	0.022
20%	-	-	-	-	-	-	0.259	0.237	0.022
10%	-	-	-	-	-	-	0.305	0.294	0.023

**No budget indicates only the CAWI mode is used. "-" means no results*. w: CAWI-only, w-Ap: nonadaptive, w-Np: adaptive designs.

Table B.2
The bias-adjusted (adj) and unadjusted (unadj) CV observations and the standard errors (se) of bias adjustment under the 100% budget level in each quarter.

Year	Quarter	w			w-Np			w-Ap		
		unadj	adj	se	unadj	adj	se	unadj	adj	se
2016	Q1				-	-	-	-	-	-
	Q2	-	-	-						
	Q3	-	-	-	-	-	-	0.093	0.084	0.021
	Q4	-	-	-	-	-	-	0.104	0.074	0.020
2017	Q1	-	-	-	-	-	-	0.092	0.078	0.021
	Q2	-	-	-	-	-	-	0.093	0.084	0.021
	Q3	-	-	-	-	-	-	0.106	0.052	0.020
	Q4	-	-	-	-	-	-	0.114	0.072	0.021

Note: w: CAWI-only, w-Ap: nonadaptive, w-Np: adaptive designs.

Table B.3
The relative cost (c) of 2017 Q1 under different budget levels for adaptive surveys.

	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
c	0.425	0.473	0.532	0.608	0.709	0.851	0.983	1.311	1.966	0.977
C	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE

Note: The allocations in each case are determined by the optimization solver "auglag" starting with initial point 0 set to 13 strata. If the convergence (C) is TRUE, the local optimum can be found, and the corresponding allocations are returned; otherwise, the process results in a false local "optimum".

References

Boonstra, H.-J. (2022). *mcmcsae*: Markov Chain Monte Carlo small area estimation. R package version 0.7.2.

Boonstra, H.-J., and van den Brakel, J.A. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 3, 395-425. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2019003/article/00005-eng.pdf.

Calinescu, M., Bhulai, S. and Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, 226(1), 115-121. DOI: https://doi.org/10.1016/j.ejor.2012.10.046.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11250-eng.pdf.

Chun, A.Y., Heeringa, S.G. and Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34(3), 581-597. DOI: https://doi.org/10.2478/jos-2018-0028.

Coffey, S., West, B.T., Wagner, J. and Elliott, M.R. (2020). What do you think? Using expert opinion to improve predictions of response propensity under a Bayesian framework. *Methods, Data, Analyses*. DOI: https://doi.org/10.12758/mda.2020.05.

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.

Linderman, S., Johnson, M.J. and Adams, R.P. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. *Advances in Neural Information Processing Systems*, 28. DOI: https://doi.org/10.48550/arXiv.1506.05843.

Ma, Y. (2021). *Optimal Stratification in Bayesian Adaptive Survey Designs*. Master Thesis, University Utrecht, The Netherlands.

Moore, J.C., Durrant, G.B. and Smith, P.W.F. (2018). Data set representativeness during data collection in three UK social surveys: Generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1), 229-248. DOI: https://doi.org/10.1111/rssa.12256.

Nishimura, R., Wagner, J. and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review*, 84(1), 43-62. DOI: https://doi.org/10.1111/INSR.12100.

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. DOI: https://doi.org/10.1198/016214508000000724.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias through case prioritization. *Survey Research Methods*, 4(1), 21-29. DOI: https://doi.org/10.18148/SRM/2010.V4I1.3037.

Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108, 1339-1349.

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/pub/12-001-x/2010002/article/11376-eng.pdf.

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

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

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P. and Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smy012.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*. Chapman and Hall/CRC.

Scott, A., and Smith, T.M. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64(327), 830-840.

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: Series B (Statistical Methodology)*, 64(4), 583-639. DOI: https://doi.org/10.1111/1467-9868.00353.

van Berkel, K., van der Doef, S. and Schouten, B. (2020). Implementing adaptive survey design with an application to the Dutch Health Survey. *Journal of Official Statistics*, 36(3), 609-629. DOI: https://doi.org/10.2478/jos-2020-0031.

Varadhan, R. (2022). *alabama: Constrained Nonlinear Optimization*. R package version 2022.4-1.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413-1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7(1), 45-55. DOI: https://doi.org/10.18148/SRM/2013.V7I1.5037.

Wagner, J., and Hubbard, F. (2013). Using propensity models during data collection for responsive designs: Issues with estimation. In 68th AAPOR conference, May (pp. 16-19).

Wagner, J.R. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. (Doctoral dissertation, University of Michigan).

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. https://arxiv.org/abs/1004.2316.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897. https://arxiv.org/abs/1208.6338.

West, B.T., Wagner, J., Coffey, S. and Elliott, M.R. (2023). Deriving priors for Bayesian prediction of daily response propensity in responsive survey design: Historical data analysis vs. literature review. *Journal of Survey Statistics and Methodology,* 11(2), 367-392.

Wu, S., Boonstra, H.-J., Moerbeek, M. and Schouten, B. (2023). Modelling time change in survey response rates: A Bayesian approach with an application to the Dutch Health Survey. *Survey Methodology*, 49, 1, 163-190. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023001/article/00010-eng.pdf.

Wu, S., Schouten, B., Meijers, R. and Moerbeek, M. (2022). Data collection expert prior elicitation in survey design: Two case studies. *Journal of Official Statistics*, 38(2), 637-662. DOI: https://doi.org/10.2478/JOS-2022-0028.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 465-486
Statistics Canada, Catalogue No. 12-001-X

Bayesian predictive inference of a finite population mean without specifying the relation between the study variable and the covariates

Ashley Lockwood and Balgobin Nandram[1]

Abstract

While we avoid specifying the parametric relationship between the study variable and covariates, we illustrate the advantage of including a spatial component to better account for the covariates in our models to make Bayesian predictive inference. We treat each unique covariate combination as an individual stratum, then we use small area estimation techniques to make inference about the finite population mean of the continuous response variable. The two spatial models used are the conditional autoregressive and simple conditional autoregressive models. We include the spatial effects by creating the adjacency matrix via the Mahalanobis distance between covariates. We also show how to incorporate survey weights into the spatial models when dealing with probability survey data. We compare the results of two non-spatial models including the Scott-Smith model and the Battese, Harter, and Fuller model to the spatial models. We illustrate the comparison between the aforementioned models with an application using BMI data from eight counties in California. Our goal is to have neighboring strata yield similar predictions, and to increase the difference between strata that are not neighbors. Ultimately, using the spatial models shows less global pooling compared to the non-spatial models, which was the desired outcome.

Key Words: Conditional autoregressive model; Hierarchical Bayesian model; Simple conditional autoregressive model; Spatial modeling.

1. Introduction

In this paper, when making inference about the finite population mean, we refrain from assuming a relationship between the response variable and the covariates. We avoid making the strong assumptions of regression models, and therefore increase the number of situations our models can be applied to. The methods we present avoid defining this relationship by considering each unique combination of the covariates in the population as an individual stratum. We accommodate the covariates by using the spatial model instead of a regression model. Then, we use small area estimation techniques to make inference about each stratum of the population based on its underlying covariates. Finally, we can understand the overall population by pooling predictions of the strata together (Rao and Molina, 2015).

We present two versions of spatial models, a conditional autoregressive (CAR) model and a simple conditional autoregressive (SCAR) model (Chung and Datta, 2022). For the spatial models, we include the spatial effects by creating the adjacency matrix via the Mahalanobis distance between the covariates for each stratum. We use these spatial models to create a neighborhood relationship between similar strata to allow for less global pooling to the overall sample mean. By allowing strata to have neighbors, we expect neighborhoods to pool together without remote strata pooling together. Using a spatial model versus a non-spatial model should provide posterior predictions with a larger variation between predicted stratum means.

1. Ashley Lockwood, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA. E-mail: anlockwood@wpi.edu; Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA. E-mail: balnan@wpi.edu.

We also present two non-spatial models for comparison, a form of the Scott-Smith model (Scott and Smith, 1969) and a form of the Battese, Harter, and Fuller (BHF) model (Battese, Harter and Fuller, 1988). The BHF model is a more general version of the Scott-Smith model that includes covariates in the model, whereas the Scott-Smith model does not. We use both non-spatial models as a baseline for comparison to see how including a spatial relationship in the models impacts the results. Appendix A contains the full technical details of the Scott-Smith model, and Appendix B contains the technical details of the BHF model.

Furthermore, Datta and Ghosh (1991) expand upon the research conducted by Battese, Harter and Fuller (1988), offering a comprehensive analysis of the nested error regression hierarchical Bayesian model, with a particular focus on small area estimation. This work significantly advances the generalization of computational formulas for deriving Bayesian predictors and their associated standard errors. However, it's notable that the approach in Datta and Ghosh (1991) continues to utilize mixed linear models, which explicitly establish the relationship between covariates and the response variable. In contrast, our methodology in this paper intentionally avoids such explicit relationship definitions, resulting in significant methodological divergence.

There are many traditional regression models that make inference about a characteristic of a population, including logistic regression, general linear models, general multivariate normal models, and classification and regression tree (CART). See Lindley and Smith (1972), Ghosh, Natarajan, Stroud and Carlin (1998), Albert and Chib (1993), Box and Tiao (1973), and Chipman, George and McCulloch (1998) for detail about each model. While these models have been widely used throughout history, the strong distribution assumptions made to properly use these models limits the types of data and situations available for application.

There also exists other models without regression coefficients that answer a similar question, including Dirichlet processes, Polya urn scheme, and Bayesian additive regression trees (BART). See Blackwell and MacQueen (1973), Antoniak (1974), Yin and Nandram (2020), Teh, Jordan, Beal and Blei (2006) and Chipman, George and McCulloch (2010) for information about these alternative models. Dirichlet processes and Polya urn schemes are popular in Bayesian modeling, however these complicated computations can lead to poor mixing in the Markov chain Monte Carlo (MCMC) algorithm. BART is a newer approach, but this method violates traditional Bayesian logic by double use of the data. The data are used in the likelihood of the BART model, and then again in a data-informed prior for two hyperparameters (Hill, Linero and Murray, 2020). We aim to improve the computation of models without regression coefficients while maintaining the coherence of the Bayesian paradigm.

For the remainder of the paper, we discuss the methodology of the two spatial models as well as an extension of including survey weights into the models in Section 2. Then in Section 3, an application using BMI data with each of the models is given, followed by a conclusion in Section 4. Appendix A contains technical details for the Scott-Smith model, and shows how to include survey weights in this model. Similarly, Appendix B contains technical details for the BHF model and shows how to include the survey weights.

2. Methodology

In this section, we show two spatial models and how we include survey weights in the spatial models. First, in Section 2.1 we present the two spatial models with the CAR model in Section 2.1.1 and the SCAR model in Section 2.1.2. Then, Section 2.2 illustrates how the survey weights can be included in the spatial models presented in Section 2.1. The methodologies for the Scott-Smith model and the BHF model can be found in Appendix A and Appendix B, respectively.

In all four models we observe a continuous response variable y_{ij} for sampling unit $j = 1,\ldots,n_i,$ belonging to stratum $i = 1,\ldots,\ell,$ and these responses are grouped together based on their covariate values. Each possible combination of covariates is taken into consideration, and each unique combination is considered to be a stratum. Therefore, each $\mathbf{y}_i$ has a unique corresponding covariate combination denoted $\mathbf{x}_i,$ where $\mathbf{y}_i$ is the aggregated vector of responses of length n_i for each stratum. The covariate matrix $\mathbf{X} = \left(\mathbf{x}_i'\right)$ has dimension $\ell \times p$ where p is the number of covariates in the data. The matrix $\mathbf{X}$ does not include an intercept column, and $\mathbf{x}_i$ corresponds to the unique rows of $\mathbf{X}$. We make inference about the finite population mean, $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i,$ based on the observed values of $\mathbf{y}_i$. Denote the sampling fraction as $f_i = n_i / N_i,$ where n_i represents the sample size and N_i represents the population size for a given stratum. The N_i in our application are unknown and we discuss later how to estimate them using inverse probability weighting. Whenever feasible, utilizing the actual population sizes, $N_i,$ is optimal and preferred. When dealing with an exceptionally high number of strata, the estimations of N_i may become increasingly susceptible to noise.

2.1 Spatial models

For the spatial models, we include the spatial effects by creating the symmetric adjacency matrix, $\mathbf{W}$ of size $\ell \times \ell,$ via the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ for $i = 1,\ldots,\ell,$ $i' = 1,\ldots,\ell,$ and $i \neq i'$. The Mahalanobis distance is defined as:

$$d_{ii'} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_{i'}\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \mathbf{x}_{i'}\right)}, \tag{2.1}$$

where $\mathbf{S}$ is the covariance matrix of $\mathbf{X},$ and $d_{ii} = 0.$ We define $\mathbf{W}$ by letting $w_{ii'} = 1$ if $d_{ii'} \leq d_0$ and $w_{ii'} = 0$ if $d_{ii'} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$). A grid search is conducted to determine the value d_0 that yields a $\mathbf{W}$ matrix that maximizes Moran's $I,$ which is defined as:

$$I = \frac{\ell}{w_{\cdot\cdot}} \frac{\sum_i \sum_{i'} w_{ii'} \left(\bar{y}_i - \bar{y}\right)\left(\bar{y}_{i'} - \bar{y}\right)}{\sum_i \left(\bar{y}_i - \bar{y}\right)^2}, \tag{2.2}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ is the response variable; $\bar{y} = \sum_{i=1}^{\ell} \bar{y}_i / \ell$ is the overall sample mean response; $w_{ii'}$ corresponds to the elements of $\mathbf{W}$; and $w_{\cdot\cdot} = \sum_i \sum_{i'} w_{ii'}$.

In Section 2.1.1 we describe the conditional autoregressive (CAR) model and in Section 2.1.2 we state the difference between this model and the simple conditional autoregressive (SCAR) model.

2.1.1 CAR model

The Bayesian hierarchical CAR model is:

$$\begin{aligned} y_{ij} \mid \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}\left(\mu_i, \sigma^2\right), \\ \boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma &\sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho}\sigma^2\left(\mathbf{R}-\gamma\mathbf{W}\right)^{-1}\right), \\ \pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \gamma &\sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \frac{1}{\lambda_1} \le \gamma \le \frac{1}{\lambda_\ell}, \quad -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\ & j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell, \end{aligned} \tag{2.3}$$

where ℓ in this case represents the total number of possible covariate combinations which are considered to be the individual strata. We discretize any continuous variables so there is a finite number of possible covariate combinations. Then, we store the continuous responses, $\mathbf{y}_i$, such that there are n_i responses for each stratum $i = 1, \ldots, \ell$. Here $\mathbf{R}$ is a diagonal $\ell \times \ell$ precision matrix defined as $\mathbf{R} = \text{diag}\left\{w_{i\cdot}\right\}_{i=1}^{\ell}$ where $w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$ is the sum of the ith row of $\mathbf{W}$. Also, λ_1 is the minimum eigenvalue of $\mathbf{R}^{-1}\mathbf{W}$ and λ_ℓ is the maximum eigenvalue of $\mathbf{R}^{-1}\mathbf{W}$, and since $\sum_{i=1}^{\ell} w_{ii} = 0$ this results in $\lambda_1 < 0 < \lambda_\ell$ (Chung and Datta, 2022). Here, $\left(\mathbf{R}-\gamma\mathbf{W}\right)$ is guaranteed to be positive definite as long as γ is in the range $\frac{1}{\lambda_1} \le \gamma \le \frac{1}{\lambda_\ell}$. To obtain samples from the joint posterior density of this model, we can integrate out $\boldsymbol{\mu}$, θ, and σ^2, and then we only need to draw γ and ρ using a griddy Gibbs sampler (Ritter and Tanner, 1992).

We can vectorize the continuous response variable, y_{ij}, to be $\mathbf{y}$ with dimension $n \times 1$ where $n = \sum_{i=1}^{\ell} n_i$ such that:

$$\mathbf{y}_{(n\times 1)} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(A\boldsymbol{\mu}, \sigma^2 \mathbb{I}_{n\times n}\right), \tag{2.4}$$

where $\mathbb{I}_{n\times n}$ is the identity matrix and A has dimension $n \times \ell$ and can be defined as

$$A = \begin{pmatrix} \mathbf{1_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{1_\ell} \end{pmatrix}, \tag{2.5}$$

and $\mathbf{1_1}$ through $\mathbf{1_\ell}$ are vectors of ones with lengths corresponding to the number of observations in that stratum. Therefore, $\mathbf{1_1}$ is a vector of ones with length n_1, $\mathbf{1_2}$ is a vector of ones with length n_2, and so on through $\mathbf{1_\ell}$. The purpose of writing the model in this way is so we can use the lemma from Section 2 in Lindley and Smith (1972) to obtain the posterior distribution of $\boldsymbol{\mu}$ which we draw samples of $\boldsymbol{\mu}$ from:

$$\begin{aligned} \boldsymbol{\mu} \mid \Omega, \mathbf{y} \sim \text{Normal}\Bigg[& \left(\text{diag}\left(n_1, \ldots, n_\ell\right) + \frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)\right)^{-1}\left(A'\mathbf{y} + \left(\frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)\right)\theta\mathbf{1}\right), \\ & \sigma^2\left(\text{diag}\left(n_1, \ldots, n_\ell\right) + \frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)\right)^{-1}\Bigg]. \end{aligned} \tag{2.6}$$

Let $\Omega = \left(\theta, \rho, \sigma^2, \gamma\right)$ for simplicity of notation.

Another way of writing this spatial model to make it simpler to integrate out $\boldsymbol{\mu}$ would be:

$$\begin{aligned} \overline{\mathbf{y}} \,|\, \boldsymbol{\mu}, \sigma^2 \;&\sim\; \text{Normal}\left(\boldsymbol{\mu}, \sigma^2 \text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right)\right), \\ \boldsymbol{\mu} \,|\, \theta, \rho, \sigma^2, \gamma \;&\sim\; \text{Normal}\left(\theta\mathbf{1}, \frac{\rho}{1-\rho}\sigma^2\left(\mathbf{R} - \gamma\mathbf{W}\right)^{-1}\right). \end{aligned} \tag{2.7}$$

Now, if we integrate out $\boldsymbol{\mu}$ from this model we are left with the posterior density:

$$\begin{aligned} \pi\left(\theta, \rho, \sigma^2, \gamma \,|\, \mathbf{y}\right) \;\propto\; & \det\left[\sigma^2\left(\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\mathbf{R} - \gamma\mathbf{W}\right)^{-1}\right)\right]^{-1/2} \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left(\overline{\mathbf{y}} - \theta\mathbf{1}\right)'\left(\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\mathbf{R} - \gamma\mathbf{W}\right)^{-1}\right)^{-1}\left(\overline{\mathbf{y}} - \theta\mathbf{1}\right)\right\} \\ & \times \left(\frac{1}{\sigma^2}\right)^{(n-\ell)/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i - 1)\, s_i^2\right\} \times \frac{1}{\sigma^2} \end{aligned} \tag{2.8}$$

where $s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}{(n_i - 1)}$. From this density, we can see that θ follows a normal distribution:

$$\theta \,|\, \sigma^2, \rho, \gamma, \overline{\mathbf{y}} \;\sim\; \text{Normal}\left(\hat{\theta}, \frac{\sigma^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right), \tag{2.9}$$

where $\hat{\theta} = \mathbf{1}'\boldsymbol{\Sigma}\overline{\mathbf{y}} / \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}$ and $\boldsymbol{\Sigma} = \left[\text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\mathbf{R} - \gamma\mathbf{W}\right)^{-1}\right]^{-1}$. We can use this fact to integrate out θ, so we have,

$$\begin{aligned} \pi\left(\rho, \sigma^2, \gamma \,|\, \mathbf{y}\right) \;\propto\; & \det\left[\sigma^2\boldsymbol{\Sigma}^{-1}\right]^{-1/2} \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right)\right\} \times \sqrt{2\pi\sigma^2 / \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} \\ & \times \left(\frac{1}{\sigma^2}\right)^{(n-\ell)/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i - 1)\, s_i^2\right\} \times \frac{1}{\sigma^2}. \end{aligned} \tag{2.10}$$

From this density, we can find the inverse-gamma distribution,

$$\sigma^2 \,|\, \rho, \gamma, \overline{\mathbf{y}} \;\sim\; \text{InvGam}\left(\frac{n-1}{2}, \left[\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right) + \sum_{i=1}^{\ell}(n_i - 1)\, s_i^2\right] \Big/ 2\right). \tag{2.11}$$

Finally, after integrating out σ^2 we have the nonstandard joint posterior density,

$$\pi\left(\rho, \gamma \,|\, \mathbf{y}\right) \;\propto\; \det\left[\boldsymbol{\Sigma}^{-1}\right]^{-1/2}\left[\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1} - \overline{\mathbf{y}}\right) + \sum_{i=1}^{\ell}(n_i - 1)\, s_i^2\right]^{\frac{-n+1}{2}}\left(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}\right)^{-1/2}. \tag{2.12}$$

Using the griddy Gibbs sampler, we can draw samples of ρ and γ from (2.12) (Ritter and Tanner, 1992). We use the same conditional posterior density to draw both parameters using a grid method, however

the grids of ρ and γ differ since their ranges of support are not equivalent. This method mixes well and converges quickly. Note that the griddy Gibbs sampler converges to an approximation of the posterior distribution of interest, rather than its exact form. Then, continuing in reverse order we can input our samples of ρ and γ to directly obtain samples of σ^2 from (2.11), next θ from (2.9), and finally $\boldsymbol{\mu}$ from (2.6). Obtaining samples of σ^2, θ, and $\boldsymbol{\mu}$ is straight-forward since they all have known distributions. Based on our samples of $\boldsymbol{\mu}$, σ^2 and the observed values of $\mathbf{y}_i$, we make inference for the finite population mean $\overline{Y}_i$, using the model:

$$\overline{Y}_i \mid \mu_i, \sigma^2, \mathbf{y}_i \overset{\text{ind}}{\sim} \text{Normal}\left(f_i \overline{y}_i + (1-f_i)\,\mu_i, (1-f_i)\,\frac{\sigma^2}{N_i} \right). \tag{2.13}$$

We examine the performance of this model in Section 3.1 with an application using BMI data.

2.1.2 SCAR model

Here we describe the simple conditional autoregressive (SCAR) model and state the differences between this model and the previously discussed CAR model. The main computational difference between the two models is that in the CAR model the matrix $\mathbf{R}$ is used in the variance of the prior on $\boldsymbol{\mu}$, and the SCAR model replaces $\mathbf{R}$ with the identity matrix, $\mathbb{I}$, hence simplifying the model. Therefore, the Bayesian hierarchical SCAR model is defined by (2.3) with $\mathbf{R} = \mathbb{I}$.

Aside from the small computational changes mentioned, the distributions and methods we use to obtain a sample from the posterior density and the method we use to make inference about the finite population mean, $\overline{Y}_i$, is the same as described in Section 2.1.1. We simply substitute $\mathbf{R}$ for the matrix $\mathbb{I}$ and use the updated values of λ_1 and λ_ℓ accordingly. We illustrate the performance of this model in Section 3.1 with an application using BMI data.

In the SCAR model, the precision matrix is set to be the identity matrix, $\mathbb{I}$. While the diagonal elements of the precision matrix are all equal, the diagonal elements of the inverse may not be all equal thus allowing for heteroscedasticity of random effects. In the CAR model, diagonal entries of the precision matrix, $\mathbf{R}$, are the number of neighbors corresponding to each stratum. Therefore, the matrix $\mathbf{R}$ weights each row by the number of neighbors it has, and acts as a normalizing matrix. Both the SCAR and CAR models assume that μ_i depends only on neighboring strata means and not on remote strata (Chung and Datta, 2022). Let $\mathbf{Q} = \mathbf{R} - \gamma\mathbf{W}$, then μ_i and μ_j for $i \neq j$ are conditionally independent, conditional on all μ_k for $k \neq i \neq j$, whenever $Q_{ij} = 0$. Note that in the CAR and SCAR models, it is important that ρ and γ are not too small, because we want to emphasize the spatial structure in order to accommodate the covariates.

2.2 Including survey weights

In this section, we show how to include survey weights in the two spatial models we are advocating for. Here we use the original survey weights, denoted v_{ij} for $j = 1,\ldots,n_i$, and $i = 1,\ldots\ell$, to calculate the effective sample size and the adjusted weights a_{ij}. First, we calculate the effective sample size, $\hat{n}$:

$$\hat{n} = \frac{\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}\right)^2}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^2}. \tag{2.14}$$

The effective sample size, $\hat{n}$, illustrates how severely the variance is increased by the unequal weighting (Nandram and Rao, 2021). Then, we calculate the adjusted weights, a_{ij}, which are used to eliminate the bias present in the original survey weights:

$$a_{ij} = \hat{n}\frac{v_{ij}}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}}, \tag{2.15}$$

where $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij} = \hat{N}$ is the Horvitz-Thompson estimator of population size, and $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} a_{ij} = \hat{n}$, the effective sample size.

These adjusted weights a_{ij} are able to be used in a model when the data do not have outliers present. However, in the BMI example in Section 3 our data do have outliers, so we use Winsorization which is an effective method to deal with outliers by trimming the survey weights (Yang, Nandram and Choi, 2023). Outliers here are defined as observed survey weights greater than $v_0 = Q_3 + 1.5(Q_3 - Q_1)$, where Q_1 is the first quartile and Q_3 is the third quartile. Let v^* denote weights after trimming,

$$v_{ij}^* = \begin{cases} v_0, & v_{ij} \geq v_0 \\ rv_{ij}, & v_{ij} < v_0 \end{cases}, \tag{2.16}$$

where r is a rescaling parameter such that $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^* = \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij} = \hat{N}$. Then we obtain the adjusted and trimmed weights a_{ij}^*,

$$\begin{aligned} \hat{n}^* &= \frac{\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^*\right)^2}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^{*2}}, \\ a_{ij}^* &= \hat{n}^*\frac{v_{ij}^*}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^*}. \end{aligned} \tag{2.17}$$

These adjusted and trimmed weights a_{ij}^* are used in both the CAR and SCAR model.

2.2.1 Including survey weights in CAR model

The CAR model with adjusted weights can be expressed by replacing the response variance in the first row of (2.3) from σ^2 to $\frac{\sigma^2}{a_{ij}^*}$, with a_{ij}^* from (2.17). We use the same logic for obtaining a sample from this model with the adjusted weights as we used in Section 2.1.1. The difference is in how we make population predictions, by including the survey weights we now need to use surrogate sampling techniques. We obtain population predictions by:

$$\overline{Y}_i \,|\, \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(\mu_i, \frac{\sigma^2}{\hat{N}_i}\right) \qquad i = 1, \ldots, \ell, \tag{2.18}$$

where $\hat{N}_i = \sum_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each stratum $i = 1, \ldots, \ell$. We no longer need to include the sample means, $\bar{y}_i$, or the sampling fraction, f_i, in this population prediction.

Previously when making population predictions, we combined both the sampled part and non-sampled part of the population to obtain a sample of $\overline{Y}_i$. However, now that we are utilizing the survey weights from the probability sample we are able to use surrogate sampling techniques and no longer need to include the sampled part. The adjusted and trimmed survey weights, $\mathbf{a}^*$, are used to simulate an unbiased sample from the population. With the adjusted and trimmed weights included in the model we must sample the entire population using surrogate sampling techniques, because the survey weights of both the sample and the nonsample are biased (Nandram, 2007; Nandram and Rao, 2021). Then, when we make population predictions in (2.18) we use the unadjusted survey weights for prediction since these weights accurately represent the entire population. We examine the performance of this CAR model with the inclusion of survey weights in Section 3.1 with an application using BMI data.

2.2.2 Including survey weights in SCAR model

Similarly, the SCAR model with adjusted weights can be expressed by replacing the response variance in the first row of (2.3) from σ^2 to $\frac{\sigma^2}{a_{ij}^*}$, with a_{ij}^* from (2.17), and letting $\mathbf{R} = \mathbb{I}$.

We use the same logic for obtaining a sample from this model with the adjusted weights as we used in Section 2.1.1. Similar to the CAR model with survey weights included we now need to use surrogate sampling techniques in the SCAR model with survey weights to make population predictions. We use (2.18) to make our population predictions in the SCAR model with survey weights included. We examine the performance of this SCAR model with the inclusion of survey weights in Section 3.1 with an application using BMI data.

3. Application using BMI data

When interested in the health of a population, the the BMI levels of individuals may be an important indicator. We illustrate our various non-spatial and spatial models using a probability sample of BMI data with 1,867 individuals from eight counties in California recorded in the Third National Health and Nutrition Examination Survey (NHANES III), (Nandram and Choi, 2005). The survey weights sum to 12,232,099, which means our sample accounts for 0.015% of the population. Examining the original survey weights, the effective sample size from (2.14) is 498, significantly lower than the observed sample size. Since the survey weights are right-skewed with a few large outliers, we apply Winsorization to trim the weights. This adjustment yields a rescaled effective sample size of 1,300 from (2.17), which is closer to the observed sample size. The adjusted and trimmed survey weights in (2.17) are more evenly distributed, and these are the weights utilized in this application.

These data have four variables we are interested in including age, race, sex, and a continuous measure of BMI in kg/m^2. The values for race are "white" or "non-white", and the values for sex are "male" or "female". The age variable is a continuous variable included in the data, and age ranges from 20 years old to 90 years old. We bin the age variable into groups of two years, therefore the bins are 20-21 years old, 22-23 years old, and so on, in order to have a finite number of possible covariate combinations. This idea of binning variables is commonly used in practice, as it lessens the need to rely on exact accuracy of the data and allows for inference to be made about a broader age group. To fully encompass all potential covariate combinations present in the population when employing this model, we advise to bin all continuous covariates. In this BMI example, where the continuous age variables has inherent lower and upper bounds, these bounds will persist in the population. However, if such bounds do not naturally emerge in the data, they should be enforced through the use of bins to prevent groups in the population being unaccounted for in the sample data. The value for the continuous BMI response variable ranges from 15.8 kg/m^2 to 58.4 kg/m^2.

After aggregating over all possible age, race, and sex combinations there are 144 strata each with its own unique set of covariate values. However, in our sample of BMI data we have 12 strata with no observations and we assume these are structural zeroes in the data. This means that we assume these 12 groups of individuals do not exist in our population. Therefore, the total number of strata, ℓ, in this case represents the total number of possible covariate combinations available in our sample and $\ell = 132$ after removing the 12 structural zeroes from the data. If we want to avoid making this assumption, we can mitigate the structural zeroes by using coarser groups of the covariates. Without making coarser groups, it would be necessary to have the known population total, N_i, for each stratum we are including in the model. In this case, we would not need to rely on the survey weights to estimate N_i using the Horvitz-Thompson estimator of population size for each stratum $i = 1, \ldots, \ell$.

When expanding the number of covariates utilized in the model, the sparse nature of the adjacency matrix, $\mathbf{W}$, facilitates the management of larger sets of covariates. If needed, coarser groups of the covariates can also be employed to decrease the number of covariate combinations and thereby mitigate computational costs. Furthermore, rather than solely focusing on maximizing Moran's I during the construction of the adjacency matrix, we have the option to prioritize achieving greater sparsity in the matrix.

In Section 3.1 we illustrate and compare the results between the two spatial and two non-spatial models, with results both including and excluding survey weights. Then in Section 3.2, we show how including the spatial component in our models reduced the amount of global pooling, compared to the non-spatial models.

3.1 BMI application model comparisons

Before sampling from either of the spatial models, we first create the symmetric adjacency matrix, $\mathbf{W}$ of size 132×132, using the Mahalanobis distance described in (2.1). To prevent the inclusion of an ordinal categorical covariate in the Mahalanobis distance calculation, we utilize the mean value of each age bin for

the age covariate. Recall that we define $\mathbf{W}$ by letting $w_{ii'} = 1$ if $d_{ii'} \leq d_0$ and $w_{ii'} = 0$ if $d_{ii'} > d_0$ with zeroes on the diagonal (i.e., $w_{ii} = 0$), where d_0 is the value yielding the $\mathbf{W}$ matrix that maximizes Moran's I from (2.2). After conducting a grid search to determine the optimal value of d_0, we attained the maximum value of Moran's I, reaching $I = 0.212$ when setting $d_0 = \text{mean}(d_{ij})/38 \approx 0.157$. In general, we found decreasing d_0 increases Moran's I up to a certain point. In our case, if we continue to decrease d_0 to be less than 0.157, we will not see any increase in Moran's I. However, if we increase d_0 to be greater than 0.157 then Moran's I will decrease.

Now that we have $\mathbf{W}$, we can proceed drawing samples from the CAR and SCAR models described in Section 2.1. Sampling from these models is extremely similar as they both begin using the griddy Gibbs sampler to obtain samples of ρ and γ simultaneously (Ritter and Tanner, 1992). For the CAR model the grid interval for γ is: (-1.94, 1) and for the SCAR model the grid interval for γ is: (-0.390, 0.169). The grid intervals for γ in these models differ since the range of γ is based on the eigenvalues of $\mathbf{R}^{-1}\mathbf{W}$ for the CAR model and the eigenvalues of $\mathbf{W}$ for the SCAR model. The grid interval from the CAR model is better in the sense that it brings γ closer to unity. We also present the results with the survey weights included in the CAR and SCAR models.

We ran 10,000 iterations of the sampler, then dropped the first 1,000 sampled values and chose every 9th sampled value to end with a final sample size of 1,000 for both parameters. In both the CAR and SCAR models, the griddy Gibbs sampler shows good performance as evident by the trace plots, auto-correlations, Geweke test of stationarity, and the effective sample sizes. For ρ and γ the P-values for the Geweke test are 0.237 and 0.286, respectively, in the CAR model and 0.938 and 0.833, respectively, in the SCAR model, meaning both parameters pass stationarity requirements in each model. As seen in Table 3.1 in the CAR model excluding survey weights, the posterior means for ρ and γ are 0.188 and 0.937, respectively, and in the SCAR model excluding survey weights, the posterior means for ρ and γ are 0.044 and 0.165, respectively. Since it is important that ρ and γ are not too small considering we want to emphasize the spatial structure that accommodates the covariates, then we prefer the CAR model which has larger values of both ρ and γ. In the CAR model, γ is close to unity which is a good sign that our spatial component will have more of an impact in the model compared to the much lower γ value in the SCAR model. As ρ and γ decrease, our posterior standard error of the population predictions will also decrease.

In Table 3.2, we observe that $\overline{Y}$ in the SCAR model has a slightly lower posterior standard error compared to the CAR model, and this is due to the lower values of ρ and γ in the SCAR model shown in Table 3.1. After successfully running the griddy Gibbs sampler to obtain values for ρ and γ, we continue to sample the remaining parameters σ^2, θ, and $\boldsymbol{\mu}$ directly from their known posterior densities for both the CAR and SCAR models. Each model contains 136 total parameters that we then use to predict the BMI of the population using (2.13). The method for obtaining samples from the CAR and SCAR models with weights included is the same as described in the models with weights excluded, and the griddy Gibbs sampler has a very similar good performance (Ritter and Tanner, 1992). Making population predictions is different in the CAR and SCAR models when we include the survey weights, as shown in (2.18).

Table 3.1
Posterior estimates of ρ and γ.

	PM	PSE	CV	95% HPDI
Models excluding survey weights				
ρ (CAR)	0.188	0.045	0.244	(0.108, 0.289)
γ (CAR)	0.937	0.050	0.054	(0.848, 1.000)
ρ (SCAR)	0.044	0.013	0.292	(0.021, 0.069)
γ (SCAR)	0.165	0.004	0.026	(0.159, 0.169)
Models including survey weights				
ρ (CAR)	0.185	0.048	0.262	(0.098, 0.282)
γ (CAR)	0.940	0.052	0.055	(0.851, 1.000)
ρ (SCAR)	0.042	0.013	0.316	(0.018, 0.068)
γ (SCAR)	0.166	0.004	0.023	(0.159, 0.169)

Notes : PM = posterior mean; PSE = posterior standard error; CV = coefficient of variation; HPDI = highest posterior density interval; CAR = conditional autoregressive; SCAR = simple conditional autoregressive.

Table 3.2
BMI population prediction model comparison.

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$	DIC
Models excluding survey weights					
CAR	27.402	0.091	0.003	(27.233, 27.584)	-73.2
SCAR	27.418	0.088	0.003	(27.237, 27.579)	-70.5
Scott-Smith	27.375	0.129	0.005	(27.117, 27.634)	-61.0
BHF	27.347	0.132	0.005	(27.109, 27.608)	-61.4
Models including survey weights					
CAR	27.070	0.100	0.004	(26.879, 27.263)	-119.9
SCAR	27.090	0.100	0.004	(26.898, 27.268)	-113.0
Scott-Smith	27.380	0.147	0.005	(27.070, 27.656)	-55.5
BHF	27.294	0.161	0.006	(27.007, 27.614)	-60.4

Notes : SE = standard error; CV = coefficient of variation; HPDI = highest posterior density interval; DIC = Deviance Information Criterion; CAR = conditional autoregressive; SCAR = simple conditional autoregressive; BHF = Battese, Harter and Fuller.

A sample from the Scott-Smith model (presented in Appendix A) can be obtained directly without the need for any MCMC algorithm. Once we sample ρ using the grid method, then we can draw samples of σ^2, θ, and $\boldsymbol{\mu}$ in order from their known conditional posterior densities. Since there are 132 strata, our model contains , $\mu_1,\ldots,\mu_{132}$, resulting in 135 total parameters sampled in this model. A sample size of 1,000 was used to predict each parameter of the Scott-Smith model. Once these parameters are obtained we are able to make predictions for the BMI of the population using (A.9).

Similar to the Scott-Smith model, a sample from the BHF model (presented in Appendix B) can also be obtained without the need for an MCMC sampler. For this model, we sample ρ using the grid method, then we can draw samples of σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\nu}$ in order from their known conditional posterior densities. This is the only model containing covariates directly in the model, so we have $\beta_0,\ldots,\beta_3$ where β_0 represents the intercept's coefficient. Therefore, there are 138 total parameters that once sampled we are able to predict the BMI of the population using (B.12). A sample size of 1,000 was used to predict each parameter of the BHF model. The Scott-Smith and BHF models with weights included are fitted by sampling from the same posterior density as in the case of the models excluding weights. Only the population predictions are different. The population predictions can be made using (A.11) and (B.14) for the Scott-Smith and BHF

models with weights included, respectively. The Scott-Smith and BHF models are used as a baseline to compare to the performance of the spatial model.

Table 3.2 contains the results of the population prediction of BMI for all four models, both excluding and including survey weights. In our application the response variable is BMI, therefore $\overline{Y}$ represents the overall mean of BMI for the population of the eight counties in California. The results from the two non-spatial models, the Scott-Smith and the BHF models, are very similar in all four measures of the posterior mean, posterior standard error (SE), coefficient of variation (CV), and highest posterior density interval (HPDI). The two spatial models, CAR and SCAR, also perform similar to each other. The CAR and SCAR models resulted in a slightly higher prediction of the posterior finite population mean BMI of the population. Both the CAR and SCAR models outperform the non-spatial models in terms of DIC, with the CAR model showing the lowest DIC value, closely followed by the SCAR model. The spatial models also have a smaller posterior standard error and CV which yields a tighter HPDI compared to the non-spatial models. Since strata are gaining strength from neighboring strata in the spatial models, we see the posterior standard errors decrease while the posterior means are more tailored to each neighborhood. Strata with a very small sample size are no longer relying on solely their limited number of observations in the spatial models, since they are now included in neighborhoods that collectively have a larger number of observations. In the models without the spatial component, predictions for strata are more general leading to more vague predictions centered around the overall sample mean with larger posterior standard error.

Table 3.2 also contains the results of the population prediction of BMI when the adjusted and trimmed survey weights are included in the models. From the table we can see that the overall population prediction for the finite population mean BMI is similar to that of the models without weights. However, the posterior standard error and the CV increase in the models with the weights compared to the models without weights. Having larger standard error in turn also leads to the models with survey weights having wider HPDIs. By including the adjusted and trimmed survey weights in the model, we do expect the posterior standard error to increase since including the weights decreases the sample size to the effective sample size. Naturally, with a smaller sample size the posterior standard error will be larger. The DIC for both spatial models is significantly lower when adjusted and trimmed weights are included, indicating a preference for the spatial models with survey weights.

Table 3.3 contains the results of the population prediction of BMI for all four models split by the eight counties included in our BMI survey data. The models were not fit to each county, rather the results were simply separated by county. The results for each county are similar to the overall results described from Table 3.2, however since the sample sizes are reduced when we group by county then the posterior standard errors will increase due to this smaller sample size. All of the counties have roughly the same sample size, except for County 3 which has an exceptionally large sample size of 795 observations, which accounts for about 43% of the total BMI data sample size. The remainder of the counties each have sample sizes ranging from 125 to 176 observations. The comparatively large sample size of County 3 yields smaller posterior standard errors compared to the other counties.

Table 3.3
County level BMI population prediction model comparison.

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$
County 1				
CAR	27.242	0.101	0.004	(27.061, 27.457)
SCAR	27.232	0.097	0.004	(27.044, 27.414)
Scott-Smith	27.194	0.149	0.005	(26.913, 27.481)
BHF	27.198	0.151	0.006	(26.891, 27.480)
County 2				
CAR	27.549	0.112	0.004	(27.308, 27.759)
SCAR	27.554	0.112	0.004	(27.336, 27.768)
Scott-Smith	27.491	0.178	0.006	(27.154, 27.852)
BHF	27.487	0.174	0.006	(27.138, 27.801)
County 3				
CAR	27.460	0.094	0.003	(27.268, 27.635)
SCAR	27.470	0.093	0.003	(27.292, 27.656)
Scott-Smith	27.424	0.140	0.005	(27.142, 27.701)
BHF	27.389	0.140	0.005	(27.100, 27.646)
County 4				
CAR	27.446	0.139	0.005	(27.178, 27.700)
SCAR	27.467	0.144	0.005	(27.164, 27.729)
Scott-Smith	27.462	0.212	0.008	(27.029, 27.842)
BHF	27.428	0.219	0.008	(27.008, 27.829)
County 5				
CAR	27.557	0.123	0.004	(27.310, 27.794)
SCAR	27.563	0.124	0.004	(27.329, 27.801)
Scott-Smith	27.551	0.188	0.007	(27.188, 27.939)
BHF	27.497	0.194	0.007	(27.125, 27.899)
County 6				
CAR	27.481	0.129	0.005	(27.236, 27.725)
SCAR	27.502	0.131	0.005	(27.260, 27.781)
Scott-Smith	27.408	0.205	0.007	(26.984, 27.766)
BHF	27.366	0.206	0.008	(26.967, 27.753)
County 7				
CAR	27.083	0.114	0.004	(26.856, 27.287)
SCAR	27.109	0.110	0.004	(26.892, 27.322)
Scott-Smith	27.113	0.163	0.006	(26.827, 27.450)
BHF	27.087	0.166	0.006	(26.769, 27.405)
County 8				
CAR	27.218	0.120	0.004	(26.983, 27.445)
SCAR	27.273	0.113	0.004	(27.050, 27.490)
Scott-Smith	27.205	0.163	0.006	(26.895, 27.503)
BHF	27.184	0.170	0.006	(26.885, 27.581)

Notes : SE = standard error; CV = coefficient of variation; HPDI = highest posterior density interval; CAR = conditional autoregressive; SCAR = simple conditional autoregressive; BHF = Battese, Harter and Fuller.

Table 3.4 contains the results of the population prediction of BMI for all four models with survey weights included split by the eight counties included in our BMI survey data. Again, the models were not fit to each county, and instead the results were simply separated by county. The results for each county are similar to the overall results described from Table 3.2. In Table 3.4, the sample sizes are being reduced by the inclusion of survey weights, in addition to the sample sizes being split by county. The posterior standard

errors continue to increase due to both factors reducing the sample sizes. Recall that all of the counties have roughly the same sample size, except for County 3 which has an exceptionally large sample size. This large sample size of County 3 yields the smallest posterior standard errors compared to the other counties.

Table 3.4
Including survey weights in county level BMI population prediction.

	Predicted $\bar{Y}$	SE of $\bar{Y}$	CV of $\bar{Y}$	95% HPDI of $\bar{Y}$
County 1				
CAR	27.158	0.114	0.004	(26.949, 27.411)
SCAR	27.185	0.111	0.004	(26.970, 27.401)
Scott-Smith	27.347	0.169	0.006	(27.011, 27.666)
BHF	27.202	0.175	0.006	(26.888, 27.575)
County 2				
CAR	26.945	0.125	0.005	(26.709, 27.202)
SCAR	27.021	0.117	0.004	(26.806, 27.259)
Scott-Smith	27.275	0.164	0.006	(26.950, 27.575)
BHF	27.416	0.200	0.007	(27.048, 27.814)
County 3				
CAR	27.196	0.102	0.004	(26.999, 27.383)
SCAR	27.199	0.104	0.004	(27.014, 27.417)
Scott-Smith	27.362	0.145	0.005	(27.070, 27.631)
BHF	27.329	0.171	0.006	(26.987, 27.634)
County 4				
CAR	26.983	0.122	0.005	(26.749, 27.229)
SCAR	26.990	0.123	0.005	(26.734, 27.206)
Scott-Smith	27.268	0.165	0.006	(26.979, 27.607)
BHF	27.352	0.253	0.009	(26.852, 27.816)
County 5				
CAR	27.162	0.118	0.004	(26.932, 27.391)
SCAR	27.171	0.117	0.004	(26.943, 27.400)
Scott-Smith	27.284	0.174	0.006	(26.953, 27.633)
BHF	27.393	0.224	0.008	(26.967, 27.859)
County 6				
CAR	27.142	0.115	0.004	(26.924, 27.379)
SCAR	27.135	0.118	0.004	(26.887, 27.344)
Scott-Smith	27.320	0.179	0.007	(26.972, 27.659)
BHF	27.298	0.233	0.009	(26.871, 27.761)
County 7				
CAR	26.935	0.121	0.004	(26.706, 27.180)
SCAR	26.923	0.120	0.004	(26.694, 27.158)
Scott-Smith	27.107	0.174	0.006	(26.794, 27.477)
BHF	27.085	0.194	0.007	(26.725, 27.459)
County 8				
CAR	26.840	0.127	0.005	(26.590, 27.090)
SCAR	26.903	0.126	0.005	(26.669, 27.150)
Scott-Smith	27.054	0.171	0.006	(26.742, 27.403)
BHF	27.156	0.198	0.007	(26.774, 27.585)

Notes : SE = standard error; CV = coefficient of variation; HPDI = highest posterior density interval; CAR = conditional autoregressive; SCAR = simple conditional autoregressive; BHF = Battese, Harter and Fuller.

3.2 Reduction in global pooling via spatial modeling

Our main goal is to make inference about the finite population mean without directly including the covariates in our models, which we have shown. A reason we chose to introduce a spatial component in our models is to have the posterior means of the individual strata result in less global pooling. We do not want the posterior mean of each individual stratum to simply approach the overall population posterior mean. Instead we would rather have strata with similar covariate attributes (i.e. neighbors in the $\mathbf{W}$ matrix) borrow strength from each other to have stratum posterior means gravitate towards the neighborhood posterior mean. Therefore, we study the $\boldsymbol{\mu}$ from the Scott-Smith model and the CAR model to show that by including the spatial component, as seen in the CAR model, we are able to increase the variability of the posterior predictions for $\boldsymbol{\mu}$. Since the results from the Scott-Smith model and the BHF model are similar we only use the Scott-Smith model in this comparison. In general, we wish to avoid using covariates in our models (as seen in the BHF model) when possible and the similar results from the Scott-Smith and BHF models are evidence that including the covariates in the model did not improve the prediction results. Similarly, since the CAR and SCAR models yield similar results and the SCAR model is a simpler version of the CAR model, then we proceed making this comparison using the CAR model.

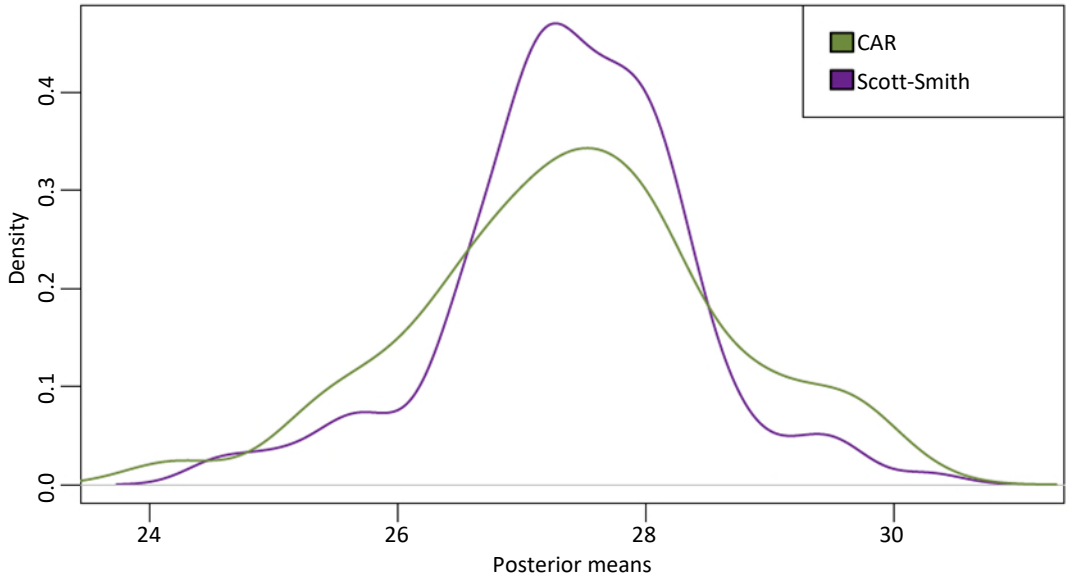
We are analyzing the $\boldsymbol{\mu}$ from each model, which means we have a sample from the posterior density of μ_i for each $i = 1,\ldots,\ell$. There are three main indications that illustrate how including the spatial component in the CAR model reduces global pooling when compared to the Scott-Smith model. First, the estimates of $\mu_1,\ldots,\mu_{132}$ from the CAR model have a standard deviation of 1.237, compared to the estimates from the Scott-Smith model which have a standard deviation of 0.967. From this we can already see increased variation in the $\boldsymbol{\mu}$ estimates in the CAR model, and this increased variation is a sign that less global pooling occurs in the CAR model as $\boldsymbol{\mu}$ contains more disperse values. Second, by looking at Figure 3.1 which contains the two kernel density curves for $\boldsymbol{\mu}$ from each model, we are able to see that $\boldsymbol{\mu}$ from the CAR model has a lower peak and heavier tails compared to the Scott-Smith model. The values of $\boldsymbol{\mu}$ in the CAR model are not converging to the overall population mean as aggressively as the Scott-Smith model.

Thirdly, we look at the shrinkage parameters in the posterior mean of $\boldsymbol{\mu}$ from each model. Recall that the posterior mean of μ_i for the Scott-Smith model in (A.3) is: $\left(\lambda_i \bar{y}_i + (1-\lambda_i)\,\theta\right)$ where $\lambda_i = n_i\rho \big/ \left((n_i - 1)\,\rho + 1\right)$ for $i = 1\ldots,\ell$. We can rewrite the posterior mean of $\boldsymbol{\mu}$ in the CAR model in (2.6) as: $\left(\Lambda\bar{\mathbf{y}} + (\mathbb{I} - \Lambda)\,\theta\mathbf{1}\right)$ where

$$\Lambda = \left(\operatorname{diag}\left(\frac{\sigma^2}{n_1}, \ldots, \frac{\sigma^2}{n_\ell} \right)^{-1} + \left(\frac{\rho}{1-\rho}\,\sigma^2 \left(\mathbf{R} - \gamma \mathbf{W} \right)^{-1} \right)^{-1} \right)^{-1} \operatorname{diag}\left(\frac{\sigma^2}{n_1}, \ldots, \frac{\sigma^2}{n_\ell} \right)^{-1}.$$

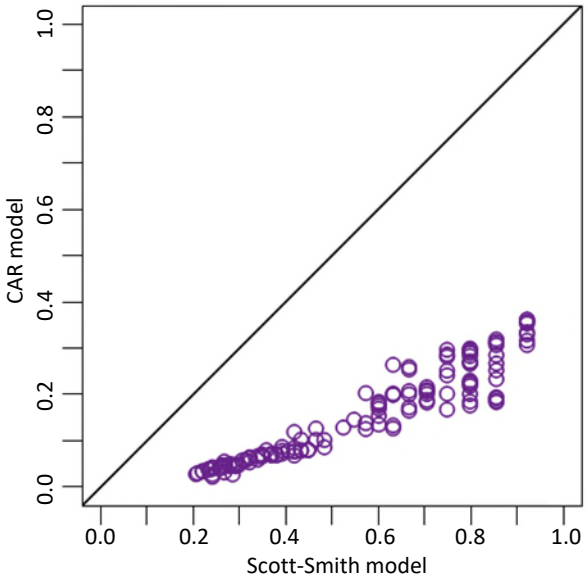
When the shrinkage parameters from the Scott-Smith model, $(1-\lambda_i)$, are greater than the sum of the row values of the shrinkage parameters from the CAR model, $(\mathbb{I} - \Lambda)$, then the non-spatial Scott-Smith model tends more towards the global pooling parameter, θ, instead of maintaining the characteristics of the individual strata. This is exactly what we see in Figure 3.2.

Figure 3.1 Comparing posterior distribution of μ.



Note : CAR = conditional autoregressive.

Figure 3.2 Comparing shrinkage parameters.



Note : CAR = conditional autoregressive.

Figure 3.2 shows that the Scott-Smith model puts significantly more weight on the global pooling parameter, θ, compared to the CAR model. This maintains our objective that the CAR model would result in less global pooling overall by including the neighborhood relationships. It is also important to point out that all of the shrinkage parameter values for both the Scott-Smith and the CAR models are in the range $[0,1]$.

4. Conclusion

Our main goal in introducing the spatial component of these models is to accommodate the covariates without using a regression model. In doing so, we also reduce the severity of global pooling and instead allow for neighbors with similar attributes to have predictions closer together. We have shown in our comparison of the $\boldsymbol{\mu}$ in the CAR and Scott-Smith models that including this spatial relationship of the strata does indeed limit the global pooling. This point was made by looking at the shrinkage parameters, the posterior densities, and the posterior variation of $\boldsymbol{\mu}$ from both models. The CAR and SCAR models both work well as small area estimation models that will reduce global pooling effects without defining the relationship between the response and the covariates. However, since it is important that ρ and γ are not too small in order to emphasize the spatial structure that accommodates the covariates, then we prefer the CAR model which has larger values of both ρ and γ.

In the CAR model, γ is close to unity which is a good sign that our spatial component will have more of an impact in the model compared to the much lower γ value in the SCAR model. As ρ and γ decrease, our posterior standard error of the population predictions will also decrease. We also presented how to use these spatial models we are advocating for with and without survey weights, and how to make population predictions in both cases. Ultimately, we are not interested in defining a relationship between the response variable $\mathbf{y}$ and $\mathbf{X}$ by having $\boldsymbol{\beta}$ in the model, as is the case in the BHF model. We avoid making strong assumptions about this relationship, and we maximize the number of potential applications our models can be applied to.

Future work includes continuing to work on this type of problem by adapting the models to cover the situation with a binary response variable instead of a continuous response variable. While the binary case is more computationally intense, it has a lot of useful applications. For an example related to the BMI application, we may be more interested in the proportion of individuals in the population who are obese (i.e. BMI ≥ 30), instead of predicting the overall BMI of the finite population. There are cases where knowing the proportion of a characteristic possessed by a population is more informative than knowing the average value of that characteristic in the population.

Appendix A

Scott-Smith model

In Appendix A we discuss the technical details of the Scott-Smith model (Scott and Smith, 1969). The adapted Scott-Smith model we use can be written as:

$$\begin{aligned} y_{ij} \mid \mu_i, \sigma^2 &\sim \text{Normal}\left(\mu_i, \sigma^2\right), \\ \mu_i \mid \theta, \rho, \sigma^2 &\sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \quad \text{(A.1)} \\ & j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell. \end{aligned}$$

Although the covariates are not present in the model, the responses are still grouped together using the covariate values, so each $\mathbf{y}_i$ has the same unique covariate combination for each stratum $i = 1, \ldots \ell$. Nandram, Toto and Choi (2011) has shown that ρ is a common intra-class correlation. Here μ_i follows a normal distribution:

$$\mu_i \,|\, \theta, \sigma^2, \rho, \mathbf{y} \;\sim\; \text{Normal}\left(\lambda_i \overline{y}_i + (1-\lambda_i)\,\theta, (1-\lambda_i)\,\rho\sigma^2 \big/ (1-\rho)\right) \tag{A.2}$$

where $\lambda_i = n_i\rho \big/ \left((n_i - 1)\,\rho + 1\right)$ for $i = 1 \ldots, \ell$.

The conditional posterior density of θ is:

$$\theta \,|\, \sigma^2, \rho, \mathbf{y} \;\sim\; \text{Normal}\left(\tilde{y}, \frac{\sigma^2\rho}{(1-\rho)\sum_{i=1}^{\ell}\lambda_i}\right). \tag{A.3}$$

Note that

$$\tilde{y} = \left(\sum_{i=1}^{\ell}\left(n_i \big/ \left((n_i - 1)\,\rho + 1\right)\right)\overline{y}_i\right) \Big/ \left(\sum_{i=1}^{\ell}\left(n_i \big/ \left((n_i - 1)\,\rho + 1\right)\right)\right)$$

is well defined for all $0 \le \rho \le 1$ and $l \ge 2$.

The conditional posterior density of σ^2 is:

$$\sigma^2 \,|\, \rho, \mathbf{y} \;\sim\; \text{InvGamma}\left(\frac{n-1}{2}, \left\{\sum_{i=1}^{\ell}(n_i - 1)\,s_i^2 \;+\; \frac{1-\rho}{\rho}\left(\sum_{i=1}^{\ell}\lambda_i\left(\overline{y}_i - \tilde{y}\right)^2\right)\right\} \Big/ 2\right). \tag{A.4}$$

Finally, once we integrate out σ^2 we are left with the nonstandard posterior density,

$$\begin{aligned}\pi_4\left(\rho \,|\, \mathbf{y}\right) \;\propto\; & (1-\rho)^{(l-2)/2}\sqrt{\frac{\prod_{i=1}^{\ell} n_i \big/ \left((n_i - 1)\,\rho + 1\right)}{\sum_{i=1}^{\ell} n_i \big/ \left((n_i - 1)\,\rho + 1\right)}} \\ & \times \frac{1}{\left\{1 + (1-\rho)\left(\sum_{i=1}^{\ell}\left(n_i \big/ \left((n_i - 1)\rho + 1\right)\right)\left(\overline{y}_i - \tilde{y}\right)^2\right) \Big/ \left(\sum_{i=1}^{\ell}(n_i - 1)\,s_i^2\right)\right\}^{(n-1)/2}}.\end{aligned} \tag{A.5}$$

This proves that the joint posterior density is proper, and this also shows how we can obtain a sample from the joint posterior density by sampling from $\pi_4\left(\rho \,|\, \mathbf{y}\right)$ first and then continuing to draw samples from their known distributions in reverse order (Nandram, Toto and Choi, 2011). Therefore, we begin by drawing samples of ρ from (A.5) using the grid method. Next, we use the sample of ρ we obtained to draw a sample of σ^2 directly from (A.4). Then we use the samples of ρ and σ^2 to draw a sample of θ from its standard distribution (A.3). Finally, we use the samples of ρ, σ^2, and θ to draw a sample of μ_i for $i = 1, \ldots, \ell$ from (A.2). Based on our samples from the posterior density and the observed values of $\mathbf{y}_i$, we make inference for the finite population mean $\overline{Y}_i$, using the model:

$$\overline{Y}_i \,|\, \mathbf{y}_i \;\overset{\text{ind}}{\sim}\; \text{Normal}\left(f_i\,\overline{y}_i + (1-f_i)\,\mu_i, (1-f_i)\,\frac{\sigma^2}{N_i}\right). \tag{A.6}$$

The results of this model are presented in Section 3.1 with an application using BMI data.

Including survey weights in Scott-Smith model

We also can include survey weights in the Scott-Smith model, and we use the same adjusted and trimmed survey weights described in Section 2.2. The Scott-Smith model with weights can be expressed by replacing the response variance in the first row of (A.1) from σ^2 to $\frac{\sigma^2}{a_{ij}^*}$, with a_{ij}^* from (2.17).

We use the same logic for obtaining a sample from this model with the adjusted weights as we used above in the model without weights. Making population predictions differs, because we obtain population predictions by:

$$\overline{Y}_i \,|\, \boldsymbol{\mu}, \sigma^2 \;\sim\; \text{Normal}\left(\mu_i, \frac{\sigma^2}{\hat{N}_i}\right) \qquad i = 1, \ldots, \ell, \tag{A.7}$$

where $\hat{N}_i = \sum_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each stratum $i = 1, \ldots, \ell$.

Appendix B

BHF model

In Appendix B we discuss the technical details of the Battese, Harter, and Fuller (BHF) model (Battese, Harter and Fuller, 1988). This non-spatial model introduces covariates and includes the random effects for each stratum. The BHF model is the only model used in this paper that specifies the relationship between the response and the covariates. In general, we want to avoid defining this relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$. The BHF model is:

$$\begin{aligned} y_{ij} \,|\, \mathbf{v}, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta} + v_i, \sigma^2\right), \\ \mathbf{v} \,|\, \rho, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \\ \pi\left(\boldsymbol{\beta}, \sigma^2, \rho\right) &\propto \frac{1}{\sigma^2}; \quad \sigma^2 > 0, \quad 0 < \rho < 1, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots \ell. \end{aligned} \tag{B.1}$$

Letting $\lambda_i = \rho n_i / ((1-\rho) + \rho n_i)$, then v_i follows a Normal distribution:

$$v_i \,|\, \boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y} \overset{\text{ind}}{\sim} \text{Normal}\left(\lambda_i\left(\overline{y}_i - \mathbf{x}_i'\boldsymbol{\beta}\right), \frac{(1-\lambda_i)\,\rho\sigma^2}{(1-\rho)}\right). \tag{B.2}$$

The conditional posterior density of $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} \,|\, \sigma^2, \rho, \mathbf{y} \;\sim\; \text{Normal}\left(\hat{\boldsymbol{\beta}}, \sigma^2\hat{\boldsymbol{\Sigma}}\right). \tag{B.3}$$

where

$$\hat{\boldsymbol{\beta}} \;=\; \hat{\boldsymbol{\Sigma}} \sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \overline{y}_i \mathbf{x}_i' \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} \;=\; \left(\sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \mathbf{x}_i \mathbf{x}_i'\right)^{-1}. \tag{B.4}$$

Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ are well defined for all ρ provided the design matrix $\mathbf{X}=\left(\mathbf{x}_i'\right)$ is full rank, where $\mathbf{x}_i'$ correspond to the rows of $\mathbf{X}$.

The conditional posterior density of σ^2 is:

$$\sigma^2 \mid \rho, \mathbf{y} \sim \text{InvGamma}\left(\frac{n-p}{2}, \frac{\sum_{i=1}^{\ell}\left[n_i\left(1-\lambda_i\right)\left(\bar{y}_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2+\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2\right]}{2}\right) \tag{B.5}$$

with $n=\sum_{i=1}^{\ell} n_i$. Therefore, after integrating out σ^2 we are finally left with the nonstandard posterior density of ρ:

$$\begin{aligned}\pi\left(\rho \mid \mathbf{y}\right) \propto & \det\left[\left(\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]^{1/2}\prod_{i=1}^{\ell}\left(1-\lambda_i\right)^{1/2} \\ & \times\left[\sum_{i=1}^{\ell} n_i\left(1-\lambda_i\right)\left(\bar{y}_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2+\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2\right]^{\frac{-n+p}{2}}.\end{aligned} \tag{B.6}$$

We are now able to directly obtain a sample from the joint posterior density by beginning with using the grid method to sample ρ. Then sampling σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\nu}$ is straight forward since each of these parameters has a standard form. Based on our samples from the posterior density and the observed values of $\mathbf{y}_i$, we make inference for the finite population mean $\bar{Y}_i$, using the model:

$$\bar{Y}_i \mid \mathbf{y}_i \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\bar{y}_i+\left(1-f_i\right)\left[\mathbf{x}_i'\boldsymbol{\beta}+\nu_i\right], \left(1-f_i\right)\frac{\sigma^2}{N_i}\right). \tag{B.7}$$

We explore the performance of this model in Section 3.1 with an application using BMI data.

Including survey weights in BHF model

We also can include survey weights in the BHF model using the same adjusted and trimmed survey weights described in Section 2.2. The BHF model with weights can be expressed by replacing the response variance in the first row of (B.1) from σ^2 to $\frac{\sigma^2}{a_{ij}^*}$, with a_{ij}^* from (2.17).

We use the same logic for obtaining a sample from this model with the adjusted weights as we used above in the model without weights. Making population predictions differs, because we obtain population predictions by:

$$\bar{Y}_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \nu_i, \sigma^2 \sim \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta}+\nu_i, \frac{\sigma^2}{\hat{N}_i}\right) \qquad i=1,\ldots,\ell, \tag{B.8}$$

where $\hat{N}_i=\sum_{j=1}^{n_i} v_{ij}$ represents the Horvitz-Thompson estimator of population size for each stratum $i=1,\ldots,\ell$.

References

Albert, J.H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669-679. https://doi.org/10.2307/2290350.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6), 1152-1174. http://www.jstor.org/stable/2958336.

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

Blackwell, D., and MacQueen, J.B. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2), 353-355. http://www.jstor.org/stable/2958020.

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Longman Higher Education. ISBN 10: 0201006227/ISBN 13: 9780201006223.

Chipman, H.A., George, E.I. and McCulloch, R.E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443), 935-948. https://doi.org/10.2307/2669832.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298. http://www.jstor.org/stable/27801587.

Chung, H.C., and Datta, G.S. (2022). Bayesian spatial models for estimating means of sampled and non-sampled small areas. *Survey Methodology*, 48, 2, 463-489. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00012-eng.pdf.

Datta, G., and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics*, 19(4), 1748-1770. https://doi.org/10.1214/aos/1176348369.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93(441), 273-282. https://doi.org/10.2307/2669623.

Hill, J., Linero, A. and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and its Application*. 7, 251-278. https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-031219-041110.

Lindley, D.V., and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society,* Series B (Methodological), 34(1), 1-41. http://www.jstor.org/stable/2985048.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, (Eds., S.K. Upadhyay, U. Singh and D. Dey), Anamaya, New Delhi, Chapter 25, 356-374.

Nandram, B., and Choi, J.W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Survey Methodology*, 31, 1, 73-84. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005001/article/8089-eng.pdf.

Nandram, B., and Rao, J.N.K. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1568-1603. http://www.asasrms.org/Proceedings/y2021/files/1912256.pdf.

Nandram, B., Toto, M.C. and Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81, 1593-1608. https://www.semanticscholar.org/paper/A-Bayesian-benchmarking-of-the-Scott%E2%80%93Smith-model-Nandram-Toto/a18cd37adaea51d06e81b2b525f61526d028fd73.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, Wiley Series in Survey Methodology.

Ritter, C., and Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419), 861-868. https://doi.org/10.2307/2290225.

Scott, A., and Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 101, 1387-1397.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566-1581. http://www.jstor.org/stable/27639773.

Yang, L., Nandram, B. and Choi, J.W. (2023). Bayesian predictive inference under nine methods for incorporating survey weights. *International Journal of Statistics and Probability*, l2, 1. https://ccsenet.org/journal/index.php/ijsp/article/view/0/48223.

Yin, J., and Nandram, B. (2020). A Bayesian small area model with Dirichlet processes on the responses. *Statistics in Transition New Series*, ISSN 2450-0291, Exeley, New York, 21, 3, 1-19, https://doi.org/10.21307/stattrans-2020-041.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 487-511
Statistics Canada, Catalogue No. 12-001-X

Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata

Jacek Wesołowski, Robert Wieczorkowski and Wojciech Wójciak[1]

Abstract

The optimum sample allocation in stratified sampling is one of the basic issues of survey methodology. It is a procedure of dividing the overall sample size into strata sample sizes in such a way that for given sampling designs in strata the variance of the *stratified π estimator* of the population total (or mean) for a given study variable assumes its minimum. In this work, we consider the optimum allocation of a sample, under lower and upper bounds imposed jointly on sample sizes in strata. We are concerned with the variance function of some generic form that, in particular, covers the case of the *simple random sampling without replacement* in strata. The goal of this paper is twofold. First, we establish (using the Karush-Kuhn-Tucker conditions) a generic form of the optimal solution, the so-called optimality conditions. Second, based on the established optimality conditions, we derive an efficient recursive algorithm, named *RNABOX*, which solves the allocation problem under study. The *RNABOX* can be viewed as a generalization of the classical recursive Neyman allocation algorithm, a popular tool for optimum allocation when only upper bounds are imposed on sample strata-sizes. We implement *RNABOX* in `R` as a part of our package `stratallo` which is available from the Comprehensive R Archive Network (CRAN) repository.

Key Words: Neyman allocation; Optimum allocation under box constraints; Optimum sample allocation; Recursive Neyman algorithm; Stratified sampling.

1. Introduction

Let us consider a finite population U of size N. Suppose the parameter of interest is the population total t of a variable y in U, i.e. $t=\sum_{k\in U}y_k$, where y_k denotes the value of y for population element $k\in U$. To estimate t, we consider the *stratified sampling* with the π *estimator*. Under this well-known sampling technique, population U is stratified, i.e. $U=\bigcup_{h\in\mathcal{H}}U_h$, where $U_h, h\in\mathcal{H}$, called strata, are disjoint and non-empty, and $\mathcal{H}$ denotes a finite set of strata labels. The size of stratum U_h is denoted $N_h, h\in\mathcal{H}$ and clearly $\sum_{h\in\mathcal{H}}N_h=N$. Probability samples $s_h\subseteq U_h$ of size $n_h\le N_h, h\in\mathcal{H}$, are selected independently from each stratum according to chosen sampling designs which are often of the same type across strata. The resulting total sample is of size $n=\sum_{h\in\mathcal{H}}n_h\le N$. It is well known that the *stratified* π *estimator* $\hat{t}_\pi$ of t and its variance are expressed in terms of the first and second order inclusion probabilities (see, e.g. Särndal, Swensson and Wretman, 1992, Result 3.7.1, page 102). In particular, for several important sampling designs

$$\mathrm{Var}\left(\hat{t}_\pi\right)=\sum_{h\in\mathcal{H}}\frac{A_h^2}{n_h}-B, \tag{1.1}$$

1. Jacek Wesołowski, Programming, Coordination of Statistical Surveys and Registers Department, Statistics Poland, Aleja Niepodległości, 208, 00-925 Warsaw, Poland, and Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland. E-mail: jacek.wesolowski@pw.edu.pl; Robert Wieczorkowski, Programming, Coordination of Statistical Surveys and Registers Department, Statistics Poland, Aleja Niepodległości 208, 00-925 Warsaw, Poland. E-mail: R.Wieczorkowski@stat.gov.pl; Wojciech Wójciak, Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland. E-mail: wojciech.wojciak.dokt@pw.edu.pl.

where $A_h > 0, B$ do not depend on $n_h, h \in \mathcal{H}$. Among the most basic and common sampling designs that give rise to the variance of the form (1.1) is the *simple random sampling without replacement* in strata (abbreviated *STSI*). In this case, the *stratified* π *estimator* of t assumes the form

$$\hat{t}_\pi = \sum_{h\in\mathcal{H}} \tfrac{N_h}{n_h} \sum_{k\in s_h} y_k, \tag{1.2}$$

which yields in (1.1): $A_h = N_h S_h$, where S_h denotes stratum standard deviation of study variable y, $h \in \mathcal{H}$, and $B = \sum_{h\in\mathcal{H}} N_h S_h^2$ (see, e.g. Särndal et al., 1992, Result 3.7.2, page 103).

The classical problem of optimum sample allocation is formulated as the determination of the allocation vector $\mathbf{n} = (n_h, h \in \mathcal{H})$ that minimizes the variance (1.1), subject to $\sum_{h\in\mathcal{H}} n_h = n$, for a given $n \le N$ (see, e.g. Särndal et al., 1992, Section 3.7.3, page 104). In this paper, we are interested in the classical optimum sample allocation problem with additional two-sided constraints imposed on sample sizes in strata. We phrase this problem in the language of mathematical optimization as Problem 1.1.

Problem 1.1. Given a finite set $\mathcal{H} \neq \varnothing$ and numbers $A_h > 0, m_h, M_h, n,$ such that $0 < m_h < M_h \le N_h, h \in \mathcal{H}$ and $\sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h$,

$$\underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\in\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad \sum_{h\in\mathcal{H}} \tfrac{A_h^2}{x_h} \tag{1.3}$$

$$\text{subject to} \quad \sum_{h\in\mathcal{H}} x_h = n \tag{1.4}$$

$$m_h \le x_h \le M_h, \quad h \in \mathcal{H}. \tag{1.5}$$

To emphasize the fact that the optimal solution to Problem 1.1 may not be an integer one, we denote the optimization variable by $\mathbf{x}$, not by $\mathbf{n}$. The assumptions about $n, m_h, M_h, h \in \mathcal{H},$ ensure that Problem 1.1 is feasible.

The upper bounds imposed on $x_h, h \in \mathcal{H},$ are natural since for instance the solution with $x_h > N_h$ for some $h \in \mathcal{H}$ is impossible. The lower bounds are necessary e.g. for estimation of population strata variances $S_h^2, h \in \mathcal{H}$. They also appear when one treats strata as domains and assigns upper bounds for variances of estimators of totals in domains. Such approach was considered e.g. in Choudhry, Rao and Hidiroglou (2012), where apart of the upper bounds constraints $x_h \le N_h, h \in \mathcal{H},$ the additional constraints $\left(\frac{1}{x_h} - \frac{1}{N_h}\right) N_h^2 S_h^2 \le R_h$, $h \in \mathcal{H},$ where $R_h, h \in \mathcal{H},$ are given constants, have been imposed. Obviously, the latter system of inequalities can be rewritten as lower bounds constraints of the form $x_h \ge m_h = \frac{N_h^2 S_h^2}{R_h + N_h S_h^2}, h \in \mathcal{H}$. The solution given in Choudhry et al. (2012) was obtained by the procedure based on the Newton-Raphson algorithm, a general-purpose root-finding numerical method. See also a related paper by Wright, Noble and Bailer (2007), where the optimum allocation problem under the constraint of the equal precision for estimation of the strata means was considered.

It is convenient to introduce the following definition for feasible solutions of Problem 1.1.

Definition 1.1. *Any vector* $\mathbf{x}=(x_h, h\in\mathcal{H})$ *satisfying (1.4) and (1.5) will be called an* allocation.

An allocation $\mathbf{x}=(x_h, h\in\mathcal{H})$ *is called a* vertex *one if and only if*

$$x_h=\begin{cases} m_h, & h\in\mathcal{L} \\ M_h, & h\in\mathcal{U}, \end{cases}$$

where $\mathcal{L},\mathcal{U}\subseteq\mathcal{H}$ *are such that* $\mathcal{L}\cup\mathcal{U}=\mathcal{H}$ *and* $\mathcal{L}\cap\mathcal{U}=\varnothing$.

An allocation which is not a vertex *one will be called a* regular *allocation.*

The solution to Problem 1.1 will be called the optimum *allocation.*

Note that an optimum allocation may be of a *vertex* or of a *regular* form. The name *vertex* allocation refers to the fact that in this case $\mathbf{x}$ is a vertex of the hyper-rectangle $\times_{h\in\mathcal{H}}[m_h, M_h]$. We note that Problem 1.1 becomes trivial if $n=\sum_{h\in\mathcal{H}}m_h$ or $n=\sum_{h\in\mathcal{H}}M_h$. In the former case, the solution is $\mathbf{x}^*=(m_h, h\in\mathcal{H})$, and in the latter $\mathbf{x}^*=(M_h, h\in\mathcal{H})$. These two are boundary cases of the *vertex* allocation. In real surveys with many strata, a *vertex* optimum allocation rather would not be expected. Nevertheless, for completeness we also consider such a case in Theorem 3.1, which describes the form of the optimum allocation vector. We also note that a *regular* optimum allocation $\mathbf{x}^*\in\times_{h\in\mathcal{H}}(m_h, M_h)$ if and only if it is the classical Tschuprow-Neyman allocation

$$\mathbf{x}^*=\left(A_h\frac{n}{\sum_{v\in\mathcal{H}}A_v}, h\in\mathcal{H}\right)$$

(see Neyman, 1934; Tschuprow, 1923).

The rest of this paper is structured as follows. Section 2 presents motivations for this research as well as a brief review of the literature. In Section 3, we identify Problem 1.1 as a convex optimization problem and then use the Karush-Kuhn-Tucker conditions to establish necessary and sufficient conditions for a solution to optimization Problem 1.1. These conditions, called the optimality conditions, are presented in Theorem 3.1. In Section 4, based on these optimality conditions, we introduce a new algorithm, *RNABOX*, and prove that it solves Problem 1.1 (see Theorem 4.1). The name *RNABOX* refers to the fact that this algorithm generalizes the recursive Neyman algorithm, denoted here *RNA*. The *RNA* is a well-established allocation procedure, commonly used in everyday survey practice. It finds a solution to the allocation Problem 2.1 (see below), which is a relaxed version of Problem 1.1. In Section 5, we discuss numerical experiments related to computational efficiency of the *RNABOX* algorithm and the *fixed-point iteration* algorithm from Münnich, Sachs and Wagner (2012). A concise summary of the results is given in Section 6, where we also briefly comment on some of the key aspects of rounding of non-integer optimum allocations. Auxiliary remarks and lemmas as well as proofs of both theorems are placed in the Appendix.

Finally, let us note that the implementation of *RNABOX* algorithm is available through our `R` package `stratallo` (Wójciak, 2023b), which is published in CRAN repository (R Core Team, 2023).

2. Motivation and literature review

An abundant body of literature is devoted to the problem of optimum sample allocation, going back to classical solution of Tschuprow (1923) and Neyman (1934), dedicated to *STSI* sampling without taking inequality constraints (1.5) into account. In spite of this fact, a thorough analysis of the literature shows that Problem 1.1 has not been completely understood yet and it suffers from the lack of fully satisfactory algorithms.

Below, we briefly review existing methods for solving Problem 1.1, including methods that provide integer-valued solutions.

2.1 Not-necessarily integer-valued allocation

An approximate solution to Problem 1.1 can be achieved through generic methods of non-linear programming (NLP) (see, e.g. the monograph Valliant, Dever and Kreuter, 2018, and references therein). These methods have been involved in the problem of optimum sample allocation since solving the allocation problem is equivalent to finding the extreme (namely, stationary points) of a certain objective function over a feasible set. Knowing the (approximate) extreme of the objective function, one can determine (approximate, yet typically sufficiently accurate) sizes of samples allocated to individual strata.

In a similar yet different approach adopted e.g. in Münnich et al. (2012), Problem 1.1 is transformed into root-finding or fixed-point-finding problems (of some properly defined function) to which the solution is obtained by general-purpose algorithms like e.g. *bisection* or *regula falsi*.

Algorithms used in both these approaches would in principle have infinitely many steps, and are stopped by an arbitrary decision, typically related to the precision of the iterates. There are two main weaknesses associated with this way of operating: failure of the method to converge or slow convergence towards the optimal solution for some poor starting points. In other words, performance of these algorithms may strongly depend on an initial choice of a starting point, and such a choice is almost always somewhat hazardous. As an example consider the *fixed-point iteration* algorithm (*FPIA*), of Münnich et al. (2012). For a population with 4 strata, such that $A_1 = 380$, $A_2 = 140$, $A_3 = 230$, $A_4 = 1{,}360$, bounds $m_h = 10$, $M_h = 50$, $h \in \{1, 2, 3, 4\}$, total sample size $n = 80$, and for starting point $\lambda_0 = 695.64$ (chosen in the way suggested in that paper), the *FPIA* does not convergence due to oscillations around the optimal solution. Another drawback of the algorithms of this type is their sensitivity to finite precision arithmetic issues that can arise in case when the stopping criterion is not expressed directly in terms of the allocation vector iterates (which is often the case).

Contrary to that, in the recursive algorithms (we are concerned with), the optimal solution is always found by recursive search of feasible candidates for the optimum allocation among subsets of $\mathcal{H}$. Hence, they stop always at the exact solution and after finitely many iterations (not exceeding the number of strata + 1, as we will see for the case of *RNABOX* in the proof of Theorem 4.1). An important example of such an algorithm, is the recursive Neyman algorithm, *RNA*, dedicated for Problem 2.1, a relaxed version of Problem 1.1.

Problem 2.1. Given a finite set $\mathcal{H} \neq \emptyset$ and numbers $A_h > 0, M_h, n > 0,$ such that $0 < M_h \leq N_h, h \in \mathcal{H}$ and $n \leq \sum_{h\in\mathcal{H}} M_h,$

$$\begin{aligned} \underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\in\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimize}} \quad & \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} \\ \text{subject to} \quad & \sum_{h\in\mathcal{H}} x_h = n \\ & x_h \leq M_h, \quad h \in \mathcal{H}. \end{aligned}$$

Although *RNA* is popular among practitioners, a formal proof of the fact that it gives the optimal solution to Problem 2.1 has been given only recently in Wesołowski, Wieczorkowski and Wójciak (2022). For other recursive approaches to Problem 2.1, see also e.g. Stenger and Gabler (2005), Kadane (2005).

To the best of our knowledge, the only non-integer recursive optimum allocation algorithm described in the literature that is intended to solve Problem 1.1 is the *noptcond* procedure proposed by Gabler, Ganninger and Münnich (2012). In contrary to *RNABOX*, this method in particular performs strata sorting. Unfortunately, the allocation computed by *noptcond* may not yield the minimum of the objective function (1.3). This fact can be illustrated with a short numerical example given in Table 2.1, which follows Wójciak (2019, Example 3.9).

Table 2.1
Two allocations for a population with two strata: Non-optimum $\mathbf{x}^{\text{noptcond}}$ and the optimum $\mathbf{x}^*$.

h	A_h	m_h	M_h	$\mathbf{x}^{\text{noptcond}}$	$\mathbf{x}^*$
1	2,000	30	50	30	50
2	3,000	40	200	130	110
total sample size $n = 160$.					

2.2 Integer-valued allocation

Integer-valued algorithms dedicated to Problem 1.1 are proposed in Friedrich, Münnich, de Vries and Wagner (2015), Wright (2017, 2020). The multivariate version of the optimum sample allocation problem under box constraints in which $m_h = m, h \in \mathcal{H},$ for a given constant m, is considered in the paper of de Moura Brito, do Nascimento Silva, Silva Semaan and Maculan (2015). The proposed procedure that solves that problem, uses binary integer programming algorithm and can be applied to the univariate case. See also Brito, Silva and Veiga (2017) for the `R`-implementation of this approach.

Integer-valued allocation methods proposed in these papers are precise (not approximate) and theoretically sound. However, they are relatively slow, when compared with not-necessarily integer-valued algorithms. For instance, at least for one-sided constraints, the integer-valued algorithm *capacity scaling* of Friedrich et al. (2015) may be thousands of times slower than the *RNA* (see Wesołowski et al., 2022, Section 4). This seems to be a major drawback of these methods as the differences in variances of estimators based on integer-rounded non-integer optimum allocation and integer optimum allocation are negligible as explained in Section 6. The computational efficiency is of particular significance when the number of strata

is large, see, e.g. application to the German census in Burgard and Münnich (2012), and it becomes even more pronounced in iterative solutions to stratification problems, when the number of iterations may count in millions (see, e.g. Khan, Nand and Ahmad, 2008; Baillargeon and Rivest, 2011; Barcaroli, 2014; Gunning and Horgan, 2004; Lednicki and Wieczorkowski, 2003).

Having all that said, the search for a new, universal, theoretically sound and computationally effective recursive algorithms of optimum sample allocation under two-sided constraints on the sample strata-sizes, is crucial both for theory and practice of survey sampling.

3. Optimality conditions

In this section we establish optimality conditions, that is, a general form of the solution to Problem 1.1. As it will be seen in Section 4, these optimality conditions are crucial for the construction of *RNABOX* algorithm.

Before we establish necessary and sufficient optimality conditions for a solution to optimization Problem 1.1, we first define a set function s, which considerably simplifies notation and calculations.

Definition 3.1. Let $\mathcal{H}, n, A_h > 0, m_h, M_h, h \in \mathcal{H}$ be as in Problem 1.1 and let $\mathcal{L}, \mathcal{U} \subseteq \mathcal{H}$ be such that $\mathcal{L} \cap \mathcal{U} = \varnothing, \mathcal{L} \cup \mathcal{U} \subsetneq \mathcal{H}$. The set function s is defined as

$$s(\mathcal{L}, \mathcal{U}) = \frac{n - \sum_{h \in \mathcal{L}} m_h - \sum_{h \in \mathcal{U}} M_h}{\sum_{h \in \mathcal{H} \setminus (\mathcal{L} \cup \mathcal{U})} A_h}. \tag{3.1}$$

Below, we will introduce the $\mathbf{x}^{(\mathcal{L},\mathcal{U})}$ vector for disjoint $\mathcal{L}, \mathcal{U} \subseteq \mathcal{H}$. It appears that the solution of the Problem 1.1 is necessarily of the form (3.2) with sets $\mathcal{L}$ and $\mathcal{U}$ defined implicitly through systems of equations/inequalities established in Theorem 3.1.

Definition 3.2. Let $\mathcal{H}, n, A_h > 0, m_h, M_h, h \in \mathcal{H}$ be as in Problem 1.1, and let $\mathcal{L}, \mathcal{U} \subseteq \mathcal{H}$ be such that $\mathcal{L} \cap \mathcal{U} = \varnothing$. We define the vector $\mathbf{x}^{(\mathcal{L},\mathcal{U})} = \left(x_h^{(\mathcal{L},\mathcal{U})}, h \in \mathcal{H}\right)$ as follows

$$x_h^{(\mathcal{L},\mathcal{U})} = \begin{cases} m_h, & h \in \mathcal{L} \\ M_h, & h \in \mathcal{U} \\ A_h\, s(\mathcal{L}, \mathcal{U}), & h \in \mathcal{H} \setminus (\mathcal{L} \cup \mathcal{U}). \end{cases} \tag{3.2}$$

The following Theorem 3.1 characterizes the form of the optimal solution to Problem 1.1 and therefore is one of the key results of this paper.

Theorem 3.1 (Optimality conditions)**.** *The optimization Problem 1.1 has a unique optimal solution. Point* $\mathbf{x}^* \in \mathbb{R}_+^{|\mathcal{H}|}$ *is a solution to optimization Problem 1.1 if and only if* $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$, *with disjoint* $\mathcal{L}^*, \mathcal{U}^* \subseteq \mathcal{H}$, *such that one of the following two cases holds:*

CASE I: $\mathcal{L}^* \cup \mathcal{U}^* \subsetneq \mathcal{H}$ *and*

$$\begin{aligned} \mathcal{L}^* &= \left\{h \in \mathcal{H}\colon\, s\left(\mathcal{L}^*, \mathcal{U}^*\right) \le \tfrac{m_h}{A_h}\right\}, \\ \mathcal{U}^* &= \left\{h \in \mathcal{H}\colon\, s\left(\mathcal{L}^*, \mathcal{U}^*\right) \ge \tfrac{M_h}{A_h}\right\}. \end{aligned} \tag{3.3}$$

CASE II: $\mathcal{L}^* \cup \mathcal{U}^* = \mathcal{H}$ *and*

$$\max_{h \in \mathcal{U}^*} \tfrac{M_h}{A_h} \le \min_{h \in \mathcal{L}^*} \tfrac{m_h}{A_h}, \qquad \text{if } \mathcal{U}^* \ne \varnothing \text{ and } \mathcal{L}^* \ne \varnothing, \tag{3.4}$$

$$\sum_{h \in \mathcal{L}^*} m_h + \sum_{h \in \mathcal{U}^*} M_h = n. \tag{3.5}$$

Remark 3.1. *The optimum allocation* $\mathbf{x}^*$ *is a* regular *one in CASE I and a* vertex *one in CASE II.*

The proof of Theorem 3.1 is given in Appendix A. Note that Theorem 3.1 describes the general form of the optimum allocation up to specification of *take-min* and *take-max* strata sets $\mathcal{L}^*$ and $\mathcal{U}^*$. The question how to identify sets $\mathcal{L}^*$ and $\mathcal{U}^*$ that determine the optimal solution $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*, \mathcal{U}^*)}$ is the subject of Section 4.

4. Recursive Neyman algorithm under box constraints

4.1 The *RNABOX* algorithm

In this section we introduce an algorithm solving Problem 1.1. In view of Theorem 3.1 its essential task is to split the set of all strata labels $\mathcal{H}$ into three subsets of *take-min* $\left(\mathcal{L}^*\right)$, *take-max* $\left(\mathcal{U}^*\right)$, and *take-Neyman* $\left(\mathcal{H} \setminus \left(\mathcal{L}^* \cup \mathcal{U}^*\right)\right)$. We call this new algorithm *RNABOX* since it generalizes existing algorithm *RNA* in the sense that *RNABOX* solves optimum allocation problem with simultaneous lower and upper bounds, while the *RNA* is dedicated for the problem with upper bounds only, i.e. for Problem 2.1. Moreover, *RNABOX* uses *RNA* in one of its interim steps. We first recall *RNA* algorithm and then present *RNABOX*. For more information on *RNA*, see Wesołowski et al. (2022, Section 2) or Särndal et al. (1992, Remark 12.7.1, page 466).

Algorithm *RNA*

Input:	$\mathcal{H}, \left(A_h\right)_{h \in \mathcal{H}}, \left(M_h\right)_{h \in \mathcal{H}}, n.$
Require:	$A_h > 0, M_h > 0, h \in \mathcal{H}, 0 < n \le \sum_{h \in \mathcal{H}} M_h.$
Step 1:	Set $\mathcal{U} = \varnothing.$
Step 2:	Determine $\tilde{\mathcal{U}} = \left\{h \in \mathcal{H} \setminus \mathcal{U}\colon\, A_h\, s\left(\varnothing, \mathcal{U}\right) \ge M_h\right\},$ where set function s is defined in (3.1).
Step 3:	If $\tilde{\mathcal{U}} = \varnothing,$ go to step 4. Otherwise, update $\mathcal{U} \leftarrow \mathcal{U} \cup \tilde{\mathcal{U}}$ and go to step 2.
Step 4:	Return $\mathbf{x}^* = \left(x_h^*, h \in \mathcal{H}\right)$ with $x_h^* = \begin{cases} M_h, & h \in \mathcal{U} \\ A_h s\left(\varnothing, \mathcal{U}\right), & h \in \mathcal{H} \setminus \mathcal{U}. \end{cases}$

Algorithm *RNABOX*

Input: $\mathcal{H}, (A_h)_{h\in\mathcal{H}}, (m_h)_{h\in\mathcal{H}}, (M_h)_{h\in\mathcal{H}}, n.$

Require: $A_h > 0, 0 < m_h < M_h, h \in \mathcal{H}, \sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h.$

Step 1: Set $\mathcal{L} = \varnothing.$

Step 2: Run $RNA\left[\mathcal{H}, (A_h)_{h\in\mathcal{H}}, (M_h)_{h\in\mathcal{H}}, n\right]$ to obtain $\left(x_h^{**}, h \in \mathcal{H}\right).$
Let $\mathcal{U} = \left\{h \in \mathcal{H}: x_h^{**} = M_h\right\}.$

Step 3: Determine $\tilde{\mathcal{L}} = \left\{h \in \mathcal{H} \setminus \mathcal{U}: x_h^{**} \le m_h\right\}.$

Step 4: If $\tilde{\mathcal{L}} = \varnothing$ go to step 5. Otherwise, update $n \leftarrow n - \sum_{h\in\tilde{\mathcal{L}}} m_h, \mathcal{H} \leftarrow \mathcal{H} \setminus \tilde{\mathcal{L}}, \ \mathcal{L} \leftarrow \mathcal{L} \cup \tilde{\mathcal{L}}$ and go to step 2.

Step 5: Return $\mathbf{x}^* = \left(x_h^*, h \in \mathcal{L} \cup \mathcal{H}\right)$ with $x_h^* = \begin{cases} m_h, & h \in \mathcal{L} \\ x_h^{**}, & h \in \mathcal{H}. \end{cases}$

We note that in real life applications numbers $(A_h)_{h\in\mathcal{H}}$ are typically unknown and therefore their estimates $\left(\hat{A}_h\right)_{h\in\mathcal{H}}$ are used instead in the input of the algorithms.

Theorem 4.1 is the main theoretical result of this paper and its proof is given in Appendix B.

Theorem 4.1. The RNABOX algorithm provides the optimal solution to Problem 1.1.

4.2 An example of performance of *RNABOX*

We demonstrate the operational behaviour of *RNABOX* algorithm for an artificial population with 10 strata and for total sample size $n =$ 5,110, as shown in Table 4.1.

Table 4.1
An example of *RNABOX* performance for a population with 10 strata and total sample size $n =$ 5,110.

h	A_h	m_h	M_h	$\mathcal{L}_1/\mathcal{U}_1$	$\mathcal{L}_2/\mathcal{U}_2$	$\mathcal{L}_3/\mathcal{U}_3$	$\mathcal{L}_4/\mathcal{U}_4$	$\mathcal{L}_5/\mathcal{U}_5$	$\mathcal{L}_6/\mathcal{U}_6$	x^*
1	2,700	750	900			□	□	□	□	750
2	2,000	450	500	■		□	□	□	□	450
3	4,200	250	300	■	■	■	■	■		261.08
4	4,400	350	400	■	■	■			□	350
5	3,200	150	200	■	■	■	■	■		198.92
6	6,000	550	600	■	■	■		□	□	550
7	8,400	650	700	■	■	■			□	650
8	1,900	50	100	■	■	■	■	■	■	100
9	5,400	850	900	■	■		□	□	□	850
10	2,000	950	1,000		□	□	□	□	□	950
SUM		5,000	5,600	0/8	1/7	3/6	4/3	5/3	7/1	5,110

Notes: Columns $\mathcal{L}_r/\mathcal{U}_r$, $r = 1, \ldots, 6,$ represent the content of sets $\mathcal{L}, \mathcal{U}$ respectively, in the r^{th} iteration of the *RNABOX* (between step 3 and step 4): symbols □ or ■ indicate that the stratum with label h is in $\mathcal{L}_r$ or $\mathcal{U}_r$, respectively.

For this example, *RNABOX* stops after 6 iterations with *take-min* strata set $\mathcal{L}^* = \{1, 2, 4, 6, 7, 9, 10\}$, *take-max* strata set $\mathcal{U}^* = \{8\}$ and *take-Neyman* strata set $\mathcal{H} \setminus \left(\mathcal{L}^* \cup \mathcal{U}^*\right) = \{3, 5\}$ (see column $\mathcal{L}_6/\mathcal{U}_6$). The

optimum allocation is a *regular* one and it is given in column x^* of Table 4.1. The corresponding value of the objective function (1.3) is 441,591.5. The details of interim allocations of strata to sets $\mathcal{L}, \mathcal{U}$ at each of 6 iterations of the algorithm are given in columns $\mathcal{L}_1 / \mathcal{U}_1 - \mathcal{L}_6 / \mathcal{U}_6$.

4.3 Possible modifications and improvements

Alternatives for *RNA* in step 2

The *RNABOX* algorithm uses *RNA* in its step 2. However, it is not hard to see that any algorithm dedicated to Problem 2.1 (like for instance *SGA* by Stenger and Gabler, 2005 or *COMA* by Wesołowski et al., 2022) could be used instead. We chose *RNA* as it allows to keep *RNABOX* free of any strata sorting.

A twin version of *RNABOX*

Let us observe that the order in which $\mathcal{L}$ and $\mathcal{U}$ are computed in the algorithm could be interchanged. Such a change, implies that the *RNA* used in step 2 of the *RNABOX*, should be replaced by its twin version, the *LRNA*, that solves optimum allocation problem under one-sided lower bounds. The *LRNA* is described in details in Wójciak (2023a).

Algorithm *LRNA*

Input: $\mathcal{H}, (A_h)_{h\in\mathcal{H}}, (m_h)_{h\in\mathcal{H}}, n.$

Require: $A_h > 0, m_h > 0, h \in \mathcal{H}, n \geq \sum_{h\in\mathcal{H}} m_h.$

Step 1: Set $\mathcal{L} = \varnothing.$

Step 2: Determine $\tilde{\mathcal{L}} = \left\{h \in \mathcal{H} \setminus \mathcal{L}\colon\ A_h\, s(\mathcal{L}, \varnothing) \leq m_h\right\},$ where set function s is defined in (3.1).

Step 3: If $\tilde{\mathcal{L}} = \varnothing,$ go to step 4. Otherwise, update $\mathcal{L} \leftarrow \mathcal{L} \cup \tilde{\mathcal{L}}$ and go to step 2.

Step 4: Return $\mathbf{x}^* = \left(x_h^*, h \in \mathcal{H}\right)$ with $x_h^* = \begin{cases} m_h, & h \in \mathcal{L} \\ A_h s(\mathcal{L}, \varnothing), & h \in \mathcal{H} \setminus \mathcal{L}. \end{cases}$

Taking into account the observation above, step 2 and step 3 of *RNABOX* would read:

Step 2: Run $LRNA\left[\mathcal{H}, (A_h)_{h\in\mathcal{H}}, (m_h)_{h\in\mathcal{H}}, n\right]$ to obtain $\left(x_h^{**}, h \in \mathcal{H}\right).$
Let $\mathcal{L} = \left\{h \in \mathcal{H}\colon\ x_h^{**} = m_h\right\}.$

Step 3: Determine $\tilde{\mathcal{U}} = \left\{h \in \mathcal{H} \setminus \mathcal{L}\colon\ x_h^{**} \geq M_h\right\}.$

The remaining steps should be adjusted accordingly.

Using prior information in *RNA* at step 2

In view of Lemma B.2, using the notation introduced in Appendix B.1, in step 2 of *RNABOX*, for $r^* \geq 2$ we have

$$\mathcal{U}_r = \left\{h \in \mathcal{H} \setminus \mathcal{L}_r\colon\ x_h^{**} = M_h\right\} \subseteq \mathcal{U}_{r-1}, \qquad r = 2, \ldots, r^*.$$

This suggests that the domain of discourse for $\mathcal{U}_r$ could be shrunk from $\mathcal{H} \setminus \mathcal{L}_r$ to $\mathcal{U}_{r-1} \subseteq \mathcal{H} \setminus \mathcal{L}_r$, i.e.

$$\mathcal{U}_r = \left\{ h \in \mathcal{U}_{r-1} \colon x_h^{**} = M_h \right\}, \qquad r = 2, \ldots, r^*. \tag{4.1}$$

Given the above observation and the fact that from the implementation point of view set $\mathcal{U}_r$ is determined internally by *RNA*, it is tempting to consider modification of *RNA* such that it makes use of the domain of discourse $\mathcal{U}_{r-1}$ for set $\mathcal{U}_r$. This domain could be specified as an additional input parameter, say $\mathcal{J} \subseteq \mathcal{H}$, and then step 2 of *RNA* algorithm would read:

Step 2: Determine $\tilde{\mathcal{U}} = \left\{ h \in \mathcal{J} \setminus \mathcal{U} \colon A_h \, s(\varnothing, \mathcal{U}) \geq M_h \right\}$.

From *RNABOX* perspective, this new input parameter of *RNA* should be set to $\mathcal{J} = \mathcal{H}$ for the first iteration, and then $\mathcal{J} = \mathcal{U}_{r-1}$ for subsequent iterations $r = 2, \ldots, r^* \geq 2$ (if any).

5. Numerical results

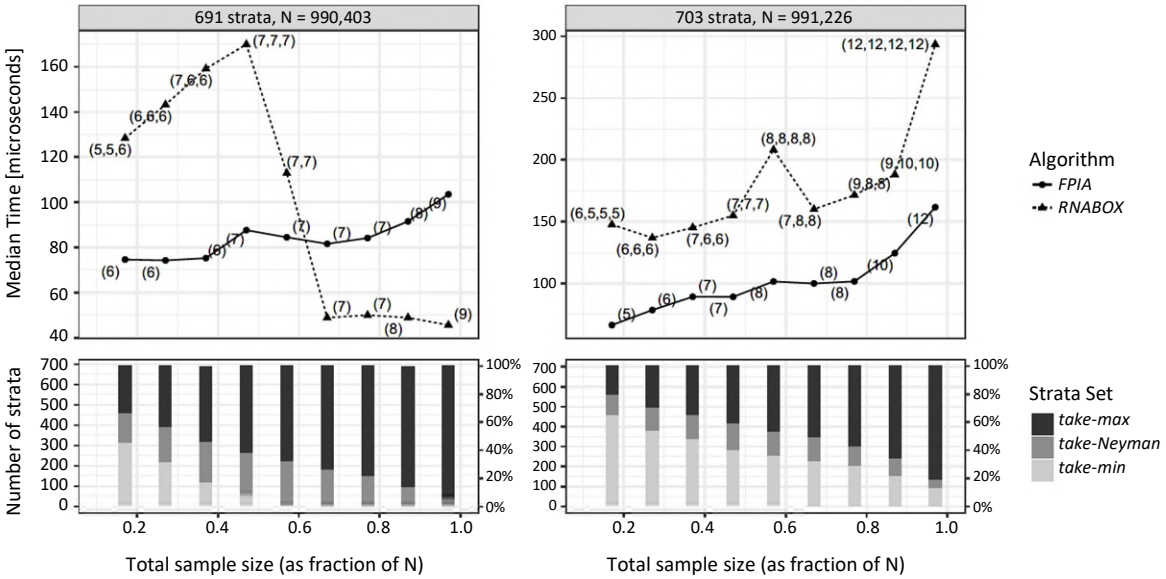
In simulations, using `R` Statistical Software (R Core Team, 2023) and `microbenchmark R` package (Mersmann, 2021), we compared the computational efficiency of *RNABOX* algorithm with the efficiency of the *FPIA* of Münnich et al. (2012). The latter one is known to be an efficient algorithm dedicated to Problem 1.1 and therefore we used it as a benchmark. The comparison was not intended to verify theoretical computational complexity, but was rather concerned with quantitative results regarding computational efficiency for the specific implementations of both algorithms.

To compare the performance of the algorithms we used the *STSI* sampling for several artificially created populations. Here, we chose to report on simulation for two such populations with 691 and 703 strata, results of which are representative for the remaining ones. These two populations were constructed by iteratively binding K sets of numbers, where K equals 100 (for the first population) and 200 (for the second population). Each set, labelled by $i = 1, \ldots, K$, contains 10,000 random numbers generated independently from log-normal distribution with parameters $\mu = 0$ and $\sigma = \log(1+i)$. For every set $i = 1, \ldots, K$, strata boundaries were determined by the geometric stratification method of Gunning and Horgan (2004) with parameter 10 being the number of strata and targeted coefficient of variation equal to 0.05. This stratification method is implemented in the `R` package `stratification`, developed by Rivest and Baillargeon (2022) and described in Baillargeon and Rivest (2011). For more details, see the `R` code with the experiments, which is placed in our GitHub repository (see Wieczorkowski, Wesołowski and Wójciak, 2023).

Results of these simulations are illustrated in Figure 5.1. From Figure 5.1 we see that, while for majority of the cases the *FPIA* is slightly faster than *RNABOX*, the running times of both of these algorithms are generally comparable. The gain in the execution time of the *FPIA* results from the fact that it typically runs through a smaller number of sets $\mathcal{L}, \mathcal{U} \subset \mathcal{H}$, than *RNABOX* in order to find the optimal $\mathcal{L}^*$ and $\mathcal{U}^*$. Although this approach usually gives correct results (as in the simulations reported in this section), it may happen that the *FPIA* misses the optimal sets $\mathcal{L}^*, \mathcal{U}^* \subsetneq \mathcal{H}$ and therefore it may not end up with the correct optimum allocation. Such a rare case was illustrated with a numerical example given in Section 2.1. As a

side note we point out that the *FPIA* is not well-defined when the optimum allocation is of a *vertex* type, i.e. when $\mathcal{L}^* \cup \mathcal{U}^* = \mathcal{H}$.

Figure 5.1 Running times of *FPIA* and *RNABOX* for two artificial populations.



Notes: Top graphs show the empirical median of execution times (calculated from 100 repetitions) for different total sample sizes. Numbers in brackets are the numbers of iterations of a given algorithm. In the case of *RNABOX*, it is a vector with number of iterations of the *RNA* (see step 2 of *RNABOX*) for each iteration of *RNABOX*. Thus, the length of this vector is equal to the number of iterations of *RNABOX*. Counts of *take-min*, *take-Neyman*, and *take-max* strata are shown on bottom graphs.

6. Concluding comments

In this paper we considered Problem 1.1 of optimum sample allocation under box constraints. The main result of this work is the mathematically precise formulation of necessary and sufficient conditions for the solution to Problem 1.1, given in Theorem 3.1, as well as the development of the new recursive algorithm, termed *RNABOX*, that solves Problem 1.1. The optimality conditions are fundamental to analysis of the optimization problem. They constitute trustworthy underlay for development of effective algorithms and can be used as a baseline for any future search of new algorithms solving Problem 1.1. Essential properties of *RNABOX* algorithm, that distinguish it from other existing algorithms and approaches to the Problem 1.1, are:

1. Universality: *RNABOX* provides optimal solution to every instance of feasible Problem 1.1 (including the case of a *vertex* optimum allocation).
2. No initialization issues: *RNABOX* does not require any initializations, pre-tests or whatsoever that could have an impact on the final results of the algorithm. This, in turn, takes places e.g. in case of NLP methods.
3. No sorting: *RNABOX* does not perform any ordering of strata.
4. Computational efficiency: *RNABOX* running time is comparable to that of *FPIA* (which is probably the fastest previously known optimum allocation algorithm for the problem considered).

5. Directness: *RNABOX* computes important quantities (including *RNA* internals) via formulas that are expressed directly in terms of the allocation vector $\mathbf{x}^{(\mathcal{L},\mathcal{U})}$ (see Definition 3.2). This reduces the risk of finite precision arithmetic issues, comparing to the algorithms that base their key operations on some interim variables on which the optimum allocation depends, as is the case of e.g. the NLP-based method.
6. Recursive nature: *RNABOX* repeatedly applies allocation step 2 and step 3 to step-wise reduced set of strata, i.e. "smaller" versions of the same problem. This translates to clarity of the routines and a natural way of thinking about the allocation problem.
7. Generalization: *RNABOX*, from the perspective of its construction, is a generalization of the popular *RNA* algorithm that solves Problem 2.1 of optimum sample allocation under one-sided bounds on sample strata sizes.

Finally, we would like to note that Problem 1.1 considered in this paper is not an integer-valued allocation problem, while the sample sizes in strata should be of course natural numbers. On the other hand, the integer-valued optimum allocation algorithms are relatively slow and hence might be inefficient in some applications, as already noted in Section 2. If the speed of an algorithm is of concern and non-necessarily integer-valued allocation algorithm is chosen (e.g. *RNABOX*), the natural remedy is to round the non-integer optimum allocation provided by that algorithm. Altogether, such procedure is still much faster than integer-valued allocation algorithms. However, a simple rounding of the non-integer solution does not, in general, yield the minimum of the objective function, and may even lead to an infeasible solution, as noted in Friedrich et al. (2015, Section 1, page 3). Since infeasibility can in fact arise only from violating constraint (1.4), it can be easily avoided by using a rounding method of Cont and Heidari (2014) that preserves the integer sum of positive numbers. Moreover, all numerical experiments that we carried out, show that the values of the objective function obtained for non-integer optimum allocation before and after rounding and for the integer optimum allocation are practically indistinguishable. For example, for the two populations used in Section 5, the ratios $V/V_{\text{int}} \in [0.999759, 1]$, while $V_{\text{round}}/V_{\text{int}} = 1$ (up to 6 decimal places), for different total sample sizes $n = 0.1N, \ldots, 0.9N$. Here, V, V_{int} and V_{round} denote variances (1.1) computed for non-integer optimum allocations, integer optimum allocations, and for rounded non-integer optimum allocations (with the rounding method of Cont and Heidari, 2014), respectively.

The above observations suggest that fast, not-necessarily integer-valued allocation algorithms, with properly rounded results, may be a good and reasonable alternative to slower integer algorithms when speed of an algorithm is crucial.

Acknowledgements

We are very grateful to the Associate Editor and the Referees for devoting their time and effort to review the first and the second version of this paper. Their comments helped us a lot in preparation of the present

version of the paper. In particular, we very much appreciate all the remarks and suggestions related to the *fixed-point iteration* algorithm.

Appendix

A. Proof of Theorem 3.1

Remark A.1. *Problem 1.1 is a convex optimization problem as its objective function* $f\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}_+$,

$$f(\mathbf{x}) = \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h}, \tag{A.1}$$

and inequality constraint functions $g_h^m\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$, $g_h^M\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$g_h^m(\mathbf{x}) = m_h - x_h, \qquad h\in\mathcal{H}, \tag{A.2}$$

$$g_h^M(\mathbf{x}) = x_h - M_h, \qquad h\in\mathcal{H}, \tag{A.3}$$

are convex functions, whilst the equality constraint function $w\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$w(\mathbf{x}) = \sum_{h\in\mathcal{H}} x_h - n$$

is affine. More specifically, Problem 1.1 is a convex optimization problem of a particular type in which inequality constraint functions (A.2)-(A.3) are affine. See Appendix D for the definition of the convex optimization problem.

Proof of Theorem 3.1. We first prove that Problem 1.1 has a unique solution. The optimization Problem 1.1 is feasible since requirements $m_h < M_h, h\in\mathcal{H},$ and $\sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h$ ensure that the feasible set $F := \{\mathbf{x}\in\mathbb{R}_+^{|\mathcal{H}|}\colon (1.4)-(1.5)$ are all satisfied$\}$ is non-empty. The objective function (1.3) attains its minimum on F since it is a continuous function and F is closed and bounded. Finally, uniqueness of the solution is due to strict convexity of the objective function on F.

As explained in Remark A.1, Problem 1.1 is a convex optimization problem in which the inequality constraint functions $g_h^m, g_h^M, h\in\mathcal{H}$ are affine. The optimal solution for such a problem can be identified through the Karush-Kuhn-Tucker (KKT) conditions, in which case they are not only necessary but also sufficient; for further references, see Appendix D.

The gradients of the objective function (A.1) and constraint functions (A.2)-(A.3) are as follows:

$$\nabla f(\mathbf{x}) = \left(-\frac{A_h^2}{x_h^2}, h\in\mathcal{H}\right), \quad \nabla w(\mathbf{x}) = \underline{1}, \quad \nabla g_h^m(\mathbf{x}) = -\nabla g_h^M(\mathbf{x}) = -\underline{1}_h, \quad h\in\mathcal{H},$$

where, $\underline{1}$ is a vector with all entries 1 and $\underline{1}_h$ is a vector with all entries 0 except the entry with the label h, which is 1. Hence, the KKT conditions (D.2) for Problem 1.1 assume the form

$$-\frac{A_h^2}{\left(x_h^*\right)^2}+\lambda-\mu_h^m+\mu_h^M=0, \qquad h\in\mathcal{H}, \tag{A.4}$$

$$\sum_{h\in\mathcal{H}} x_h^* - n = 0, \tag{A.5}$$

$$m_h \le x_h^* \le M_h, \qquad h\in\mathcal{H}, \tag{A.6}$$

$$\mu_h^m\left(m_h - x_h^*\right)=0, \qquad h\in\mathcal{H}, \tag{A.7}$$

$$\mu_h^M\left(x_h^* - M_h\right)=0, \qquad h\in\mathcal{H}. \tag{A.8}$$

To prove Theorem 3.1, it suffices to show that for $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$ with $\mathcal{L}^*, \mathcal{U}^*$ satisfying conditions of CASE I or CASE II, there exist $\lambda\in\mathbb{R}$ and $\mu_h^m, \mu_h^M \ge 0, h\in\mathcal{H},$ such that (A.4)-(A.8) hold. It should also be noted that the requirement $m_h < M_h, h\in\mathcal{H},$ guarantees that $\mathcal{L}^*$ and $\mathcal{U}^*$ defined in (3.3) and (3.4) are disjoint. Therefore, $\mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$ is well-defined according to Definition 3.2.

CASE I: Take $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$ with $\mathcal{L}^*$ and $\mathcal{U}^*$ as in (3.3). Then, (A.5) is clearly met after referring to (3.2) and (3.1), while (A.6) follows directly from (3.2) and (3.3), since (3.3) for $h\in\mathcal{H}\setminus\left(\mathcal{L}^*\cup\mathcal{U}^*\right)$ specifically implies $m_h < A_h s\left(\mathcal{L}^*,\mathcal{U}^*\right) < M_h$. Take $\lambda = \frac{1}{s^2(\mathcal{L}^*,\mathcal{U}^*)}$ and

$$\mu_h^m = \begin{cases} \lambda - \frac{A_h^2}{m_h^2}, & h\in\mathcal{L}^* \\ 0, & h\in\mathcal{H}\setminus\mathcal{L}^*, \end{cases} \qquad \mu_h^M = \begin{cases} \frac{A_h^2}{M_h^2} - \lambda, & h\in\mathcal{U}^* \\ 0, & h\in\mathcal{H}\setminus\mathcal{U}^*. \end{cases} \tag{A.9}$$

Note that (3.3) along with requirement $n \ge \sum_{h\in\mathcal{H}} m_h$ (the latter needed if $\mathcal{U}^* = \varnothing$) ensure $s\left(\mathcal{L}^*,\mathcal{U}^*\right) > 0,$ whilst (3.3) alone implies $\mu_h^m, \mu_h^M \ge 0, h\in\mathcal{H}.$ After referring to (3.2), it is a matter of simple algebra to verify (A.4), (A.7) and (A.8) for $\lambda, \mu_h^m, \mu_h^M, h\in\mathcal{H}$ defined above.

CASE II: Take $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$ with $\mathcal{L}^*, \mathcal{U}^*$ satisfying (3.4) and (3.5). Then, condition (A.5) becomes (3.5), while (A.6) is trivially met due to (3.2). Assume that $\mathcal{L}^* \ne \varnothing$ and $\mathcal{U}^* \ne \varnothing$ (for empty $\mathcal{L}^*$ or $\mathcal{U}^*$, (A.4), (A.7) and (A.8) are trivially met). Take an arbitrary $\tilde{s} > 0$ such that

$$\tilde{s} \in \left[\max_{h\in\mathcal{U}^*} \frac{M_h}{A_h}, \min_{h\in\mathcal{L}^*} \frac{m_h}{A_h}\right]. \tag{A.10}$$

Note that (3.4) ensures that the interval above is well-defined. Let $\lambda = \frac{1}{\tilde{s}^2}$ and $\mu_h^m, \mu_h^M, h\in\mathcal{H}$ be as in (A.9). Note that (A.10) ensures that $\mu_h^m, \mu_h^M \ge 0$ for all $h\in\mathcal{H}.$ Then it is easy to check, similarly as in CASE I, that (A.4), (A.7) and (A.8) are satisfied.

B. Auxiliary lemmas and proof of Theorem 4.1

B.1 Notation

Throughout the Appendix B, by $\mathcal{U}_r, \mathcal{L}_r, \tilde{\mathcal{L}}_r,$ we denote sets $\mathcal{U}, \mathcal{L}, \tilde{\mathcal{L}}$ respectively, as they are in the r th iteration of *RNABOX* algorithm after step 3 and before step 4. The iteration index r takes on values from

set $\{1,\ldots,r^*\}$, where $r^* \geq 1$ indicates the final iteration of the algorithm. Under this notation, we have $\mathcal{L}_1 = \varnothing$ and in general, for subsequent iterations, if any $\left(\text{i.e. if } r^* \geq 2\right)$, we get

$$\mathcal{L}_r = \mathcal{L}_{r-1} \cup \tilde{\mathcal{L}}_{r-1} = \bigcup_{i=1}^{r-1} \tilde{\mathcal{L}}_i, \qquad r = 2,\ldots,r^*. \tag{B.1}$$

As *RNABOX* iterates, objects denoted by symbols n and $\mathcal{H}$ are being modified. However, in this Appendix B, whenever we refer to n and $\mathcal{H}$, they always denote the unmodified total sample size and the set of strata labels as in the input of *RNABOX*. In particular, this is also related to set function s (defined in (3.1)) which depends on n and $\mathcal{H}$.

For convenient notation, for any $\mathcal{A} \subseteq \mathcal{H}$ and any set of real numbers $z_h, h \in \mathcal{A}$, we denote

$$z_{\mathcal{A}} = \sum_{h \in \mathcal{A}} z_h.$$

B.2 Auxiliary remarks and lemmas

We start with a lemma describing important monotonicity properties of function s.

Lemma B.1. *Let* $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{H}$ *and* $\mathcal{C} \subseteq \mathcal{D} \subseteq \mathcal{H}$.

1. *If* $\mathcal{B} \cup \mathcal{D} \subsetneq \mathcal{H}$ *and* $\mathcal{B} \cap \mathcal{D} = \varnothing$, *then*

$$s(\mathcal{A},\mathcal{C}) \geq s(\mathcal{B},\mathcal{D}) \quad \Leftrightarrow \quad s(\mathcal{A},\mathcal{C})\left(A_{\mathcal{B}\setminus\mathcal{A}} + A_{\mathcal{D}\setminus\mathcal{C}}\right) \leq m_{\mathcal{B}\setminus\mathcal{A}} + M_{\mathcal{D}\setminus\mathcal{C}}. \tag{B.2}$$

2. *If* $\mathcal{A} \cup \mathcal{D} \subsetneq \mathcal{H}$, $\mathcal{A} \cap \mathcal{D} = \varnothing$, $\mathcal{B} \cup \mathcal{C} \subsetneq \mathcal{H}$, $\mathcal{B} \cap \mathcal{C} = \varnothing$, *then*

$$s(\mathcal{A},\mathcal{D}) \geq s(\mathcal{B},\mathcal{C}) \quad \Leftrightarrow \quad s(\mathcal{A},\mathcal{D})\left(A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}}\right) \leq m_{\mathcal{B}\setminus\mathcal{A}} - M_{\mathcal{D}\setminus\mathcal{C}}. \tag{B.3}$$

Proof. Clearly, for any $\alpha \in \mathbb{R}, \beta \in \mathbb{R}, \delta \in \mathbb{R}, \gamma > 0, \gamma + \delta > 0$, we have

$$\frac{\alpha+\beta}{\gamma+\delta} \geq \frac{\alpha}{\gamma} \quad \Leftrightarrow \quad \frac{\alpha+\beta}{\gamma+\delta}\,\delta \leq \beta. \tag{B.4}$$

To prove (B.2), take

$$\alpha = n - m_{\mathcal{B}} - M_{\mathcal{D}} \qquad \beta = m_{\mathcal{B}\setminus\mathcal{A}} + M_{\mathcal{D}\setminus\mathcal{C}}$$
$$\gamma = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{D}} \qquad \delta = A_{\mathcal{B}\setminus\mathcal{A}} + A_{\mathcal{D}\setminus\mathcal{C}}.$$

Then, $\frac{\alpha}{\gamma} = s(\mathcal{B},\mathcal{D})$, $\frac{\alpha+\beta}{\gamma+\delta} = s(\mathcal{A},\mathcal{C})$, and hence (B.2) holds as an immediate consequence of (B.4).
Similarly for (B.3), take

$$\alpha = n - m_{\mathcal{B}} - M_{\mathcal{C}} \qquad \beta = m_{\mathcal{B}\setminus\mathcal{A}} - M_{\mathcal{D}\setminus\mathcal{C}}$$
$$\gamma = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{C}} \qquad \delta = A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}},$$

and note that $\gamma+\delta = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{C}} + A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}} = A_{\mathcal{H}} - A_{\mathcal{B}} - A_{\mathcal{C}} + A_{\mathcal{B}} - A_{\mathcal{A}} - A_{\mathcal{D}} + A_{\mathcal{C}} = A_{\mathcal{H}} - A_{\mathcal{A}\cup\mathcal{D}} > 0$ due to the assumptions made for $\mathcal{A}, \mathcal{D}, \mathcal{B}, \mathcal{C},$ and $A_h > 0, h \in \mathcal{H}.$ Then, $\frac{\alpha}{\gamma} = s(\mathcal{B}, \mathcal{C})$, $\frac{\alpha+\beta}{\gamma+\delta} = s(\mathcal{A}, \mathcal{D})$, and hence (B.3) holds as an immediate consequence of (B.4).

The remark below describes some relations between sets $\mathcal{L}_r$ and $\mathcal{U}_r, r = 1, \ldots, r^* \geq 1,$ appearing in *RNABOX* algorithm. These relations are particularly important for understanding computations involving the set function s (recall, that it is defined only for such two disjoint sets, the union of which is a proper subset of $\mathcal{H}$).

Remark B.1. *For* $r^* \geq 1,$

$$\mathcal{L}_r \cap \mathcal{U}_r = \varnothing, \qquad r = 1, \ldots, r^*, \tag{B.5}$$

and for $r^* \geq 2,$

$$\mathcal{L}_r \cup \mathcal{U}_r \subsetneq \mathcal{H}, \qquad r = 1, \ldots, r^* - 1. \tag{B.6}$$

Moreover, let $\mathbf{x}^*$ *be as in step 5 of* RNABOX *algorithm. Then, for* $r^* \geq 1,$

$$\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} \subsetneq \mathcal{H} \quad \Leftrightarrow \quad \mathbf{x}^* \textit{ is a } \text{regular } \textit{allocation,} \tag{B.7}$$

and

$$\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H} \quad \Leftrightarrow \quad \mathbf{x}^* \textit{ is a } \text{vertex } \textit{allocation.} \tag{B.8}$$

Proof. From the definition of set $\mathcal{U}$ in step 2 of *RNABOX*, for $r^* \geq 1,$

$$\mathcal{U}_r \subseteq \mathcal{H} \setminus \mathcal{L}_r, \qquad r = 1, \ldots, r^*, \tag{B.9}$$

which proves (B.5). Following (B.1), for $r^* \geq 2,$

$$\mathcal{L}_r = \bigcup_{i=1}^{r-1} \tilde{\mathcal{L}}_i \subseteq \mathcal{H}, \qquad r = 2, \ldots, r^*, \tag{B.10}$$

where the inclusion is due to definition of set $\tilde{\mathcal{L}}$ in step 3 of *RNABOX*, i.e. $\tilde{\mathcal{L}}_r \subseteq \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r)$ for $r = 1, \ldots, r^*.$ Inclusions (B.9), (B.10) with $\mathcal{L}_1 = \varnothing$ imply

$$\mathcal{L}_r \cup \mathcal{U}_r \subseteq \mathcal{H}, \qquad r = 1, \ldots, r^* \geq 1. \tag{B.11}$$

Given that $r^* \geq 2,$ step 4 of the algorithm ensures that set $\tilde{\mathcal{L}}_r \subseteq \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r)$ is non-empty for $r = 1, \ldots,$ $r^* - 1,$ which implies $\mathcal{L}_r \cup \mathcal{U}_r \neq \mathcal{H}.$ This fact combined with (B.11) gives (B.6). Equivalences (B.7) and (B.8) hold trivially after referring to Definition 1.1 of *regular* and *vertex* allocations.

The following two remarks summarize some important facts arising from step 2 of *RNABOX* algorithm. These facts will serve as starting points for most of the proofs presented in this section.

Remark B.2. *In each iteration* $r=1,\ldots,r^*\geq 1,$ *of* RNABOX *algorithm, a vector* $\left(x_h^{**}, h\in\mathcal{H}\setminus\mathcal{L}_r\right)$ *obtained in step 2, has the elements of the form*

$$x_h^{**}=\begin{cases} M_h, & h\in\mathcal{U}_r\subseteq\mathcal{H}\setminus\mathcal{L}_r \\ A_h s\left(\mathcal{L}_r,\mathcal{U}_r\right)<M_h, & h\in\mathcal{H}\setminus\left(\mathcal{L}_r\cup\mathcal{U}_r\right), \end{cases} \tag{B.12}$$

where the set function s *is defined in (3.1). Equation (B.12) is a direct consequence of Theorem C.1.*

Remark B.3. *Remark B.2 together with Theorem C.1, for* $r^*\geq 2$ *yield*

$$\mathcal{U}_r=\left\{h\in\mathcal{H}\setminus\mathcal{L}_r\colon\ A_h s\left(\mathcal{L}_r,\mathcal{U}_r\right)\geq M_h\right\},\quad r=1,\ldots,r^*-1, \tag{B.13}$$

whilst for $r^*\geq 1,$

$$\mathcal{U}_{r^*}=\left\{h\in\mathcal{H}\setminus\mathcal{L}_{r^*}\colon\ A_h s\left(\mathcal{L}_{r^*},\mathcal{U}_{r^*}\right)\geq M_h\right\}, \tag{B.14}$$

if and only if $\mathbf{x}^*$ *(computed at step 5 of* RNABOX *algorithm) is: a* regular *allocation or a* vertex *allocation with* $\mathcal{L}_{r^*}=\mathcal{H}.$

Moreover, for $r^*\geq 1,$

$$\tilde{\mathcal{L}}_r=\left\{h\in\mathcal{H}\setminus\left(\mathcal{L}_r\cup\mathcal{U}_r\right)\colon\ A_h s\left(\mathcal{L}_r,\mathcal{U}_r\right)\leq m_h\right\},\quad r=1,\ldots,r^*. \tag{B.15}$$

Note that in Remark B.3, function s is well-defined due to Remark B.1. The need to limit the scope of (B.14) to *regular* allocations only, is dictated by the fact that in the case of a *vertex* allocation we have $\mathcal{L}_{r^*}\cup\mathcal{U}_{r^*}=\mathcal{H}$ (see (B.8)) and therefore $s\left(\mathcal{L}_{r^*},\mathcal{U}_{r^*}\right)$ is not well-defined.

Lemma B.2 and Lemma B.3 reveal certain monotonicity properties of sequence $\left(\mathcal{U}_r\right)_{r=1}^{r^*}$ and sequence $\left(s\left(\mathcal{L}_r,\mathcal{U}_r\right)\right)_{r=1}^{r^*}$, respectively. These properties will play a crucial role in proving Theorem 4.1.

Lemma B.2. *Sequence* $\left(\mathcal{U}_r\right)_{r=1}^{r^*}$ *is non-increasing, that is, for* $r^*\geq 2,$

$$\mathcal{U}_r\supseteq\mathcal{U}_{r+1},\qquad r=1,\ldots,r^*-1. \tag{B.16}$$

Proof. Let $r^*\geq 2$ and $r=1,\ldots,r^*-1.$ Then, by (B.6), $\mathcal{L}_r\cup\mathcal{U}_r\subsetneq\mathcal{H}.$ Following (B.13), the domain of discourse for $\mathcal{U}_r$ is $\mathcal{H}\setminus\mathcal{L}_r,$ and in fact it is $\mathcal{H}\setminus\left(\mathcal{L}_r\cup\tilde{\mathcal{L}}_r\right)=\mathcal{H}\setminus\mathcal{L}_{r+1},$ since $\mathcal{U}_r\not\subset\tilde{\mathcal{L}}_r$ as ensured by step 3 of *RNABOX*. That is, both $\mathcal{U}_r$ and $\mathcal{U}_{r+1}$ have essentially the same domain of discourse, which is $\mathcal{H}\setminus\mathcal{L}_{r+1}.$ Given this fact and the form of the set-builder predicate in (B.13)-(B.14) as well as equality $\mathcal{U}_{r^*}=\mathcal{H}\setminus\mathcal{L}_{r^*}$ for the case when $\mathbf{x}^*$ is a *vertex* allocation (for which (B.14) does not apply), we conclude that only one of the following two distinct cases is possible: $\mathcal{U}_r\supseteq\mathcal{U}_{r+1}$ or $\mathcal{U}_r\subsetneq\mathcal{U}_{r+1}.$

The proof is by contradiction, that is, assume that (B.16) does not hold. Therefore, in view of the above observation, there exists $r\in\left\{1,\ldots,r^*-1\right\}$ such that $\mathcal{U}_r\subsetneq\mathcal{U}_{r+1}.$ Then,

$$\varnothing \neq \left(\mathcal{U}_{r+1} \setminus \mathcal{U}_r\right) \subsetneq \mathcal{H} \setminus \left(\mathcal{L}_r \cup \mathcal{U}_r\right), \tag{B.17}$$

and hence, due to (B.12),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) < M_h, \qquad h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r. \tag{B.18}$$

On the other hand, from (B.15),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) \leq m_h, \qquad h \in \tilde{\mathcal{L}}_r. \tag{B.19}$$

Summing sidewise: (B.18) over $h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r$, (B.19) over $h \in \tilde{\mathcal{L}}_r$, and then all together, we get

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right)\left(A_{\tilde{\mathcal{L}}_r} + A_{\mathcal{U}_{r+1} \setminus \mathcal{U}_r}\right) < m_{\tilde{\mathcal{L}}_r} + M_{\mathcal{U}_{r+1} \setminus \mathcal{U}_r}. \tag{B.20}$$

Vector $\mathbf{x}^*$ *is a* regular *allocation*: In this case, following Remark B.1, we see that inequality (B.20) is the right-hand side of equivalence (B.2) with

$$\mathcal{A} = \mathcal{L}_r \subseteq \left(\mathcal{L}_r \cup \tilde{\mathcal{L}}_r\right) = \mathcal{L}_{r+1} = \mathcal{B} \subsetneq \mathcal{H}, \tag{B.21}$$

$$\mathcal{C} = \mathcal{U}_r \subsetneq \mathcal{U}_{r+1} = \mathcal{D} \subsetneq \mathcal{H}. \tag{B.22}$$

Then, following Lemma B.1, inequality (B.20) is equivalent to

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right) > s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right). \tag{B.23}$$

Combining

$$s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right) \geq \tfrac{M_h}{A_h}, \qquad h \in \mathcal{U}_{r+1}, \tag{B.24}$$

(it follows from (B.13)-(B.14)) with inequalities (B.23) and (B.18), we get the contradiction

$$\tfrac{M_h}{A_h} > s\left(\mathcal{L}_r, \mathcal{U}_r\right) > s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right) \geq \tfrac{M_h}{A_h}, \quad h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r. \tag{B.25}$$

Therefore, (B.16) holds true, given that $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} \subsetneq \mathcal{H}$.

Vector $\mathbf{x}^*$ *is a* vertex *allocation*: Since $\mathcal{L}_{r+1} \cup \mathcal{U}_{r+1} \subsetneq \mathcal{H}$ for $r = 1, \ldots, r^* - 2$, the proof of (B.16) for such r is identical to the proof for the case of *regular* allocation. Hence, we only need to show that (B.16) holds for $r = r^* - 1$. For this purpose, we will exploit inequality (B.20), which in view of Definition 3.1 of set function s, assumes the following form for $r = r^* - 1$,

$$\tfrac{n - m_{\mathcal{L}_{r^*-1}} - M_{\mathcal{U}_{r^*-1}}}{A_{\mathcal{H}} - A_{\mathcal{L}_{r^*-1} \cup \mathcal{U}_{r^*-1}}}\left(A_{\tilde{\mathcal{L}}_{r^*-1}} + A_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}}\right) < m_{\tilde{\mathcal{L}}_{r^*-1}} + M_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}}. \tag{B.26}$$

Since $A_{\mathcal{H}} - A_{\mathcal{L}_{r^*-1} \cup \mathcal{U}_{r^*-1}} = A_{\tilde{\mathcal{L}}_{r^*-1}} + A_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}}$, for $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H}$, inequality (B.26) simplifies to

$$n < m_{\tilde{\mathcal{L}}_{r^*-1}} + m_{\mathcal{L}_{r^*-1}} + M_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}} + M_{\mathcal{U}_{r^*-1}} = m_{\mathcal{L}_{r^*}} + M_{\mathcal{U}_{r^*}} = n, \tag{B.27}$$

which is a contradiction. Note that the last equality follows from step 2 of the *RNABOX* after referring to (C.3) and using the fact that $\mathcal{U}_{r^*} = \mathcal{H} \setminus \mathcal{L}_{r^*}$ for a *vertex* allocation. Therefore, (B.16) holds true also for $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H}$.

Lemma B.3. *Let* $r^* \geq 3$. *Then*

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right) \geq s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right), \quad r = 1, \ldots, r^* - 2. \tag{B.28}$$

Moreover, if $\mathbf{x}^*$ *(computed at step 5 of* RNABOX *algorithm) is a* regular *allocation and* $r^* \geq 2$, *then*

$$s\left(\mathcal{L}_{r^*-1}, \mathcal{U}_{r^*-1}\right) \geq s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right). \tag{B.29}$$

Proof. We first prove (B.28). Let $r^* \geq 3$ and $r = 1, \ldots, r^* - 2$. Following Lemma B.2 and using (B.13),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) \geq M_h, \qquad h \in \mathcal{U}_r \setminus \mathcal{U}_{r+1}. \tag{B.30}$$

On the other hand, from (B.15),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) \leq m_h, \qquad h \in \tilde{\mathcal{L}}_r. \tag{B.31}$$

Multiplying both sides of inequality (B.30) by -1, summing it sidewise over $h \in \mathcal{U}_r \setminus \mathcal{U}_{r+1}$ and then adding it to (B.31), which is previously summed sidewise over $h \in \tilde{\mathcal{L}}_r$, we get

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right)\left(A_{\tilde{\mathcal{L}}_r} - A_{\mathcal{U}_r \setminus \mathcal{U}_{r+1}}\right) \leq m_{\tilde{\mathcal{L}}_r} - M_{\mathcal{U}_r \setminus \mathcal{U}_{r+1}}. \tag{B.32}$$

Relation (B.32) is the second inequality in (B.3) with

$$\mathcal{A} = \mathcal{L}_r \subsetneq \left(\mathcal{L}_r \cup \tilde{\mathcal{L}}_r\right) = \mathcal{L}_{r+1} = \mathcal{B} \subsetneq \mathcal{H}, \tag{B.33}$$

$$\mathcal{C} = \mathcal{U}_{r+1} \subseteq \mathcal{U}_r = \mathcal{D} \subsetneq \mathcal{H}. \tag{B.34}$$

Based on Remark B.1, we see that $\mathcal{A} \cup \mathcal{D} \subsetneq \mathcal{H}$, $\mathcal{A} \cap \mathcal{D} = \varnothing$, and $\mathcal{B} \cup \mathcal{C} \subsetneq \mathcal{H}$, $\mathcal{B} \cap \mathcal{C} = \varnothing$, and thus the first inequality in (B.3) follows, that is

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right) \geq s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right). \tag{B.35}$$

Hence (B.28) is proved.

If $\mathbf{x}^*$ is a *regular* allocation, in view of Remark B.1, the same reasoning leading to inequality (B.35) clearly remains valid for $r = r^* - 1, r^* \geq 2$.

B.3 Proof of Theorem 4.1

To prove Theorem 4.1, we have to show that:

(I) the algorithm terminates in a finite number of iterations, i.e. $r^* < \infty$,

(II) the solution computed at r^* is optimal.

The proof of part (I) is relatively straightforward. In every iteration $r = 1, \ldots, r^* - 1, r^* \geq 2,$ the set of strata labels $\mathcal{H}$ is reduced by subtracting $\tilde{\mathcal{L}}_r$. Therefore, $r^* \leq |\mathcal{H}| + 1 < \infty,$ where $r^* = |\mathcal{H}| + 1$ if and only if $\left|\tilde{\mathcal{L}}_r\right| = 1, r = 1, \ldots, r^* - 1.$ In words, the algorithm stops after at most $|\mathcal{H}| + 1$ iterations.

In order to prove part (II), following Theorem 3.1 and Remark 3.1, it suffices to show that when $\mathbf{x}^*$ (computed at step 5 of *RNABOX* algorithm) is a *regular* allocation, for all $h \in \mathcal{H},$

$$h \in \mathcal{L}_{r^*} \quad \Leftrightarrow \quad s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \leq \tfrac{m_h}{A_h}, \tag{B.36}$$

$$h \in \mathcal{U}_{r^*} \quad \Leftrightarrow \quad s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \geq \tfrac{M_h}{A_h}, \tag{B.37}$$

and when $\mathbf{x}^*$ is a *vertex* allocation

$$\max_{h \in \mathcal{U}_{r^*}} \tfrac{M_h}{A_h} \leq \min_{h \in \mathcal{L}_{r^*}} \tfrac{m_h}{A_h}, \qquad \text{when } \mathcal{U}_{r^*} \neq \varnothing \text{ and } \mathcal{L}_{r^*} \neq \varnothing, \tag{B.38}$$

$$m_{\mathcal{L}_{r^*}} + M_{\mathcal{U}_{r^*}} = n. \tag{B.39}$$

Vector $\mathbf{x}^*$ *is a* regular *allocation*: Note that Remark B.1, implies that $s\left(\mathcal{L}_r, \mathcal{U}_r\right)$ is well-defined. We start with equivalence (B.36).

Necessity: For $r^* = 1,$ we have $\mathcal{L}_{r^*} = \varnothing$ and hence, the right-hand side of equivalence (B.36) is trivially met. Let $r^* \geq 2,$ and $h \in \mathcal{L}_{r^*} = \bigcup_{r=1}^{r^*-1} \tilde{\mathcal{L}}_r$. Thus, $h \in \tilde{\mathcal{L}}_r$ for some $r \in \left\{1, \ldots, r^* - 1\right\}$ and then, due to (B.15), we have $s\left(\mathcal{L}_r, \mathcal{U}_r\right) \leq \tfrac{m_h}{A_h}$. Consequently, (B.28) with (B.29) yield $s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \leq \tfrac{m_h}{A_h}$.

Sufficiency: Since $\tilde{\mathcal{L}}_{r^*} = \varnothing,$ (B.15) implies

$$\left\{h \in \mathcal{H} \setminus \left(\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*}\right) \colon s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \leq \tfrac{m_h}{A_h}\right\} = \varnothing, \quad r^* \geq 1. \tag{B.40}$$

On the other hand, (B.14) along with $m_h < M_h, h \in \mathcal{H},$ yield

$$s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \geq \tfrac{M_h}{A_h} > \tfrac{m_h}{A_h}, \qquad h \in \mathcal{U}_{r^*}, \tag{B.41}$$

and hence, (B.40) reads

$$\left\{h \in \mathcal{H} \setminus \mathcal{L}_{r^*} \colon s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \leq \tfrac{m_h}{A_h}\right\} = \varnothing, \qquad r^* \geq 1. \tag{B.42}$$

The proof of necessity in (B.37) is immediate in view of (B.14), whilst sufficiency follows by contradiction. Indeed, let $r^* \geq 1.$ Assume that the right-hand side of equivalence (B.37) holds and $h \notin \mathcal{U}_{r^*}.$ Then, in view of Remark B.1, either $h \in \mathcal{H} \setminus \left(\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*}\right)$ and then from (B.12)

$$s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) < \tfrac{M_h}{A_h}, \tag{B.43}$$

a contradiction, or $h \in \mathcal{L}_{r^*}$ and then from (B.36), in view of $m_h < M_h, h \in \mathcal{H}$,

$$s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \leq \frac{m_h}{A_h} < \frac{M_h}{A_h}, \tag{B.44}$$

a contradiction.

Vector $\mathbf{x}^*$ *is a* vertex *allocation*: For $r^* = 1,$ the only possibility is that $\mathcal{U}_{r^*} = \mathcal{H}, \mathcal{L}_{r^*} = \varnothing.$ Then, (B.38) is clearly met, while (B.39) follows from step 2 of the *RNABOX* after referring to (C.3). Let $r^* \geq 2.$ Then, by (B.13) we have

$$s\left(\mathcal{L}_{r^*-1}, \mathcal{U}_{r^*-1}\right) \geq \frac{M_h}{A_h}, \quad h \in \mathcal{U}_{r^*-1} \supseteq \mathcal{U}_{r^*}, \tag{B.45}$$

where the set inclusion is due to Lemma B.2. On the other hand, from (B.15), we get

$$s\left(\mathcal{L}_{r^*-1}, \mathcal{U}_{r^*-1}\right) \leq \frac{m_h}{A_h}, \quad h \in \mathcal{L}_{r^*-1} \cup \tilde{\mathcal{L}}_{r^*-1} = \mathcal{L}_{r^*}, \tag{B.46}$$

where the fact that the above inequality is met for $h \in \mathcal{L}_{r^*-1}$ follows from (B.28). By comparing (B.45) and (B.46) we clearly see that (B.38) is satisfied. Lastly, equation (B.39) is fulfilled due to

$$n - m_{\tilde{\mathcal{L}}_1} - \ldots - m_{\tilde{\mathcal{L}}_{r^*-1}} = n - m_{\mathcal{L}_{r^*}} = M_{\mathcal{U}_{r^*}}, \tag{B.47}$$

where the first equality follows from (B.1) while the second one follows from step 2 of the *RNABOX* after referring to (C.3) and using the fact that $\mathcal{U}_{r^*} = \mathcal{H} \setminus \mathcal{L}_{r^*}$ for a *vertex* allocation.

C. Optimality conditions for Problem 2.1

The following Theorem C.1 provides necessary and sufficient conditions for the optimal solution to Problem 2.1. It was originally given as Theorem 1.1 in Wesołowski et al. (2022) and it is crucial for the proof of Theorem 4.1. Here, we will quote it in a slightly expanded form so that it also covers the case of $\mathcal{U}^* = \mathcal{H}.$ As usual, the set function s is defined as in Definition 3.1. The algorithm that solves Problem 2.1 is *RNA* and it is given in Section 4 of this paper.

Theorem C.1. *The optimization Problem 2.1 has a unique optimal solution. Point* $\mathbf{x}^* = \left(x_h^*, h \in \mathcal{H}\right) \in \mathbb{R}_+^{|\mathcal{H}|}$ *is a solution to optimization Problem 2.1 if and only if* $\mathbf{x}^*$ *has entries of the form*

$$x_h^* = \begin{cases} M_h, & h \in \mathcal{U}^* \\ A_h s\left(\varnothing, \mathcal{U}^*\right), & h \in \mathcal{H} \setminus \mathcal{U}^*, \end{cases} \tag{C.1}$$

with $\mathcal{U}^* \subseteq \mathcal{H},$ *such that one of the following two cases holds:*

CASE I: $\mathcal{U}^* \subsetneq \mathcal{H}$ *and*

$$\mathcal{U}^* = \left\{h \in \mathcal{H}: \; A_h s\left(\varnothing, \mathcal{U}^*\right) \geq M_h\right\}. \tag{C.2}$$

CASE II: $\mathcal{U}^* = \mathcal{H}$ *and*

$$n = \sum_{h \in \mathcal{H}} M_h. \tag{C.3}$$

D. Convex optimization scheme and the KKT conditions

A convex optimization problem is an optimization problem in which the objective function is a convex function and the feasible set is a convex set. In standard form it is written as

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{D}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & w_i(\mathbf{x}) = 0, \quad i = 1, \ldots, k \\ & g_j(\mathbf{x}) \le 0, \quad j = 1, \ldots, \ell, \end{aligned} \tag{D.1}$$

where $\mathbf{x}$ is the optimization variable, $\mathcal{D} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_+$, the objective function $f\colon \mathcal{D}_f \subseteq \mathbb{R}^p \to \mathbb{R}$ and inequality constraint functions $g_j\colon \mathcal{D}_{g_j} \subseteq \mathbb{R}^p \to \mathbb{R}$, $j = 1, \ldots, \ell$, are convex, whilst equality constraint functions $w_i\colon \mathcal{D}_{w_i} \subseteq \mathbb{R}^p \to \mathbb{R}$, $i = 1, \ldots, k$, are affine. Here, $\mathcal{D} = \mathcal{D}_f \cap \bigcap_{i=1}^{k} \mathcal{D}_{w_i} \cap \bigcap_{j=1}^{\ell} \mathcal{D}_{g_j}$ denotes a common domain of all the functions. Point $\mathbf{x} \in \mathcal{D}$ is called *feasible* if it satisfies all of the constraints, otherwise the point is called *infeasible*. An optimization problem is called *feasible* if there exists $\mathbf{x} \in \mathcal{D}$ that is *feasible*, otherwise the problem is called *infeasible*.

In the context of the optimum allocation Problem 1.1 discussed in this paper, we are interested in a particular type of the convex problem, i.e. (D.1) in which all inequality constraint functions g_j, $j = 1, \ldots, \ell$, are affine. It is well known, see, e.g. the monograph Boyd and Vandenberghe (2004), that the solution for such an optimization problem can be identified through the set of equations and inequalities known as the Karush-Kuhn-Tucker (KKT) conditions, which in this case are not only necessary but also sufficient.

Theorem D.1 (KKT conditions for convex optimization problem with affine inequality constraints)**.** *A point* $\mathbf{x}^* \in \mathcal{D} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_+$, *is a solution to the convex optimization problem (D.1) in which functions* g_j, $j = 1, \ldots, \ell$, *are affine if and only if there exist numbers* $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, k$, *and* $\mu_j \ge 0$, $j = 1, \ldots, \ell$, *called KKT multipliers, such that*

$$\begin{aligned} \nabla f\left(\mathbf{x}^*\right) + \sum_{i=1}^{k} \lambda_i \nabla w_i\left(\mathbf{x}^*\right) + \sum_{j=1}^{\ell} \mu_j \nabla g_j\left(\mathbf{x}^*\right) &= \mathbf{0} \\ w_i\left(\mathbf{x}^*\right) = 0, \quad & i = 1, \ldots, k \\ g_j\left(\mathbf{x}^*\right) \le 0, \quad & j = 1, \ldots, \ell \\ \mu_j g_j\left(\mathbf{x}^*\right) = 0, \quad & j = 1, \ldots, \ell. \end{aligned} \tag{D.2}$$

References

Baillargeon, S., and Rivest, L.-P. (2011). The construction of stratified designs in R with the package *stratification*. *Survey Methodology*, 37, 1, 53-65. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11447-eng.pdf.

Barcaroli, G. (2014). `SamplingStrata`: An R Package for the Optimization of Stratified Sampling. *Journal of Statistical Software*, 61(4), 1-24. https://www.jstatsoft.org/index.php/jss/article/view/v061i04.

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Brito, J., Silva, P. and Veiga, T. (2017). `stratbr`: *Optimal Stratification in Stratified Sampling*. R package version 1.2. https://CRAN.R-project.org/package=stratbr.

Burgard, J.P., Münnich, R.T. (2012). Modelling over and undercounts for design-based Monte Carlo studies in small area estimation: An application to the German register-assisted census. *Computational Statistics & Data Analysis*, 56, 2856-2863. https://www.sciencedirect.com/science/article/pii/S0167947310004305.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11682-eng.pdf.

Cont, R., and Heidari, M. (2014). Optimal rounding under integer constraints. https://arxiv.org/abs/1501.00014.

de Moura Brito, J.A., do Nascimento Silva, P.L., Silva Semaan, G. and Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41, 2, 427-442. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-eng.pdf.

Friedrich, U., Münnich, R., de Vries, S. and Wagner, M. (2015). Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. *Computational Statistics & Data Analysis*, 92, 1-12. https://www.sciencedirect.com/science/article/pii/S0167947315001413.

Gabler, S., Ganninger, M., Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75(2), 151-161. DOI: https://doi.org/10.1007/s00184-010-0319-3.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7749-eng.pdf.

Kadane, J.B. (2005). Optimal dynamic sample allocation among strata. *Journal of Official Statistics*, 21(4), 531-541. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/optimal-dynamic-sample-allocation-among-strata.pdf.

Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 2, 205-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10761-eng.pdf.

Lednicki, B., and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6(2), 287-305. https://stat.gov.pl/download/gfx/portalinformacyjny/en/defaultstronaopisowa/3432/1/1/sit_volume_4-7.zip.

Mersmann, O. (2021). `microbenchmark`: *Accurate Timing Functions*. R package version 1.4.9. https://CRAN.R-project.org/package=microbenchmark.

Münnich, R.T., Sachs, E.W. and Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Advances in Statistical Analysis*, 96(3), 435-450. DOI: https://doi.org/10.1007/s10182-011-0176-z.

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

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rivest, L.-P., and Baillargeon, S. (2022). `stratification`: *Univariate Stratification of Survey Populations*. R package version 2.2-7. https://CRAN.R-project.org/package=stratification.

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

Stenger, H., and Gabler, S. (2005). Combining random sampling and census strategies − Justification of inclusion probabilities equal to 1. *Metrika*, 61(2), 137-156. DOI: https://doi.org/10.1007/s001840400328.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation (Chapters 4-6). *Metron*, 2(4), 636-680.

Valliant, R., Dever, J.A. and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, 2nd ed., Springer Cham.

Wesołowski, J., Wieczorkowski, R. and Wójciak, W. (2022). Optimality of the Recursive Neyman Allocation. *Journal of Survey Statistics and Methodology*, 10(5), 1263-1275. https://academic.oup.com/jssam/article-pdf/10/5/1263/46878255/smab018.pdf.

Wieczorkowski, R., Wesołowski, J. and Wójciak, W. (2023). Numerical Performance of the *RNABOX* Algorithm. https://github.com/rwieczor/recursive_Neyman_rnabox.

Wright, T. (2017). Exact optimal sample allocation: More efficient than Neyman. *Statistics & Probability Letters*, 129, 50-57. https://www.sciencedirect.com/science/article/pii/S0167715217301657.

Wright, T. (2020). A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, 165, 108829. https://www.sciencedirect.com/science/article/pii/S0167715220301322.

Wright, S.E., Noble, R. and Bailer, A.J. (2007). Equal-precision allocations and other constraints in stratified random sampling. *Journal of Statistical Computation and Simulation*, 77(12), 1081-1089. DOI: https://doi.org/10.1080/10629360600897191.

Wójciak, W. (2019). *Optimal Allocation in Stratified Sampling Schemes*. Master's thesis, Warsaw University of Technology, Warsaw, Poland. http://home.elka.pw.edu.pl/~wwojciak/msc_optimal_allocation.pdf.

Wójciak, W. (2023a). Another solution for some optimum allocation problem. *Statistics in Transition New Series*, 24(5), 203-219. DOI: https://doi.org/10.59170/stattrans-2023-071.

Wójciak, W. (2023b). `stratallo`: *Optimum Sample Allocation in Stratified Sampling*. R package version 2.2.1. https://CRAN.R-project.org/package=stratallo.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 513-541
Statistics Canada, Catalogue No. 12-001-X

Daily rhythm of data quality: Evidence from the Survey of Unemployed Workers in New Jersey

Jorge González Chapela[1]

Abstract

This paper investigates whether survey data quality fluctuates over the day. After laying out the argument theoretically, panel data from the Survey of Unemployed Workers in New Jersey are analyzed. Several indirect indicators of response error are investigated, including item nonresponse, interview completion time, rounding, and measures of the quality of time diary data. The evidence that we assemble for a time of day of interview effect is weak or nonexistent. Item nonresponse and the probability that interview completion time is among the 5% shortest appear to increase in the evening, but a more thorough assessment requires instrumental variables.

Key Words: Panel data; Survey data quality; Survey of Unemployed Workers in New Jersey; Time of day.

1. Introduction

That surveys are an essential tool for empirical research seems as indisputable as seems that measurement error can compromise the quality of survey data. Among the tenets which appear to underlie the measurement error literature is the principle that the survey respondent must perform a series of cognitive operations before answering a question (e.g., Tourangeau, Rips and Rasinski, 2000, Chapter 1). Each of those operations can be quite complex, involving a great deal of cognitive work (Krosnick, 1999). Extensive research (summarized among others by Schmidt, Collette, Cajochen and Peigneux, 2007) has shown that human performance on a wide range of cognitive tasks fluctuates over the day. Yet, the impact that these fluctuations may have on the quality of survey data remains largely ignored.

This paper attempts to identify problematic times of day for survey data quality by exploiting high-frequency longitudinal microdata from the Survey of Unemployed Workers in New Jersey (SUWNJ). The SUWNJ interviewed online every week for up to 24 weeks some 6,000 workers who were unemployed at the beginning of the survey in October 2009. Although SUWNJ respondents selected themselves to answer the survey at their most convenient times, the availability of repeated observations on each respondent makes it possible to remove the many unobserved factors that remained constant over the relatively short survey period of the SUWNJ (as compared with other large-scale longitudinal surveys).

The paper is organized as follows. Section 2 provides background and context to this research. Section 3 describes the data, the construction of the main variables, and the selection of the sample. Section 4 discusses the methodology. The results are presented in Section 5. Section 6 summarizes the findings and suggests directions for future research.

1. Jorge González Chapela, Centro Universitario de la Defensa de Zaragoza, Academia General Militar, Ctra. de Huesca s/n, 50090 Zaragoza, Spain. E-mail: jorgegc@unizar.es.

2. Background and context

2.1 Background

Psychologists and survey methodologists have characterized a series of cognitive steps in answering survey questions. Tourangeau et al. (2000, page 8) distinguish four steps (comprehension of the question, retrieval of relevant information, use of that information to make required judgments, and selection and reporting of an answer), and provide an illustrative list of mental processes that may be involved in the answering process. Attention and memory are part of that list, both of which have been shown to fluctuate over the day.

The search for time of day fluctuations in human cognitive performance has increasingly been based on the so-called two-process model of sleep-wake regulation (Blatter and Cajochen, 2007; Schmidt et al., 2007). This model postulates that the influence of time of day on cognitive performance is mediated by sleepiness, which in turn is determined by the interacting influences of two propensities. The homeostatic propensity for sleep continuously accumulates during time spent awake and continuously decreases during sleep. The nearly 24-hr (or circadian) oscillatory wake propensity balances the accumulated homeostatic sleep drive during wakefulness.

The circadian wake propensity, which is the result of an internal clock that is synchronized by signals created by the Earth's rotation (light, temperature, etc.) (Roenneberg, Kuehnle, Juda, Kantermann, Allebrandt, Gordijn and Merrow, 2007), reaches its maximum in the evening and its minimum in the early morning. So, for a person who usually sleeps from 23:00 to 07:00, cognitive performance would be at a lower level during nighttime and early morning, a better level occurs around noon, there is a decrease after lunch (e.g., Bes, Jobert and Schulz, 2009), and higher levels occur during afternoon and evening hours (Valdez, 2019). Yet, this time course can be modulated by the kind of task and inter-individual differences in task performance (Blatter and Cajochen, 2007).

The phase of the circadian wake propensity and that of the signals differ across individuals, creating a relationship between internal and external time called phase of entrainment. People who differ in the phase of entrainment are referred to as different chronotypes. The alignment between chronotype and time of day enhances a number of cognitive functions, giving rise to the so-called synchrony effect (e.g., Hasher, Goldstein and May, 2005; Hornik and Tal, 2010; Salehinejad, Wischnewski, Ghanavati, Mosayebi-Samani, Kuo and Nitsche, 2021; Guarana, Stevenson, Gish, Ryu and Crawley, 2022). Thus, if people responded to surveys during hours aligned with their chronotype (as the evidence in Fordsham, Moss, Krumholtz, Roggina, Robinson and Litman, 2019 suggests), the effect of the time of day would be positively moderated by the sorting of respondents into optimal times.

2.2 Context

A careful and comprehensive performance of each of the four steps in the survey answering process can require a substantial amount of mental effort. Hence, according to Krosnick's (1991) satisficing theory,

survey respondents may simply provide a satisfactory answer, the likelihood of which decreases with respondent ability. This insight promoted studies investigating the link between cognitive ability and data quality, the former understood as a stable or slowly changing trait. See, e.g., Kaminska, McCutcheon and Billiet (2010), Kroh, Lüdtke, Düzel and Winter (2016), Gideon, Helppie-McFall and Hsu (2017), Olson, Smyth and Ganshert (2019), Truebner (2021), Angrisani and Couper (2022), Bais, Schouten and Toepoel (2022), and Phillips and Stenger (2022). As predicted by Krosnick (1991), cognitive ability and satisficing appear generally as inversely related.

Time of day fluctuations in cognitive performance may be another aspect of respondent ability related to satisficing. However, this potential link has been little studied. Ziniel (2008, Chapter 4) investigates whether the proportion of "don't knows" provided by respondents to the Health and Retirement Study is sensitive to the time of day, reaching a negative conclusion. Binder (2022) recruited participants from Amazon Mechanical Turk (MTurk) to examine whether inflation expectations and responses to questions with objectively correct answers differ depending on the time of day, finding little differences. On the other hand, a survey carried out on suppliers competing for public contracts in Ireland (Flynn, 2018) reveals that the time of day respondents started the survey predicts survey completeness.

A limitation of these previous studies is that respondents selected themselves to answer the survey at their most convenient times. Hence, and as recognized by Ziniel (2008, Chapter 4), inter-individual differences in cognitive ability or chronotype may interfere with potential time of day fluctuations in cognitive performance. To be sure that factors like these do not interfere with the time of day, Dickinson and McElroy (2010) randomize the survey response window, finding that the time of day (as represented by a binary variable equal to unity for response times from 1:00 to 5:00 a.m. and zero for response times from noon to 7:00 p.m.) has no effect on iterative reasoning.

Identifying problematic times of day for survey data quality is relevant first of all for survey practice, as further measures to reduce the extent of measurement error could be implemented. For example, the e-mailing of invitations/reminders for completing surveys or even the collection of data could be programmed at times of day that were best suited for the increases of data quality. However, forcing respondents to complete surveys at particular times of day could raise nonresponse error (e.g., Weeks, Kulka and Pierson, 1987; Durrant, D'Arrigo and Steele, 2011), so under the total survey error framework (e.g., Lyberg and Stukel, 2017) it would be necessary to study the tradeoff between measurement error and nonresponse error.

Besides the papers that we have already mentioned, our work is related to other strands of literature. Some studies have investigated the characteristics and behaviors of online survey participants as a function of the time of day of participation (e.g., Arechar, Kraft-Todd and Rand, 2017; Casey, Chandler, Levine, Proctor and Strolovitch, 2017; Binder, 2022). Although certain respondent characteristics may be associated with data quality, we focus on data quality and develop effects net of unobserved individual factors and optimal times of participation. The time-of-day fluctuations in cognitive performance have been blamed for the across-the-day variation in a wide spectrum of economic decisions and abilities; see, e.g., Carrell, Maghakian and West (2011), Dickinson and McElroy (2017), Williams and Shapiro (2018), Collinson,

Mathmann and Chylinski (2020), Dickinson, Chaudhuri and Greenaway-McGrevy (2020), and Guarana et al. (2022). But whether survey data quality is modulated by the time of day remains largely ignored. Last, but not least, by exploiting start and end times of each interview, we relate to the literature using paradata to investigate measurement error (reviewed by Yan and Olson, 2013).

3. Data, measures, and sample selection

3.1 SUWNJ

The data for this study are taken from the SUWNJ, a longitudinal Internet-based survey of unemployed workers conducted by the Princeton University Survey Research Center between October 2009 and April 2010. Here, we describe the main features of this survey, referring to Krueger and Mueller (2010, 2011) for the survey questionnaire, the data set, and a more complete description of the SUWNJ. The Stata code needed to proceed from the raw data to the results is available from the author upon request.

3.1.1 Sampling and invitation

The individuals sampled were selected from the universe of unemployment insurance (UI) benefit recipients in the state of New Jersey as of September 28, 2009. During 2009 and 2010, New Jersey's unemployment rate closely mirrored the U.S. average, although its population of UI recipients was more female, older, and more educated than in the wider U.S. The sample was selected through stratified random sampling with strata defined by initial duration of unemployment and availability of an e-mail address. Those unemployed 60 weeks or longer and those with an e-mail address were oversampled.

The selected individuals were invited to participate in the survey for 12 consecutive weeks, although the long-term unemployed were invited to participate in an extended study for an additional 12 weeks. The initial invitation was sent by e-mail or (to those without e-mail address) physical letter. The e-mail (letter) contained a link to the online questionnaire. Individuals contacted by letter were required to enter a valid e-mail address in order for them to receive e-mail invitations for the follow-up weekly interviews. If a respondent did not have an e-mail address, he/she could nevertheless participate in the weekly interviews by logging into the same access web page. According to the October 2009 Current Population Survey, 15% of New Jersey's unemployed workers lived in households where no one used the Internet, but no further arrangements were made to secure the participation of Internet non-users. The invitation e-mails (sent in the morning) asked individuals to complete the survey within two days and even if they had already found a job.

3.1.2 Participation and weighting

The AAPOR (2023) RR6 response rate for the first interview was 9.7% (6,025 persons). These respondents completed an average of 4.1 follow-up interviews out of a maximum of 11 (excluding the longer-term follow-up), responding to 24,638 (37.2%) of the potential follow-up interviews. Only 302

individuals completed 12 interviews, so 95.0% attrited from the initial study. The RR6 response rate for the first interview of the extended study was larger, 56.8% (1,148 persons). These respondents completed an average of 6.4 follow-up interviews out of a maximum of 11, responding to 7,390 (58.5%) of the potential follow-up interviews. 115 individuals completed 12 interviews, so 90.0% attrited from the extended study. All this yields 39,201 interviews.

The low response rates created noticeable differences between the universe of New Jersey UI recipients and the respondents. Krueger and Mueller (2011) created inverse propensity weights based on administrative data from the UI system. The weights labeled "current week weights" adjust for differential sampling probabilities and response rates over the 12 weeks of the survey (or 24 weeks, for those who participated in the extended study). The regressors utilized to create "current week weights" were strata indicator variables and time-invariant demographics.

3.1.3 Survey instrument

The SUWNJ questionnaire consists of two parts: an entry survey, administered in the first week, with demographic, income, and wealth questions, and a weekly survey, administered in the first and each subsequent week, with questions related to life satisfaction, food expenditure, job search activities, and time use. The time use information is for the day previous to the interview, and is collected by means of a self-completed time diary from 07:00 to 23:00 and two questions asking wake-up and going-to-bed times. To complete the diary, respondents could select up to 2 activities for each hour from a pre-designated list of 21 activities.

After initiating an interview, respondents could move back and forth through the questionnaire, as well as stop the interview and return to it later. Although completion via phone browser was possible, the questionnaire was not optimized to be taken on a mobile device. The data set includes the date and time (recorded to the second) that each interview was initiated and completed, plus the end time of the time-use section (the third of the five sections of the weekly survey).

3.2 Measures

3.2.1 Time of day

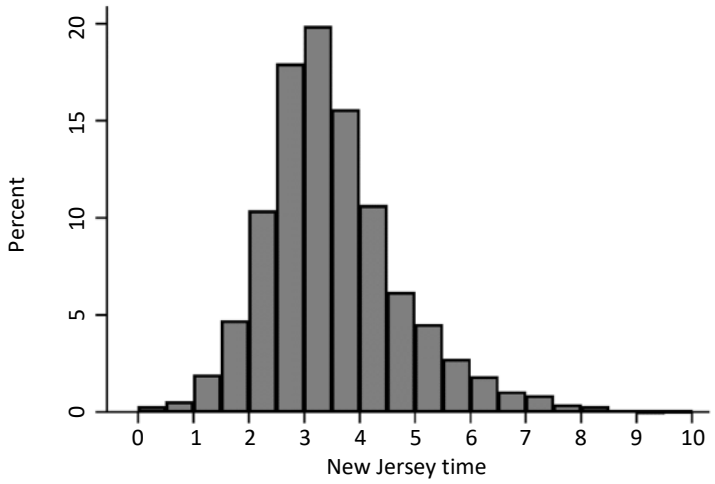
Times are local times of New Jersey, measured continuously from midnight and expressed in hours (e.g., 9.5 for 09:30). The time of day of interview (denoted D) is approximated by the mid-time between the start and end times of the interview. In a robustness check, it will be approximated by randomly selected points within the start and end times of the interview (Ahn, Peng, Park and Jeon, 2012).

3.2.2 Chronotype

Roenneberg et al. (2007) use the Munich ChronoType Questionnaire to assess chronotype, measured as the half-way point between sleep-onset and sleep-end (or mid-sleep) on free days corrected for oversleep (MSF_{sc}). A proxy measure for chronotype can be constructed along those lines using the SUWNJ time-use

information. Sleep duration is estimated as the time between going to bed and wake up, and its half-way point is averaged over free days. For individuals who sleep longer on free days than on workdays, the difference between sleep duration on free days and its weekly average (assuming a 5-day workweek) is subtracted from the mid-sleep on free days. The resulting measure is denoted MSF_{sc}^{e}. Sleep timing and sleep duration are essentially independent traits (Roenneberg et al., 2007). The correlation between MSF_{sc}^{e} and average sleep duration is 0.06 (although statistically different from zero at 5% level). Figure 3.1 shows the distribution of MSF_{sc}^{e} in the sample.

Figure 3.1 Chronotype (MSF_{sc}^{e}).



Source : SUWNJ.

3.2.3 Data quality

We analyze four sets of measures of data quality (Juster, 1986; Malhotra, 2008; Fricker and Tourangeau, 2010): i) the percent item nonresponse, ii) measures of the quality of time-diary data (the number and variety of activities and the number of hours not coded in the diary), iii) time to complete the interview, and iv) rounded values of mood at home, food expenditure at home, and expenditure on eating out. The SUWNJ questionnaire seems to contain insufficient items to investigate response errors caused by social desirability or extreme, midpoint, or nondifferentiated responding (see, e.g., Baumgartner and Steenkamp, 2001; Chang and Krosnick, 2009). The data file contains completed interviews, which precludes analyzing survey breakoff (e.g., Peytchev, 2009).

We define the percent item nonresponse (P_{INR}) as the percentage of missing values for questions administered to all respondents at a certain interview. This excludes follow-up questions plus questions that can be postponed to the next time the person is interviewed.

To count the number and variety of activities recorded in the time diary, we follow the convention that if an activity intervenes in the middle of some other activity (e.g., shopping on the way home from a job interview), the number of activities increases by two units and the variety of activities by one unit (Juster, 1986). We present results for the number of activities (denoted NumAct), as those for the variety follow the

same patterns. If no activity is recorded in an hour, the hour is considered not coded. The variable counting the number of hours not coded is denoted HMissing. Juster (1986) notices that only weekday (Monday–Thursday) diaries suffer significant quality deterioration to the extent that they involve more than 24-hour recall, a finding that will be helpful for interpreting some of our results.

The relationship between interview completion time and data quality in Internet-based surveys is complex, as both short and long completion times may be a symptom of respondent inattention (Malhotra, 2008; Read, Wolters and Berinsky, 2021). Hence, besides a continuous measure of completion time (denoted IvDur), we analyze dummy variables for the 5% lowest and 5% highest completion times, denoted $P_{IVDUR5L}$ and $P_{IVDUR5H}$ respectively. These dummies are created by calculating the corresponding percentiles separately for first and subsequent interviews after removing outliers (see Section 3.3).

Information on mood at home is collected with the question: "Now we would like to know how you feel and what mood you are in when you are at home. When you are at home, what percentage of the time are you: in a bad mood, a little low or irritable, in a mildly pleasant mood, in a very good mood?" Respondents are asked to indicate the percentage of time that they experienced each mood category. We created dummy variables indicating respondents for whom *all* four reported percentages are multiples of 50 (leading to answers of 0, 50, or 100), 25, or 10. The three binary variables are denoted P_{MOOD50}, P_{MOOD25}, and P_{MOOD10}, respectively.

Two questions gather information on expenditure on food: "In the last 7 days, how much did you and anyone else in your family spend on food that you use at home? Please include food bought with food stamps", and "In the last 7 days, how much did you and anyone else in your family spend on eating out?" We created dummy variables indicating respondents for whom a certain expenditure is multiple of 100 or 50, denoted $P_{FOODAH100}$, $P_{FOODAH50}$, $P_{EATING\text{-}OUT100}$, and $P_{EATING\text{-}OUT50}$. Zero expenditure could be reflecting rounding, a corner solution, or infrequency of purchase. We present results assuming that zero expenditure reflects rounding, and analyze their robustness to assuming that zero expenditure does not reflect rounding.

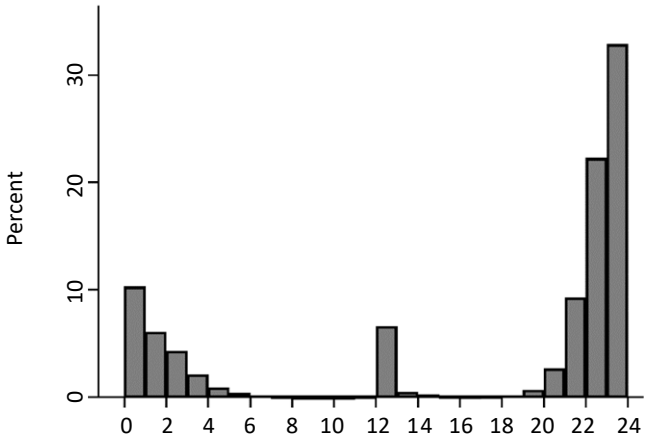
3.3 Sample selection

The distribution of interview completion time is heavily right-skewed, with median (mean) completion time of 13.2 (144.3) minutes for first interviews and 11.8 (75.6) minutes for subsequent interviews. To avoid introducing much error into our measure of D, interviews with completion time greater than 60 minutes are discarded, representing 5.7 and 4.4% of first and subsequent interviews. Moreover, we discarded first or subsequent interviews with completion times below the corresponding 1st percentile (4.8 and 3.6 minutes, respectively). Based on our own reading time, this lower bound discards interviews in which the respondent cannot have read the questionnaire. (Including these interviews leaves the conclusions unchanged.)

We also discarded interviews presenting missing or inconsistent data in some variable used in this study. Here, an issue requires some discussion. Going-to-bed time was reported using three drop-down menus of hour, minute, and AM/PM period. The AM/PM menu was set by default to PM, and Figure 3.2 suggests that this may have induced error. While going to bed between 11:00 and 11:59 a.m. is reported in 0.04% of

interviews, 6.6% report going to bed between 12:00 and 12:59 p.m. We probed the time diary for an inconsistent going-to-bed time when this was between 12:00 and 02:59 p.m. When an inconsistency was found, the interview was discarded. A total of 2,578 interviews were discarded for this reason. When assessing robustness, we shall include them in the sample assuming that the AM period applies, and a dummy indicating those cases (denoted $P_{PM\text{-}AM}$) will be analyzed for time of day of interview effects.

Figure 3.2 Going-to-bed time.



Source : SUWNJ.

Finally, the last interview of a person who completed 25 interviews is discarded because it is not clear whether he ended up attriting. All this leaves us with 5,531 persons and a total of 33,000 interviews. Figure 3.3 provides a histogram of the number of interviews contributed by each person. The mean (median) number of interviews is 6.0 (4). Table 3.1 provides descriptive statistics for the main variables used in this study.

Figure 3.3 Number of interviews contributed to the sample.

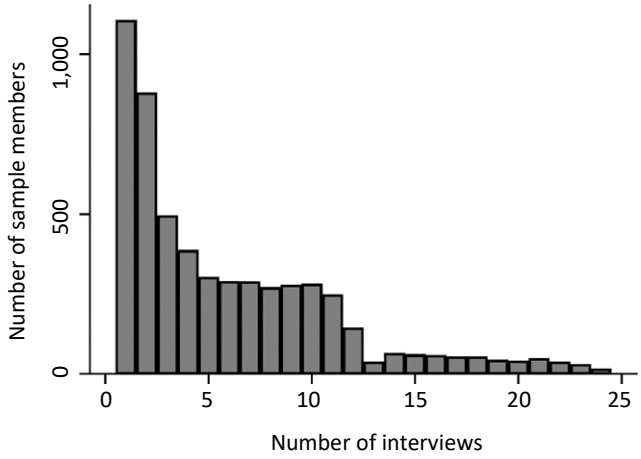


Table 3.1
Descriptive statistics.

	Observations	Mean	Standard deviation	Min	Max
P_{INR}[a]	33,000	2.64	6.58	0	60.87
NumAct	33,000	16.77	7.14	1	32
HMissing	33,000	0.53	2.14	0	15
IvDur (minutes)	33,000	13.95	8.19	3.57	59.95
$P_{IVDUR5L}$[a]	33,000	4.99			
$P_{IVDUR5H}$[a]	33,000	4.99			
P_{MOOD10}[a]	32,877	50.40			
P_{MOOD25}[a]	32,877	15.84			
P_{MOOD50}[a]	32,877	10.79			
$P_{FOODAH50}$[a]	31,949	51.64			
$P_{FOODAH100}$[a]	31,949	30.63			
$P_{EATING\text{-}OUT50}$[a]	29,084	45.74			
$P_{EATING\text{-}OUT100}$[a]	29,084	34.67			
Time of day of interview	33,000	12.94	4.80 (3.89) [3.45]	0	23.99
MSF_{sc}^{c}	5,531	3.56	1.65	0	23.99
Day of interview	33,000				
Monday[b]		8.56			
Tuesday[b]		23.39			
Wednesday[b]		16.18			
Thursday[b]		14.51			
Friday[b]		17.93			
Saturday[b]		12.76			
Sunday[b]		6.67			
Worked[b]	33,000	14.00			
Sleep duration (hours)	33,000	8.35	2.10	0.50	23.58
No. of previous interviews	33,000	5.34	5.00	0	23
Weeks between $t-2$ and $t-1$	33,000	1.43	1.40	0	16

Notes: The data pertain to 5,531 individuals. The sample variation of time of day of interview is made up of "within" (or time series) variation (shown in parentheses) and "between" (cross-section) variation (shown in brackets). Worked and Sleep duration are for the diary day. [a]: Binary indicator for the outcome given in the name's subscript scaled as a percentage. [b]: Binary indicator scaled as a percentage.

4. Methods

4.1 Baseline specification

As respondents select themselves to answer surveys at their most convenient times, it is unlikely that a simple comparison of data quality outcomes by time of day of interview can identify a causal effect. The availability of repeated observations on each SUWNJ respondent allows us to control for unobserved time-constant factors such as cognitive ability or chronotype. Measurement error (as defined in Biemer, Groves, Lyberg, Mathiowetz and Sudman, 2004, page xvii) also arises from the method of data collection and the questionnaire, but since these features are fixed across interviews, they cannot interfere with our estimates.

The following unobserved effects panel data model (Wooldridge, 2010, Chapter 10) is estimated:

$$y_{it} = \alpha\left(D_{it}\right) + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \tag{4.1}$$

where y_{it} denotes some data quality measure for individual i $\left(i = 1, 2, \ldots, N\right)$ at interview number t $\left(t = 1, 2, \ldots, T_i\right)$, $\alpha\left(D_{it}\right)$ is a scalar function of time of day of interview, $\mathbf{x}_{it}$ is a vector of interview-variant observable controls, $\boldsymbol{\beta}$ is a vector of unknown parameters, c_i is an unobserved individual effect arbitrarily correlated with D_{it} and $\mathbf{x}_{it}$, and u_{it} is an idiosyncratic error term.

Besides an intercept, and following Binder (2022) and Juster (1986), included in $\mathbf{x}_{it}$ are dummy variables for day of week of interview and a dummy for whether the respondent worked on the diary day (this information is not available for the day of interview). Cumulative insufficient sleep (e.g., Lowe, Safati, and Hall, 2017) and synchrony effects could also affect y_{it}. Hence, sleep duration on the diary day and the interaction between $\text{MSF}_{\text{sc}}^{\text{e}}$ and single-hour dummies for D_{it} are included in $\mathbf{x}_{it}$. The single-hour dummies are constructed by rounding D_{it} to the nearest integer hour, producing 24 dummies to be interacted with $\text{MSF}_{\text{sc}}^{\text{e}}$. Yet, one dummy is excluded because of collinearity with c_i. The median $\text{MSF}_{\text{sc}}^{\text{e}}$ is subtracted from $\text{MSF}_{\text{sc}}^{\text{e}}$ so $\alpha\left(D_{it}\right)$ represents the median chronotype.

Panel conditioning effects can operate in a longitudinal survey, which may entail positive or negative consequences for data quality (e.g., Bach, 2021). Respondents may gain a better understanding of the meaning of the questions with repeated administration of the questionnaire, increasing the reliability of their responses (Kroh, Winter and Schupp, 2016). On the other hand, respondents may learn to falsely respond some questions to skip follow-up questions, lowering the quality of the data (e.g., Davis, 2011). To account for panel conditioning effects, a complete set of dummy variables for the number of previous interviews is included in $\mathbf{x}_{it}$. This number can be 0, 1, 2, …, 23, producing 24 dummies. Yet, one dummy is excluded because of collinearity with the intercept.

Let $\mathbf{z}_{it}\boldsymbol{\theta} \equiv \alpha(D_{it}) + \mathbf{x}_{it}\boldsymbol{\beta}$ with $K = \dim(\boldsymbol{\theta})$. Under the strict exogeneity assumption $E(u_{it} \mid \mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iT_i}, c_i) = 0$, $\boldsymbol{\theta}$ can be estimated by ordinary least squares (OLS) of

$$\Delta y_{it} = \Delta \mathbf{z}_{it}\boldsymbol{\theta} + e_{it}, \qquad t = 2, 3, \ldots, T_i, \tag{4.2}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta \mathbf{z}_{it} = \mathbf{z}_{it} - \mathbf{z}_{i,t-1}$, and $e_{it} = u_{it} - u_{i,t-1}$. The e_{it} are assumed to be independently distributed across individuals but no restrictions are placed on the form of the autocovariances for a given individual. Heteroskedasticity and serial correlation consistent standard errors are obtained from the following variance matrix estimator (Wooldridge, 2010, pages 172 and 318):

$$\hat{\mathrm{V}}\left(\hat{\boldsymbol{\theta}}\right) = \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \Delta \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i' \Delta \mathbf{Z}_i\right) \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \Delta \mathbf{Z}_i\right)^{-1} \tag{4.3}$$

where $\Delta \mathbf{Z}_i$ is the $\left(T_i - 1\right) \times K$ matrix obtained by stacking $\Delta \mathbf{z}_{it}$ from $t = 2, 3, \ldots, T_i$ and $\hat{\mathbf{e}}_i$ is the $\left(T_i - 1\right) \times 1$ vector of OLS residuals $\hat{e}_{it}$, $t = 2, 3, \ldots, T_i$. Alternatively, a working correlation matrix for modeling within-individual correlations can be specified, and the resulting model can be estimated by population-averaged methods, called feasible generalized least squares (FGLS) estimators in econometrics. We provide the results of two FGLS estimators in a separate supplement (González Chapela, 2024). They reveal essentially the same patterns reported here.

To assess the strict exogeneity of $\left\{D_{it}\colon t = 1, \ldots, T_i\right\}$, $\alpha\left(D_{it}\right)$ will be added to equation (4.2) and then its statistical significance tested using the Wald test (Wooldridge, 2010, page 325).

4.2 Model types and model selection

Our objective is to arrive at a reasonable, parsimonious representation of $\alpha\left(D_{it}\right)$. Hence, an information criterion is employed to select a model for $\alpha\left(D_{it}\right)$ out of three linear-in-parameters model types: piecewise constant functions (specifically, those of Arechar et al., 2017; Binder, 2022; Durrant et al., 2011; Flynn, 2018; Valdez, 2019 and Weeks et al., 1987), a polynomial of degree three, and the cosinor model

$$\alpha\left(D_{it}\right)=\alpha_1 \sin\left(D_{it}\times 2\pi/24\right)+\alpha_2 \cos\left(D_{it}\times 2\pi/24\right), \tag{4.4}$$

where α_1 and α_2 are unknown parameters.

The cosinor model is a type of Fourier series representation in which sines and cosines are used to approximate complex mathematical waveforms (Brown and Czeisler, 1992; Cornelissen, 2014). Given the waveform character of the homeostatic and circadian propensities for sleep, cosinor may provide an appropriate representation of $\alpha\left(D_{it}\right)$. The cosinor model has 1 peak and 1 trough separated by 12 hours and equal in amplitude and width, the locations of which are determined by α_1 and α_2. Twice the amplitude of the cosinor wave, or $2\times\sqrt{\alpha_1^2+\alpha_2^2}$, provides a measure of the extent of predictable change within the day.

The degree three polynomial is less restrictive than cosinor because the peak and the trough may not be separated by 12 hours and the amplitude and width of the peak may differ from those of the trough. On the other hand, a polynomial may not be periodic, i.e. its values may not repeat themselves every 24 hours. To ensure periodicity, the restriction $\alpha\left(0\right)=\alpha\left(24\right)$ is imposed, yielding

$$\alpha\left(D_{it}\right)=\alpha_1 D_{it}\left(1-\left(D_{it}/24\right)^2\right)+\alpha_2 D_{it}^2\left(1-D_{it}/24\right). \tag{4.5}$$

To select among models, Schwarz's (1978) Bayesian information criterion

$$\mathrm{BIC}=\ln \mathrm{SSR}+\frac{K\ln\left(\sum_{i=1}^{N}\left(T_i-1\right)\right)}{\left(\sum_{i=1}^{N}\left(T_i-1\right)\right)} \tag{4.6}$$

is used, where SSR denotes a model's sum of squared residuals. BIC is preferred to other popular criteria when some modelling alternatives are nested (Nishii, 1988). The specification of $\mathbf{x}_{it}$ is kept the same throughout the selection process. Schwarz (1978) establishes the validity of BIC for independent and identically distributed observations. To guard against possible biases created by correlated e_{it}, the BIC values were recalculated using N in place of $\sum_{i=1}^{N}\left(T_i-1\right)$ (StataCorp, 2019, page 104), producing the same selection of models.

4.3 Attrition

If attrition is driven by unobserved factors that do not change over the survey period, then removing c_i would correct for attrition bias. Nevertheless, one might still be concerned about attrition as a consequence of unobserved interview-variant factors. We use a variant of the procedure proposed by Wooldridge (2010,

page 837) to test and correct for attrition bias, though we note that this procedure does not correct for individuals selected to participate in the SUWNJ who never responded. As the data for each individual are organized by interview number, attrition is an absorbing state.

Let s_{it} denote the interview completion indicator, with $s_{it}=1$ if individual i completed the t interview and $s_{it}=0$ if i abandoned the survey right after the $t-1$ interview. The completion equation for interview t conditional on $s_{i,t-1}=1$ is

$$s_{it}=1\left[\mathbf{w}_{it}\boldsymbol{\delta}+v_{it}>0\right], \qquad t=2,3,\ldots,T_i, \tag{4.7}$$

where $1[\cdot]$ is the indicator function, $\mathbf{w}_{it}$ is a set of variables that are observed whether or not the individual attrited, $\boldsymbol{\delta}$ is a vector of unknown parameters, and v_{it} is a standard normal error term assumed independent of $\left(\Delta\mathbf{z}_{it},\mathbf{w}_{it},s_{i,t-1}=1\right)$. Nonrandom attrition occurs when v_{it} and e_{it} are correlated.

Assuming that e_{it} is independent of $\left(\Delta\mathbf{z}_{it},\mathbf{w}_{it}\right)$ and that $\mathrm{E}\left(e_{it}\,\middle|\,v_{it},s_{i,t-1}=1\right)=\rho_t v_{it}$, ρ_t being an unknown parameter, the unknown parameters of equation (4.1) can be estimated by OLS of

$$\Delta y_{it}=\Delta\mathbf{z}_{it}\boldsymbol{\theta}+\rho_2 d2_t\hat{\lambda}_{it}+\ldots+\rho_{24}d24_t\hat{\lambda}_{it}+\varepsilon_{it}, \qquad t=2,3,\ldots,T_i. \tag{4.8}$$

In this expression, $d2_t,\ldots,d24_t$ are interview dummies so that $dj_t=1$ if $t=j$ and $dj_t=0$ if $t\neq j$, $\hat{\lambda}_{it}\equiv\lambda\left(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}\right)=\phi\left(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}\right)\Big/\Phi\left(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}\right)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of the standard normal distribution, is the estimated inverse Mills ratio, and ε_{it} is an error term.

An estimator of $\boldsymbol{\delta}$ is available from pooled probit estimation of the interview completion equation:

$$\mathrm{P}\left(s_{it}=1\,\middle|\,\mathbf{w}_{it},s_{i,t-1}=1\right)=\Phi\left(\mathbf{w}_{it}\boldsymbol{\delta}\right), \qquad t=2,3,\ldots,T_i. \tag{4.9}$$

We use pooled probit because $\boldsymbol{\delta}$ is assumed to be constant across interviews. If $\boldsymbol{\delta}$ was allowed to change (as in Wooldridge's original formulation), a probit would be estimated for each t. However, this approach is problematic because in many occasions the variables included in $\mathbf{w}_{it}$ perfectly predict one of the outcomes. The vector $\mathbf{w}_{it}$ comprises single-hour dummies for $D_{i,t-1}$, $\mathbf{x}_{i,t-1}$, and the number of weeks passed between $t-2$ and $t-1$. (For $t=2$, we count the number of weeks between the week when the initial invitations to participate in the survey were sent and the week of the first interview.)

Attrition bias can be tested by a joint test of H_0: $\rho_t=0$, $t\geq 2$, in equation (4.8). If H_0 is rejected, standard errors are corrected for the presence of estimated parameters in $\hat{\lambda}_{it}$ drawing upon Arellano and Meghir (1992).

4.4 Weighting

Since the regressors utilized to create "current week weights" are absorbed in c_i, model (4.1) includes all the design variables and thus the sampling design can be considered ignorable (Pfeffermann, 1993). Hence, the main analysis is conducted without sampling weights. However, reporting weighted estimates is useful as a misspecification check, as the failure to model heterogenous effects can generate significant

contrasts between weighted and unweighted estimates (e.g., Solon, Haider and Wooldridge, 2015). Hence, equation (4.2) will be re-estimated by weighted least squares (WLS).

4.5 Multiple inference

Nearly all of our groupings of data quality measures contain more than one measure. Consequently, significant effects may emerge by chance for some measure even if no effect on the grouping exists. To control for this, Bonferroni corrections are performed and significance is declared at level $0.05/M$, M being the number of measures in the grouping.

5. Results

5.1 Model selection

Table 5.1 lists the best-fitting models of $\alpha(D_{it})$. The cosinor model is the preferred option for analyzing most of the data quality measures. However, Binder's (2022) piecewise constant function (indicators for 06:00 to 11:59, 12:00 to 18:59, and 19:00 to 05:59) is the best fitting alternative for the number of hours not coded in the diary (HMissing), the probability of reporting all mood at home categories in multiples of 50 (P_{MOOD50}), and the probability of reporting expenditure on eating out in multiples of 50 ($P_{EATING\text{-}OUT50}$). The degree three polynomial is favored for the probability of being among the 5% highest completion times ($P_{IVDUR5H}$) and the probability of reporting expenditure on eating out in multiples of 100 ($P_{EATING\text{-}OUT100}$). For the probability of reporting all mood at home categories in multiples of 25 (P_{MOOD25}), Durrant et al.'s (2011) piecewise constant function (indicators for 00:00 to 11:59, 12:00 to 16:59, and 17:00 to 23:59) is preferred.

Table 5.1
Model selected for $\alpha(D_{it})$.

Dependent variable	Model	BIC value
P_{INR}	Cosinor	13.259
NumAct	Cosinor	13.585
HMissing	Piecewise constant (Binder, 2022)	11.081
IvDur	Cosinor	14.327
$P_{IVDUR5L}$	Cosinor	16.476
$P_{IVDUR5H}$	Degree 3 polynomial	16.716
P_{MOOD10}	Cosinor	18.031
P_{MOOD25}	Piecewise constant (Durrant et al., 2011)	17.063
P_{MOOD50}	Piecewise constant (Binder, 2022)	16.670
$P_{FOODAH50}$	Cosinor	18.096
$P_{FOODAH100}$	Cosinor	18.018
$P_{EATING\text{-}OUT50}$	Piecewise constant (Binder, 2022)	17.956
$P_{EATING\text{-}OUT100}$	Degree 3 polynomial	17.759

5.2 Baseline results

The results of estimating equation (4.2) with the functional forms listed in Table 5.1 are presented in Tables 5.2 and 5.3. Table 5.2 shows the results for the percent item nonresponse (P_{INR}), the time-diary measures, and interview completion time. Table 5.3 gathers the results for the indicators of rounding. The bottom rows of both tables list the p-values for the tests of significance of $\alpha(D_{it})$ and strict exogeneity of $\{D_{it}: t = 1, \ldots, T_i\}$.

A statistically significant $\alpha(D_{it})$ is detected in some of the regressions, which suggests the existence of some effects on data quality of D_{it}. In a p-value sense, the strongest evidence is found in the regressions for the number of activities (NumAct) and the probability of being among the 5% lowest completion times ($P_{IVDUR5L}$). The null of no effect is also rejected at 5% in the regressions for P_{INR} and $P_{EATING\text{-}OUT100}$. No statistically significant effect is detected in the remaining cases.

In the case of $P_{EATING\text{-}OUT100}$, the rejection of the null does not hold if zero expenditure (reported in 28% of the interviews) is assumed not to reflect rounding (p-value 0.55). In addition, the effect on $P_{EATING\text{-}OUT100}$ does not survive a Bonferroni correction for two simultaneous tests in the group of measures assessing expenditure on eating out, which would require p-value < 0.025.

Table 5.2
Time of day of interview effects on data quality.

	(1) P_{INR}		(2) NumAct		(3) HMissing		(4) IvDur (min)		(5) $P_{IVDUR5L}$		(6) $P_{IVDUR5H}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59					0.049	0.029						
19:00–05:59					0.048	0.039						
$D_{it}(1-(D_{it}/24)^2)$											-0.169	0.238
$D_{it}^2(1-D_{it}/24)$											0.039	0.024
$\sin(D_{it}\times 2\pi/24)$	-0.147*	0.070	0.259*	0.087			-0.020	0.135	-0.405	0.361		
$\cos(D_{it}\times 2\pi/24)$	0.040	0.064	-0.191*	0.086			-0.282*	0.138	0.814*	0.373		
Tuesday	-0.190	0.108	1.480*	0.137	-0.103*	0.039	0.733*	0.192	-0.251	0.491	1.013	0.624
Wednesday	-0.157	0.118	1.207*	0.150	-0.079	0.041	0.430*	0.204	-0.960	0.547	-0.046	0.656
Thursday	0.015	0.131	1.233*	0.159	-0.048	0.043	0.653*	0.226	-0.380	0.573	1.146	0.762
Friday	-0.008	0.118	1.132*	0.152	-0.075	0.042	0.272	0.206	0.188	0.505	-0.138	0.656
Saturday	0.012	0.124	0.755*	0.153	-0.027	0.041	0.387	0.246	0.514	0.605	0.140	0.781
Sunday	0.289*	0.140	0.054	0.171	0.001	0.045	-0.380	0.242	0.151	0.660	-0.882	0.759
Worked	0.408*	0.137	-1.711*	0.158	-0.120*	0.037	-0.174	0.173	1.378*	0.609	0.623	0.555
Sleep duration	-0.024	0.018	-0.080*	0.025	-0.016*	0.007	-0.131*	0.028	0.400*	0.089	-0.178	0.095
Significance of $\alpha(D_{it})$	[0.04]		[0.00]		[0.22]		[0.10]		[0.01]		[0.14]	
Strict exogeneity of $\{D_{it}\}$	[0.01]		[0.73]		[0.10]		[0.03]		[0.03]		[0.79]	
Observations	25,184		25,184		25,184		25,184		25,184		25,184	

Notes: Estimations are conducted using first differencing, and include complete sets of first-differenced dummies for number of previous interviews and first-differenced single-hour dummies for D_{it} interacted with MSF_{sc}^{e}. The dependent variables whose name start with P are binary indicators for the outcome given in the name's subscript scaled as a percentage. Standard errors take account of heteroskedasticity and clustering at individual level. Probability values are in brackets. *: Significant at 5%.

The estimated effects on P_{INR}, NumAct, and $P_{IVDUR5L}$, calculated by zeroing out all the controls and varying D_{it}, are depicted in Figure 5.1. The three graphs tell a rather consistent story: The quality of the data peaks in the early morning and is worst in the evening. The estimated change within the day is 0.30

percentage points (pps) for P_{INR}, 0.64 activities for NumAct, and 1.82 pps for $P_{IVDUR5L}$, representing 11, 4, and 36% of the corresponding mean.

Table 5.3
Time of day of interview effects on data quality.

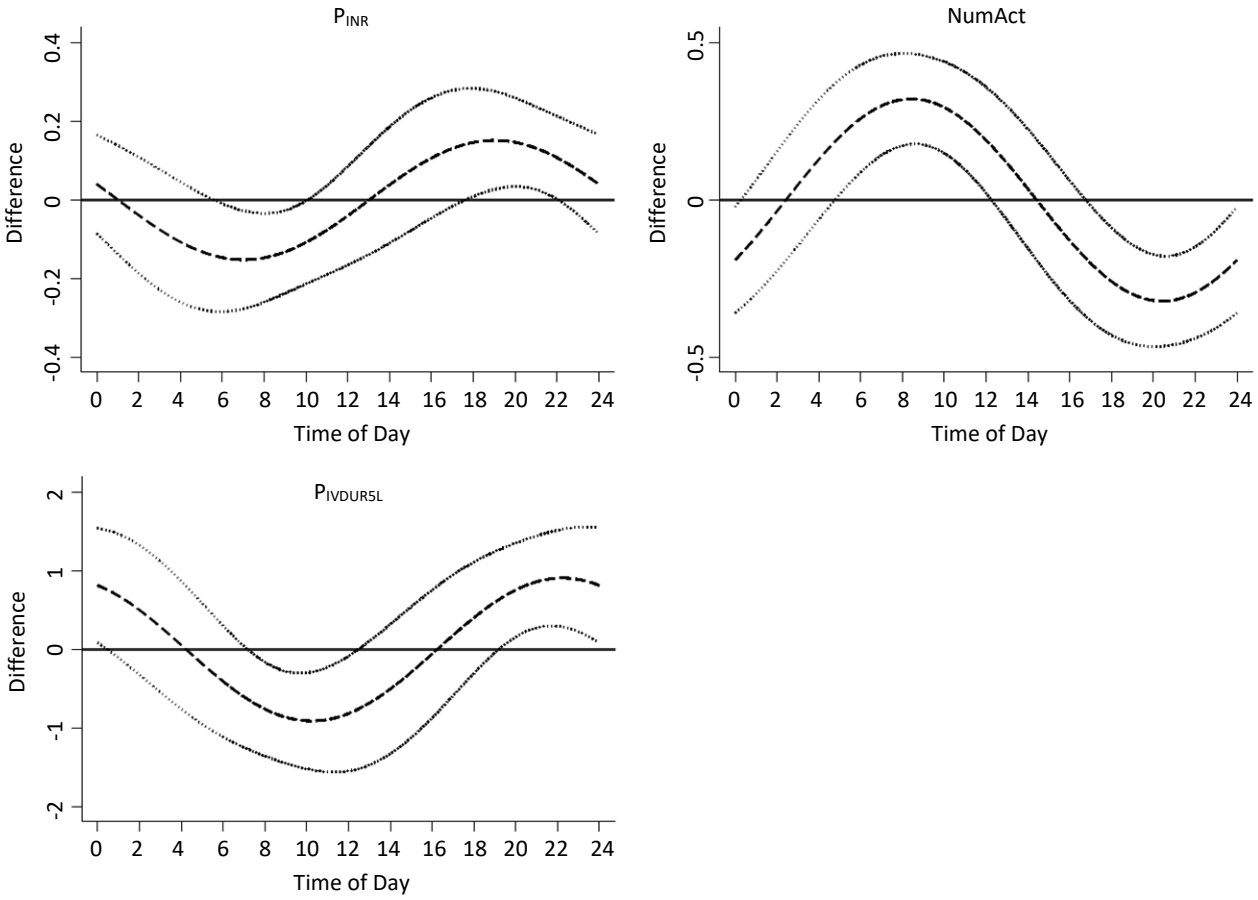
	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING-OUT50}$		$P_{EATING-OUT100}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59					-0.109	0.500					-1.594	1.108		
19:00–05:59					-0.023	0.675					-1.740	1.446		
12:00–16:59			0.589	0.589										
17:00–23:59			-0.502	0.733										
$D_{it}(1-(D_{it}/24)^2)$													0.839	0.449
$D_{it}^2(1-D_{it}/24)$													-0.032	0.045
$\sin(D_{it}\times 2\pi/24)$	-0.714	0.828					-0.277	0.898	-0.280	0.822				
$\cos(D_{it}\times 2\pi/24)$	0.820	0.820					-1.598	0.837	-0.883	0.799				
Tuesday	-0.531	1.210	0.176	0.770	-0.706	0.662	-1.212	1.276	0.833	1.239	-0.501	1.378	0.273	1.267
Wednesday	-1.326	1.361	-0.768	0.842	-1.661*	0.722	-1.028	1.433	0.720	1.389	1.332	1.552	3.122*	1.417
Thursday	-0.220	1.358	0.374	0.867	-0.873	0.737	-1.294	1.413	-0.873	1.388	0.669	1.575	1.909	1.416
Friday	-0.444	1.277	0.243	0.826	-0.455	0.663	-1.570	1.374	0.163	1.338	1.603	1.485	2.055	1.353
Saturday	-0.162	1.394	1.094	0.885	-0.068	0.721	-0.116	1.476	0.282	1.447	0.380	1.554	0.971	1.424
Sunday	-0.907	1.544	-0.408	0.920	-1.711*	0.776	-3.351*	1.569	-2.001	1.539	0.264	1.692	2.627	1.529
Worked	-2.307*	1.103	0.721	0.678	-0.160	0.573	-1.491	1.272	-2.048	1.238	-2.277	1.360	-0.664	1.201
Sleep duration	0.215	0.182	0.223	0.117	0.166	0.097	-0.378	0.194	-0.326	0.192	0.193	0.216	0.244	0.196
Significance of $\alpha(D_{it})$	[0.28]		[0.28]		[0.97]		[0.16]		[0.54]		[0.30]		[0.04]	
Strict exogeneity of $\{D_{it}\}$	[0.87]		[0.91]		[0.42]		[0.25]		[0.55]		[0.27]		[0.55]	
Observations	25,083		25,083		25,083		23,957		23,957		20,874		20,874	

Notes: See notes to Table 5.2.

The number of activities might be lower when the diary is completed in the evening due to the longer period of recall. To disentangle the effect of D_{it} from that of the recall period, the sample is split into weekday (Monday–Thursday) and weekend (Friday–Sunday) diaries. The results of re-estimating the equation for NumAct in each of the two subsamples of diaries are presented in Table 5.4. (Remember that the day indicated in the tables is the interview day.) $\alpha(D_{it})$ becomes insignificant in the subsample of weekend diaries, although this conclusion is partly driven by the imprecision of the estimates. Moreover, the extent of change within a weekend day comes out much smaller than within a weekday: 0.43 vs. 1.01 activities, representing 2.7 and 5.9% of the corresponding mean. Thus, a large extent of the daily rhythm of NumAct is driven by the period of recall.

As for the effects of the controls, the number of activities is higher in Monday–Thursday diaries, and interviews appear to be longer on Tuesdays, Wednesdays, and Thursdays. Working and sleeping longer on the diary day have contradictory effects on the quality of time-diary data, as they tend to reduce both the number of activities and the number of hours not coded. These effects are probably reflecting that working and sleeping longer reduce the time available for other activities, and the reduction of activities facilitates their recalling. Working on the diary day increases the likelihood that the interview is among the 5% shortest by 1.4 pps (or 28%).

Figure 5.1 Time of day of interview effects on data quality.



Notes: The effects (dashed lines) are calculated from the corresponding estimations in Table 5.2. Dotted lines delimit the 95% confidence interval.

Table 5.4
Time of day of interview effects on NumAct, by diary day.

	(1) Monday–Thursday diaries		(2) Friday–Sunday diaries	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**
$\sin(D_{it} \times 2\pi/24)$	0.476*	0.110	0.213	0.218
$\cos(D_{it} \times 2\pi/24)$	-0.171	0.112	-0.028	0.209
Monday			Ref.	
Tuesday	0.236	0.165		
Wednesday	0.099	0.172		
Thursday	0.088	0.165		
Friday	Ref.			
Saturday			0.238	0.263
Sunday			-0.415	0.304
Worked	-1.626*	0.213	-1.374*	0.345
Sleep duration	-0.093*	0.030	-0.021	0.057
Significance of $\alpha(D_{it})$	[0.00]		[0.59]	
Observations	14,904		3,073	

Notes: Estimations are conducted using first differencing, and include complete sets of first-differenced dummies for number of previous interviews and first-differenced single-hour dummies for D_{it} interacted with MSF_{sc}^{e}. Standard errors take account of heteroskedasticity and clustering at individual level. Probability values are in brackets. *: Significant at 5%.

5.3 Supplementary analyses

5.3.1 Strict exogeneity

We have been assuming that the variation in D_{it} within respondents is strictly exogenous. This assumption would be questioned if, for example, respondents rush through the survey or become distracted at times of day when the opportunity cost of completing the interview is highest. The p-value for the test of strict exogeneity of $\{D_{it}: t = 1, \ldots, T_i\}$ is shown in the next-to-last row of Tables 5.2 and 5.3. At 5% level, exogeneity is questioned in the regressions for P_{INR}, completion time (IvDur), and $P_{IVDUR5L}$. Since the within (or fixed effects) estimator tends to be more robust to the violation of strict exogeneity, we re-estimated equation (4.1) with the OLS estimator from the regression

$$y_{it} - \overline{y}_i = \left(\alpha\left(D_{it}\right) - \overline{\alpha\left(D_i\right)}\right) + \left(\mathbf{x}_{it} - \overline{\mathbf{x}}_i\right)\boldsymbol{\beta} + u_{it} - \overline{u}_i \tag{5.1}$$

where $\overline{y}_i = T_i^{-1}\sum_{t=1}^{T_i} y_{it}$, $\overline{\alpha\left(D_i\right)} = T_i^{-1}\sum_{t=1}^{T_i}\alpha\left(D_{it}\right)$, and so on. The null hypothesis H_0: $\alpha\left(D_{it}\right) = 0$ is rejected in the regression for P_{INR} (p-value 0.01), but not rejected in the regressions for IvDur and $P_{IVDUR5L}$ (p-value 0.39 in both cases). Note, however, that both the first-difference and the within estimators may be biased when strict exogeneity fails.

5.3.2 Robustness

The estimates change little when sleep duration is excluded from $\mathbf{x}_{it}$, or when D_{it} is approximated by the end time of the time-use section of the questionnaire or by randomly selected points within the start and end times of the interview (results not shown). When the 2,578 interviews presenting inconsistent going-to-bed time are included in the sample, the preferred model for $\alpha\left(D_{it}\right)$ changes in some cases (Table A.1 in the Appendix). A statistically significant $\alpha\left(D_{it}\right)$ is detected in the regressions for P_{INR}, NumAct, HMissing, and IvDur, whereas $\alpha\left(D_{it}\right)$ becomes insignificant in the regression for $P_{IVDUR5L}$ (Tables A.2 and A.3 in the appendix). When an effect is detected, it suggests that data quality peaks in the early morning.

5.3.3 Attrition

Table 5.5 presents probit estimation output for the decision to complete an interview. It shows selected $\boldsymbol{\delta}$ coefficients plus average marginal effects (AMEs) calculated by averaging marginal effects across observations. Completing the $t-1$ interview on Tuesday–Saturday increases the probability of completing the t interview. Working on the diary day increases that probability by 1.8 pps, whereas one more hour of sleep reduces it by 0.6 pps. The number of weeks passed between $t-2$ and $t-1$ is a strong predictor for completing the t interview, whose likelihood reduces by 3.0 pps with every week passed. None of the single-hour dummies for $D_{i,t-1}$ attains significance at 5% (not shown).

After correcting for nonrandom attrition, the cosinor model becomes the preferred option for analyzing P_{MOOD50}, while Binder's (2022) piecewise constant function comes out as the best fitting alternative for the probability of reporting expenditure on food at home in multiples of 100 ($P_{FOODAH100}$). The null hypothesis of no attrition bias is questioned in the regressions for IvDur, $P_{IVDUR5L}$, and $P_{FOODAH100}$. However, the

attrition-corrected estimates (reported in Tables A.4 and A.5 in the appendix) reveal essentially the same patterns as the non-attrition-corrected ones. The correction for nonrandom attrition makes less dubious the assumption of strict exogeneity of $\alpha(D_{it})$ in the regressions for IvDur and $P_{IVDUR5L}$ (p-value 0.12 in both cases).

Table 5.5
Probit for interview completion.

	Dependent variable: s_{it}, $t \geq 2$			
Explanatory variables $(t-1)$	**Coef.**	**S.E.**	**AME**	**S.E.**
Tuesday	0.154*	0.034	0.034*	0.008
Wednesday	0.175*	0.038	0.039*	0.008
Thursday	0.184*	0.038	0.040*	0.008
Friday	0.215*	0.037	0.047*	0.008
Saturday	0.174*	0.039	0.038*	0.009
Sunday	0.058	0.042	0.014	0.010
Worked	0.089*	0.028	0.018*	0.005
Sleep duration	-0.027*	0.004	-0.006*	0.001
Weeks between $t-2$ and $t-1$	-0.145*	0.006	-0.030*	0.001
Intercept	1.121*	0.105		
R-squared		0.070		
Observations		32,779		
Mean of s_{it}		0.859		

Notes: Observations for the last interview are excluded because individuals did certainly not continue in the survey. Includes single-hour dummies for $D_{i,t-1}$, dummies for number of previous interviews, and MSF^{e}_{sc} interacted with single-hour dummies for $D_{i,t-1}$. Standard errors take account of heteroskedasticity and clustering at individual level. R-squared equals one minus the ratio of the log likelihood of the fitted function to the log likelihood of a function with only an intercept. *: Significant at 5%.

5.3.4 Weights

Tables 5.6 and 5.7 present the WLS estimates. A statistically significant $\alpha(D_{it})$ is not detected in most of the regressions shown. While in some cases (e.g., the regression for NumAct), the WLS estimated coefficients are smaller than the OLS ones, in most cases the inference is driven by the larger standard errors. A statistically significant $\alpha(D_{it})$ is detected in the regression for $P_{IVDUR5H}$ (p-value 0.03), but this effect does not survive a Bonferroni correction for simultaneous tests in the group of measures assessing completion time. The null of no effect is also rejected at 5% in the regressions for $P_{EATING\text{-}OUT50}$ and $P_{EATING\text{-}OUT100}$ (p-value 0.01 in both cases), but in both cases the rejection of the null does not hold if zero expenditure is assumed not to reflect rounding (p-values 0.43 and 0.35 respectively).

5.3.5 Subpopulations

Finally, we split the sample by educational attainment (at most some college vs. college diploma) to investigate time of day of interview effects with certain types of individuals. Although cognitive abilities are important predictors of educational attainment, we do not expect to find big differences between demographic groups as our estimates are net of synchrony and cognitive ability effects. Indeed, although the best-fitting model of $\alpha(D_{it})$ changes for most of the dependent variables in both subpopulations, the main conclusions are preserved (results not shown).

Table 5.6
Time of day of interview effects on data quality. Weighted estimates.

	(1) P_{INR}		(2) NumAct		(3) HMissing		(4) IvDur (min)		(5) $P_{IVDUR5L}$		(6) $P_{IVDUR5H}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59					0.224	0.124						
19:00–05:59					0.063	0.113						
$D_{it}(1-(D_{it}/24)^2)$											-0.993*	0.474
$D_{it}^2(1-D_{it}/24)$											0.153*	0.058
$\sin(D_{it}\times 2\pi/24)$	-0.298	0.249	0.061	0.207			-0.581*	0.295	0.363	0.662		
$\cos(D_{it}\times 2\pi/24)$	0.133	0.139	-0.111	0.157			-0.314	0.293	0.082	1.023		
Tuesday	-0.241	0.242	1.222*	0.260	-0.133	0.093	1.341*	0.374	-0.299	1.306	1.439	1.417
Wednesday	-0.124	0.334	1.248*	0.359	-0.185	0.138	1.028*	0.438	0.077	2.241	0.812	1.507
Thursday	0.483	0.331	0.881*	0.312	0.020	0.098	1.025*	0.409	0.543	1.653	1.879	1.619
Friday	0.512	0.336	0.703*	0.291	0.104	0.127	0.621	0.431	1.067	1.422	0.262	1.533
Saturday	0.354	0.291	-0.009	0.411	0.038	0.128	-0.243	0.483	1.911	2.031	-1.101	1.531
Sunday	-0.034	0.358	0.109	0.339	-0.075	0.130	-0.959*	0.485	0.381	1.556	-3.991*	1.663
Worked	0.200	0.250	-2.582*	0.260	-0.158	0.084	-0.706*	0.320	3.005*	1.494	-1.619	1.025
Sleep duration	-0.057	0.043	-0.077	0.046	-0.030	0.017	-0.097	0.062	0.656*	0.209	0.040	0.276
Significance of $\alpha(D_{it})$	[0.38]		[0.69]		[0.14]		[0.14]		[0.84]		[0.03]	
Strict exogeneity of $\{D_{it}\}$	[0.26]		[0.92]		[0.36]		[0.02]		[0.25]		[0.23]	
Observations	25,184		25,184		25,184		25,184		25,184		25,184	

Notes: Estimations are conducted using first differencing, and include complete sets of first-differenced dummies for number of previous interviews and first-differenced single-hour dummies for D_{it} interacted with MSF_{sc}^{e}. The dependent variables whose name start with P are binary indicators for the outcome given in the name's subscript scaled as a percentage. Standard errors take account of heteroskedasticity and clustering at individual level. Probability values are in brackets. *: Significant at 5%.

Table 5.7
Time of day of interview effects on data quality. Weighted estimates.

	(1) P_{MOOD10}		(2) P_{MOOD25}		(3) P_{MOOD50}		(4) $P_{FOODAH50}$		(5) $P_{FOODAH100}$		(6) $P_{EATING-OUT50}$		(7) $P_{EATING-OUT100}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59					-1.514	1.575					-1.255	2.295		
19:00–05:59					0.127	1.500					-8.426*	3.040		
12:00–16:59			-1.515	1.582										
17:00–23:59			-3.035*	1.436										
$D_{it}(1-(D_{it}/24)^2)$													1.074	0.889
$D_{it}^2(1-D_{it}/24)$													0.037	0.096
$\sin(D_{it}\times 2\pi/24)$	-1.143	1.758					-2.348	2.000	-2.029	2.399				
$\cos(D_{it}\times 2\pi/24)$	0.236	2.192					-4.367	2.250	-4.183	2.698				
Tuesday	4.451	2.827	0.293	1.508	-0.836	1.199	2.645	2.530	3.310	2.620	-5.478	2.900	-2.843	2.732
Wednesday	1.520	2.894	-0.776	1.728	-0.807	1.502	1.089	3.120	0.796	3.130	-0.336	3.387	3.425	3.026
Thursday	5.568	3.929	0.631	1.738	-0.993	1.292	3.577	2.871	2.227	3.014	0.251	4.177	1.842	3.670
Friday	7.726*	3.824	-0.968	1.819	-1.234	1.152	2.891	3.447	3.313	3.682	4.003	3.041	2.739	2.822
Saturday	3.580	3.349	0.399	1.659	-0.234	1.321	3.190	3.127	3.793	3.692	-0.065	3.192	1.163	3.009
Sunday	5.240	3.578	-0.992	1.842	-2.572	1.450	2.211	3.327	0.165	3.645	-1.928	3.554	0.044	3.236
Worked	-3.964	2.072	0.514	1.111	-1.931*	0.918	-0.547	2.809	-0.775	2.856	-3.397	2.839	-2.032	2.383
Sleep duration	0.282	0.361	0.174	0.219	0.149	0.178	-0.855*	0.368	-0.576	0.440	0.330	0.441	0.280	0.432
Significance of $\alpha(D_{it})$	[0.77]		[0.11]		[0.35]		[0.09]		[0.28]		[0.01]		[0.01]	
Strict exogeneity of $\{D_{it}\}$	[0.60]		[0.68]		[0.97]		[0.13]		[0.34]		[0.41]		[0.92]	
Observations	25,083		25,083		25,083		23,957		23,957		20,874		20,874	

Notes: See notes to Table 5.6.

6. Summary and discussion

The analysis of high-frequency longitudinal microdata from the SUWNJ reveals no evidence of a time of day of interview effect on the quality of time-diary data (beyond the effect exerted by the length of the recall period), or on the tendency to report rounded values of subjective probabilities or food expenditure. As regards the period of recall, we found that self-completing a yesterday diary in the evening reduces the number of activities reported, whereas the amount of time not coded suffers no meaningful daily fluctuation. Thus, it appears that some activities are underreported and the duration of others is overestimated, introducing error in the measurement of the use of time. All these findings have been developed accounting for inter-individual differences in cognitive ability and synchrony effects, which may explain why they persist across education groups. They also appear to be robust to a range of alternative specifications assessing the impact of nonrandom attrition, unmodeled heterogenous effects, and different measures of time of day of interview. Although there is some evidence to indicate that item nonresponse and the probability that interview completion time is among the 5% shortest increase when the survey is completed in the evening, a more thorough assessment requires instrumental variables.

Our most reliable results support the conclusion of previous research that survey data quality is insensitive to the time of day of interview (Ziniel, 2008; Dickinson and McElroy, 2010; Binder, 2022), but disagree with those of Flynn (2018), who found that respondents who start a survey in the evening answer significantly more questions than those who start it in the morning/afternoon. Yet, Flynn's (2018) sample is made up of firm representatives, and completing a survey outside of regular office hours might benefit from reduced time pressures. As the unemployed (as compared to the employed) do not have to adhere to the limitations of work hours, their time of day of interview can be more evenly spread over the 24 hours, facilitating the identification of effects around the clock. It is also worth noting that, in contrast to MTurk samples (e.g., Binder, 2022), interviews appear to be longer on Thursdays (plus Tuesdays and Wednesdays), and that the number of activities reported is higher in Monday–Thursday diaries as in Juster (1986).

Overall, therefore, it appears that beyond the effect exerted by the length of the recall period, inducing respondents to complete surveys at specific times of the day might have limited impacts on measurement error. Thus, survey practitioners should not worry much about the consequences for measurement error of seeking to interview subjects at times of the day they are most likely to be contactable.

All that said, we recognize some limitations of this study. As regards the question of whether we uncover causal effects for the population being studied, it must be noted that we lack data on the situational context in which the interviews were completed (e.g., where the respondent was and what he/she was doing), and as argued by Bison and Zhao (2023) the temporal and situational contexts might be correlated. However, it is difficult to suggest instrumental variables sufficiently correlated with time of day of interview but uncorrelated with idiosyncratic errors, as most variables in the SUWNJ refer to days other than the interview day. Also, although the percentage of SUWNJ interviews completed from a mobile device must have been low (Callegaro, 2010, for example, reports that among all respondents who attempted to complete an online customer satisfaction survey conducted in North America in June 2010, 2.6% did so from a mobile device),

if completing an interview from a mobile device affects the quality of the data (as the evidence reviewed in Toninelli and Revilla, 2020 suggests) and depends on the time of day, our results might contain bias. As regards the predictive value of our findings in a different context, it must be noted that the results obtained for the unemployed might not be representative for broader populations if, for example, the activities conducted before completing the survey interact with sleepiness/fatigue.

In addition, insufficient data prevented us from investigating the existence of time of day of interview effects on alternative measures of data quality, such as survey breakoff and response errors caused by social desirability or extreme, midpoint, or nondifferentiated responding. As regards the effects of the length of the recall period, it seems worth investigating whether the administration of a yesterday diary by an interviewer (who could foster respondents' attention and motivation), or the "own words" reporting of activities by respondents (which avoids the process of mapping the answer onto the appropriate response option), could improve the quality of time-diary data.

Acknowledgements

This paper has greatly benefited from the comments and suggestions of Jean-Francois Beaumont, Kristen Olson, and, especially, an anonymous Associate Editor and several anonymous Referees. Thanks also to Andreas Mueller for assistance with SUWNJ data. This study was supported by the Government of Aragón, grant S32-23R.

Appendix

Table A.1
Model selected for $\alpha(D_{it})$. Including observations with inconsistent going-to-bed time.

Dependent variable	Model	BIC value
P_{INR}	Piecewise constant (Binder, 2022)	13.405
NumAct	Cosinor	13.731
HMissing	Piecewise constant (Binder, 2022)	11.271
IvDur	Piecewise constant (Binder, 2022)	14.460
$P_{IVDUR5L}$	Cosinor	16.590
$P_{IVDUR5H}$	Degree 3 polynomial	16.849
P_{MOOD10}	Piecewise constant (Binder, 2022)	18.162
P_{MOOD25}	Piecewise constant (Durrant et al., 2011)	17.211
P_{MOOD50}	Cosinor	16.820
$P_{FOODAH50}$	Degree 3 polynomial	18.226
$P_{FOODAH100}$	Piecewise constant (Binder, 2022)	18.145
$P_{EATING\text{-}OUT50}$	Piecewise constant (Binder, 2022)	18.080
$P_{EATING\text{-}OUT100}$	Degree 3 polynomial	17.887

Table A.2
Time of day of interview effects on data quality. Including observations with inconsistent going-to-bed time.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{INR}		NumAct		HMissing		IvDur (min)		$P_{IVDUR5L}$		$P_{IVDUR5H}$		P_{PM-AM}	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59	0.203*	0.073			0.070*	0.027	-0.368*	0.140						
19:00–05:59	0.221*	0.101			0.054	0.036	-0.461*	0.189						
$D_{it}(1-(D_{it}/24)^2)$											-0.106	0.219		
$D_{it}^2(1-D_{it}/24)$											0.027	0.022		
$\sin(D_{it}\times 2\pi/24)$			0.264*	0.084					-0.099	0.355			0.611	0.499
$\cos(D_{it}\times 2\pi/24)$			-0.184*	0.080					0.427	0.333			-1.069*	0.495
Tuesday	-0.177	0.101	1.420*	0.131	-0.086*	0.037	0.772*	0.184	-0.160	0.447	1.114	0.602	-2.738*	0.696
Wednesday	-0.145	0.109	1.180*	0.141	-0.056	0.039	0.434*	0.196	-0.672	0.502	0.075	0.635	-2.417*	0.763
Thursday	0.024	0.121	1.247*	0.150	-0.011	0.041	0.780*	0.213	-0.469	0.519	1.255	0.724	-1.503	0.788
Friday	-0.009	0.113	1.091*	0.145	-0.057	0.042	0.355	0.196	0.095	0.476	0.241	0.626	-2.161*	0.748
Saturday	0.072	0.121	0.755*	0.147	-0.002	0.041	0.507*	0.233	0.838	0.561	0.482	0.755	-1.156	0.815
Sunday	0.270*	0.136	-0.025	0.163	0.029	0.046	-0.283	0.228	-0.199	0.613	-0.538	0.728	-0.759	0.879
Worked	0.369*	0.129	-1.687*	0.152	-0.123*	0.035	-0.206	0.164	1.401*	0.565	0.454	0.531	-2.210*	0.659
Sleep duration	-0.028	0.018	-0.056*	0.023	-0.015*	0.007	-0.105*	0.026	0.361*	0.082	-0.131	0.090	-1.803*	0.124
Significance of $\alpha(D_{it})$	[0.01]		[0.00]		[0.03]		[0.01]		[0.35]		[0.30]		[0.02]	
Strict exogeneity of $\{D_{it}\}$	[0.02]		[0.61]		[0.17]		[0.57]		[0.01]		[0.64]		[0.65]	
Observations	28,576		28,576		28,576		28,576		28,576		28,576		28,576	

Notes: Estimations are conducted using first differencing, and include complete sets of first-differenced dummies for number of previous interviews and first-differenced single-hour dummies for D_{it} interacted with MSF_{sc}^{c}. The dependent variables whose name start with P are binary indicators for the outcome given in the name's subscript scaled as a percentage. Standard errors take account of heteroskedasticity and clustering at individual level. Probability values are in brackets. *: Significant at 5%.

A

Table A.3
Time of day of interview effects on data quality. Including observations with inconsistent going-to-bed time.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING-OUT50}$		$P_{EATING-OUT100}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59	1.065	0.895							-0.600	0.911	-1.238	1.043		
19:00–05:59	2.380*	1.197							-0.187	1.204	-0.950	1.373		
12:00–16:59			0.525	0.573										
17:00–23:59			0.206	0.686										
$D_{it}(1-(D_{it}/24)^2)$							0.283	0.451					0.578	0.411
$D_{it}^2(1-D_{it}/24)$							0.022	0.047					-0.022	0.041
$\sin(D_{it}\times 2\pi/24)$					0.111	0.440								
$\cos(D_{it}\times 2\pi/24)$					0.511	0.374								
Tuesday	0.208	1.142	-0.282	0.723	-0.864	0.619	-0.697	1.205	0.729	1.168	-0.468	1.312	0.398	1.194
Wednesday	-0.735	1.277	-0.951	0.791	-1.703*	0.672	-0.453	1.363	0.947	1.320	1.463	1.471	3.295*	1.339
Thursday	-0.176	1.270	0.191	0.814	-0.966	0.692	-1.183	1.351	-1.337	1.305	-0.120	1.473	1.251	1.322
Friday	0.034	1.207	-0.163	0.782	-0.636	0.623	-1.346	1.313	0.104	1.266	1.540	1.405	2.252	1.279
Saturday	0.078	1.300	0.353	0.840	-0.369	0.677	-0.448	1.392	-0.298	1.352	0.382	1.471	1.028	1.342
Sunday	-0.489	1.440	-0.529	0.876	-1.590*	0.727	-2.606	1.462	-1.390	1.451	0.611	1.607	2.820	1.441
Worked	-2.167*	1.067	0.377	0.659	-0.363	0.542	-1.344	1.200	-2.024	1.162	-2.238	1.277	-0.372	1.147
Sleep duration	0.152	0.171	0.200	0.110	0.115	0.095	-0.401*	0.179	-0.336	0.178	0.139	0.204	0.190	0.187
Significance of $\alpha(D_{it})$	[0.13]		[0.65]		[0.39]		[0.10]		[0.79]		[0.49]		[0.17]	
Strict exogeneity of $\{D_{it}\}$	[0.64]		[0.99]		[0.79]		[0.25]		[0.25]		[0.42]		[0.93]	
Observations	28,461		28,461		28,461		27,187		27,187		23,612		23,612	

Notes: See notes to Table A.2.

Table A.4
Time of day of interview effects on data quality. Attrition-corrected estimates.

	(1)		(2)		(3)		(4)		(5)		(6)	
	P_{INR}		NumAct		HMissing		IvDur (min)		$P_{IVDUR5L}$		$P_{IVDUR5H}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59					0.048	0.029						
19:00–05:59					0.046	0.039						
$D_{it}(1-(D_{it}/24)^2)$											-0.157	0.238
$D_{it}^2(1-D_{it}/24)$											0.038	0.024
$\sin(D_{it}\times 2\pi/24)$	-0.143*	0.070	0.256*	0.087			0.026	0.135	-0.457	0.364		
$\cos(D_{it}\times 2\pi/24)$	0.041	0.064	-0.191*	0.086			-0.293*	0.138	0.827*	0.374		
Tuesday	-0.175	0.113	1.475*	0.142	-0.096*	0.040	0.944*	0.202	-0.505	0.501	1.100	0.658
Wednesday	-0.147	0.121	1.209*	0.154	-0.074	0.043	0.645*	0.211	-1.195*	0.559	0.048	0.680
Thursday	0.033	0.136	1.231*	0.163	-0.040	0.044	0.858*	0.231	-0.681	0.576	1.243	0.774
Friday	0.003	0.123	1.130*	0.155	-0.070	0.044	0.470*	0.212	-0.052	0.512	-0.055	0.673
Saturday	0.023	0.127	0.755*	0.155	-0.022	0.042	0.542*	0.249	0.319	0.610	0.187	0.795
Sunday	0.289*	0.140	0.058	0.172	0.003	0.045	-0.336	0.242	0.121	0.661	-0.896	0.762
Worked	0.414*	0.137	-1.720*	0.158	-0.118*	0.037	-0.117	0.174	1.282*	0.608	0.717	0.556
Sleep duration	-0.025	0.018	-0.079*	0.025	-0.016*	0.007	-0.142*	0.028	0.416*	0.088	-0.188*	0.095
Attrition bias	[0.91]		[0.20]		[0.29]		[0.00]		[0.01]		[0.38]	
Significance of $\alpha(D_{it})$	[0.04]		[0.00]		[0.24]		[0.07]		[0.01]		[0.16]	
Strict exogeneity of $\{D_{it}\}$	[0.00]		[0.54]		[0.08]		[0.12]		[0.12]		[0.95]	
Observations	25,184		25,184		25,184		25,184		25,184		25,184	

Notes: Estimations are conducted using first differencing, and include a complete set of first-differenced dummies for number of previous interviews, first-differenced single-hour dummies for D_{it} interacted with MSF_{sc}^{e}, and the inverse Mills ratio interacted with dummies for interview number. The dependent variables whose name start with P are binary indicators for the outcome given in the name's subscript scaled as a percentage. Standard errors take account of heteroskedasticity and clustering at individual level and correct for generated regressors. Probability values are in brackets. *: Significant at 5%.

Table A.5
Time of day of interview effects on data quality. Attrition-corrected estimates.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING\text{-}OUT50}$		$P_{EATING\text{-}OUT100}$	
Explanatory variables	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**	**Coef**	**S.E.**
12:00–18:59									-1.027	0.994	-1.642	1.113		
19:00–05:59									-1.107	1.281	-1.824	1.446		
12:00–16:59			0.621	0.589										
17:00–23:59			-0.467	0.734										
$D_{it}(1-(D_{it}/24)^2)$													0.836	0.448
$D_{it}^2(1-D_{it}/24)$													-0.031	0.045
$\sin(D_{it}\times 2\pi/24)$	-0.741	0.831			-0.039	0.458	-0.316	0.899						
$\cos(D_{it}\times 2\pi/24)$	0.827	0.820			0.084	0.400	-1.588	0.836						
Tuesday	-0.758	1.250	0.140	0.803	-0.844	0.685	-1.400	1.328	0.792	1.290	-0.348	1.429	0.201	1.304
Wednesday	-1.539	1.394	-0.823	0.874	-1.789*	0.744	-1.175	1.486	0.745	1.443	1.505	1.609	3.053*	1.459
Thursday	-0.437	1.387	0.324	0.888	-1.011	0.749	-1.447	1.445	-0.919	1.429	0.822	1.620	1.867	1.451
Friday	-0.621	1.305	0.183	0.850	-0.570	0.679	-1.740	1.416	0.180	1.379	1.759	1.534	2.022	1.389
Saturday	-0.342	1.409	1.065	0.903	-0.162	0.733	-0.280	1.501	0.240	1.475	0.494	1.589	0.907	1.446
Sunday	-0.976	1.544	-0.396	0.926	-1.741*	0.779	-3.374*	1.575	-2.103	1.543	0.286	1.702	2.611	1.533
Worked	-2.281*	1.106	0.653	0.684	-0.214	0.576	-1.511	1.277	-1.992	1.241	-2.132	1.369	-0.491	1.203
Sleep duration	0.221	0.184	0.232*	0.117	0.176	0.097	-0.366	0.195	-0.333	0.193	0.176	0.216	0.247	0.195
Attrition bias	[0.79]		[0.30]		[0.26]		[0.67]		[0.04]		[0.96]		[0.08]	
Significance of $\alpha(D_{it})$	[0.27]		[0.28]		[0.97]		[0.16]		[0.54]		[0.28]		[0.04]	
Strict exogeneity of $\{D_{it}\}$	[0.85]		[0.98]		[0.68]		[0.25]		[0.61]		[0.33]		[0.50]	
Observations	25,083		25,083		25,083		23,957		23,957		20,874		20,874	

Notes: See notes to Table A.4.

References

Ahn, J., Peng, M., Park, C. and Jeon, Y. (2012). A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining*, 5, 336-348.

American Association for Public Opinion Research (AAPOR) (2023). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, 10th Edition.* AAPOR.

Angrisani, M., and Couper, M. (2022). A simple question goes a long way: A wording experiment on bank account ownership. *Journal of Survey Statistics and Methodology*, 10, 1172-1182.

Arechar, A., Kraft-Todd, G. and Rand, D. (2017). Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, 3, 1-11.

Arellano, M., and Meghir, C. (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *Review of Economic Studies*, 59, 537-557.

Bach, R. (2021). A methodological framework for the analysis of panel conditioning effects. In *Measurement Error in Longitudinal Data*, (Eds., Alexandru Cernat and Joseph Sakshaug), 19-42. Oxford: OUP.

Bais, F., Schouten, B. and Toepoel, V. (2022). Is undesirable answer behaviour consistent across surveys? An investigation into respondent characteristics. *Survey Methodology*, 48, 1, 191-224. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022001/article/00001-eng.pdf.

Baumgartner, H., and Steenkamp, J.-B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156.

Bes, F., Jobert, M. and Schulz, H. (2009). Modeling napping, post-lunch dip, and other variations in human sleep propensity. *Sleep*, 32(3), 392-398.

Biemer, P., Groves, R., Lyberg, L. Mathiowetz, N. and Sudman, S. (eds.) (2004). *Measurement Errors in Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Binder, C. (2022). Time-of-day and day-of-week variations in Amazon Mechanical Turk survey responses. *Journal of Macroeconomics*, 71, Article 103378.

Bison, I., and Zhao, H. (2023). Factors impacting the quality of user answers on smartphones. *CEUR Workshop Proceedings*, 3456, 208-213.

Blatter, K., and Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology and Behavior*, 90, 196-208.

Brown, E., and Czeisler, C. (1992). The statistical analysis of circadian phase and amplitude in constant-routine core-temperature data. *Journal of Biological Rhythms*, 7(3), 177-202.

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*, 3(6). https://doi.org/10.29115/SP-2010-0028.

Carrell, S., Maghakian, T. and West, J. (2011). A's from zzzz's? The causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3, 62-81.

Casey, L., Chandler, J., Levine, A., Proctor, A. and Strolovitch, D. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7(2), 1-15.

Chang, L., and Krosnick, J. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73(4), 641-678.

Collinson, J., Mathmann, F. and Chylinski, M. (2020). Time is money: Field evidence for the effect of time of day and product name on product purchase. *Journal of Retailing and Consumer Services*, 54, Article 102064.

Cornelissen, G. (2014). Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling*, 11, Article 16.

Davis, S. (2011). Comment on: Krueger, A., and A. Mueller. Job search, emotional well-being and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 42(1), 58-70.

Dickinson, D., and McElroy, T. (2010). Rationality around the clock: Sleep and time-of-day effects on guessing game responses. *Economics Letters*, 108, 245-248.

Dickinson, D., and McElroy, T. (2017). Sleep restriction and circadian effects on social decisions. *European Economic Review*, 97, 57-71.

Dickinson, D., Chaudhuri, A.·and Greenaway-McGrevy, R. (2020). Trading while sleepy? Circadian mismatch and mispricing in a global experimental asset market. *Experimental Economics*, 23, 526-553.

Durrant, G., D'Arrigo, J. and Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society. Series A*, 174(4), 1029-1049.

Flynn, A. (2018). e-Surveying and respondent behaviour: Insights from the public procurement field. *Electronic Journal of Business Research Methods*, 16(1), 38-53.

Fordsham, N., Moss, A., Krumholtz, S., Roggina, T., Robinson, J. and Litman, L. (2019). Variation among Mechanical Turk workers across time of day presents an opportunity and a challenge for research. PsyArXiv. doi:10.31234/osf.io/p8bns.

Fricker, S., and Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5), 934-955.

Gideon, M., Helppie-McFall, B. and Hsu, J. (2017). Heaping at round numbers on financial questions: The role of satisficing. *Survey Research Methods*, 11(2), 189-214.

González Chapela, J. (2024). Supplement to "Daily rhythm of data quality: Evidence from the survey of unemployed workers in New Jersey". Available at https://drive.google.com/file/d/14YPt9BmXlxFfuURatCQY0ak-OW9z1c7B/view?usp=sharing.

Guarana, C., Stevenson, R. Gish, J. Ryu, J.W. and Crawley, R. (2022). Owls, larks, or investment sharks? The role of circadian process in early-stage investment decisions. *Journal of Business Venturing*, 37, Article 106165.

Hasher, L., Goldstein, D. and May, C. (2005). It's about time: Circadian rhythms, memory, and aging. In *Human Learning and Memory: Advances in Theory and Applications*, (Eds., Chizuko Izawa and Nobuo Ohta), 199-217. New York: Lawrence Erlbaum Associates Publishers.

Hornik, J., and Tal, A. (2010). The effect of synchronizing consumers' diurnal preferences with time of response on data reliability. *Marketing Letters*, 21, 1-15.

Juster, T. (1986). Response errors in the measurement of time use. *Journal of the American Statistical Association*, 81(394), 390-402.

Kaminska, O., McCutcheon, A. and Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74(5), 956-984.

Kroh, M., Winter, F. and Schupp, J. (2016). Using person-fit measures to assess the impact of panel conditioning on reliability. *Public Opinion Quarterly*, 80(4), 914-942.

Kroh, M., Lüdtke, D., Düzel, S. and Winter, F. (2016). Response error in a web survey and a mailed questionnaire: The role of cognitive functioning. SOEPpapers on Multidisciplinary Panel Data Research, No. 888. DIW, Berlin.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Krosnick, J. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.

Krueger, A., and Mueller, A. (2010). *Survey of Unemployed Workers in New Jersey* (version Nov. 12, 2013) [Data set]. Data Archive at the Office of Population Research, Princeton University. https://oprdata.princeton.edu/archive/njui/.

Krueger, A., and Mueller, A. (2011). Job search, emotional well-being and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 42(1), 1-57.

Lowe, C., Safati, A., and Hall, P. (2017). The neurocognitive consequences of sleep restriction: A meta-analytic review. *Neuroscience and Biobehavioral Reviews*, 80, 586-604.

Lyberg, L., and Stukel, D. (2017). The roots and evolution of the total survey error concept. In *Total Survey Error in Practice*, (Eds., Paul Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker and Brady West), 1-22. Hoboken, NJ: John Wiley & Sons, Inc.

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, 72(5), 914-934.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27, 392-403.

Olson, K., Smyth, J. and Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, 7, 275-308.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, 73(1), 74-97.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2), 317-337.

Phillips, A., and Stenger, R. (2022). The effect of burdensome survey questions on data quality in an omnibus survey. *Journal of Official Statistics*, 38(4), 1019-1050.

Read, B., Wolters, L. and Berinsky, A. (2021). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*. Available at https://doi.org/10.1017/pan.2021.32.

Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M. and Merrow, M. (2007). Epidemiology of the human circadian clock. *Sleep Medicine Reviews*, 11, 429-438.

Salehinejad, M., Wischnewski, M., Ghanavati, E., Mosayebi-Samani, M., Kuo, M.-F. and Nitsche, M. (2021). Cognitive functions and underlying parameters of human brain physiology are associated with chronotype. *Nature Communications*, 12, Article 4672.

Schmidt, C., Collette, F., Cajochen, C. and Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology*, 24(7), 755-789.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.

Solon, G., Haider, S. and Wooldridge, J. (2015). What are we weighting for? *Journal of Human Resources*, 50(2), 301-316.

StataCorp (2019). *Stata Base Reference Manual. Release 16*. College Station, TX: Stata Press.

Toninelli, D., and Revilla, M. (2020). How mobile device screen size affects data collected in web surveys. In *Advances in Questionnaire Design, Development, Evaluation and Testing*, (Eds., Paul Beatty, Debbie Collins, Lyn Kaye, Jose Luis Padilla, Gordon Willis and Amanda Wilmot), 349-373. Hoboken, NJ: John Wiley & Sons, Inc.

Tourangeau, R., Rips, L. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Truebner, M. (2021). The dynamics of "neither agree nor disagree" answers in attitudinal questions. *Journal of Survey Statistics and Methodology*, 9, 51-72.

Valdez, P. (2019). Homeostatic and circadian regulation of cognitive performance. *Biological Rhythm Research*, 50, 85-93.

Weeks, M., Kulka, R. and Pierson, S. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, 51, 540-549.

Williams, K., and Shapiro, T.M. (2018). Academic achievement across the day: Evidence from randomized class schedules. *Economics of Education Review*, 67, 158-170.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Second edition. Cambridge, MA: MIT Press.

Yan, T., and Olson, K. (2013). Analyzing paradata to investigate measurement error. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, (Ed., Frauke Kreuter), 73-95. Hoboken, NJ: John Wiley & Sons, Inc.

Ziniel, S. (2008). *Cognitive Aging and Survey Measurement*. PhD dissertation, University of Michigan.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 543-551
Statistics Canada, Catalogue No. 12-001-X

Exploring a skewness conjecture: Expanding Cochran's rule to a proportion estimated from a complex sample

Phillip S. Kott and Burton Levine[1]

Abstract

Cochran's rule states that a standard (Wald) two-sided 95% confidence interval around a sample mean drawn from a population with positive skewness is reasonable when the sample size is greater than 25 times the square of the skewness coefficient of the population. We investigate whether a variant of this crude rule applies for a proportion estimated from a stratified simple random sample.

Key Words: Effective sample size; Skewness coefficient; Suppression rule; Third central moment.

1. Introduction

In his celebrated survey-statistics textbook William Cochran (1977) suggests that the standard (Wald) two-sided 95% confidence interval for an estimated population proportion p (having a property) based on a simple random sample works reasonably well when the absolute value of the estimate's (Fisher) skewness coefficient is less than 0.2. In particular, Cochran suggested that the true population proportion P will fall within the standard confidence interval at least 94% of the time over repeated samples *although the fraction of misses on either side of the interval need not be equal*.

Cochran's rule for simple random sampling, given that name (as far as we can tell) by Sugden, Smith and Jones (2000), is that the sample size n should exceed $25\mathrm{G}_1{}^2$, where G_1 is the skewness coefficient of the distribution from which the sample is drawn. We are using the definition of the coefficient of skewness Cochran used, which can be found in Evans, Hastings and Peacock (2000, page 15): the ratio of the third central moment of the distribution in the numerator and the second central moment of the distribution raised to the $3/2$ power in the denominator.

The skewness coefficient of the sample proportion p is then $\mathrm{G}(p) = \mathrm{G}_1/n^{1/2}$ ignoring finite-population correction, so the rule translates to $|\mathrm{G}(p)| < 0.2$. Cochran's crude rule (Cochran called his original suggestion "crude") applies to any sample mean with a positive skewness. Here mostly we limit discussion to an estimated proportion p having a positive skewness. Note that the estimated proportion $1-p$ is symmetric to p and has a negative skewness. Consequently, we conjecture that the standard two-sided confidence interval is reasonable when $|\mathrm{G}(p)| < 0.2$ for any nearly (design) unbiased estimated proportion p computed from a complex sample. We investigate this conjecture empirically for unbiased estimates based on virtual stratified simple random samples in Section 3. In Section 4, we offer a discussion of the practical repercussions of our conjecture realizing that operationally $\mathrm{G}(p)$ will need to be replaced by an estimate. We provide some statistical background from probability-sampling theory (often called "design-based sampling theory") in the next section.

1. Phillip S. Kott, RTI International, retired (US). E-mail: philkott1@gmail.com; Burton Levine, RTI International (US).

2. Some background

For a stratified simple random sample, let $h = 1, \ldots, H$ denote strata; $n_h > 1$ the sample size in stratum h; N_h the population size of stratum h; $n = \sum_{h=1}^{H} n_h$; and $N = \sum_{h=1}^{H} N_h$. Let P_h the population proportion in stratum h, and p_h the sample proportion in stratum h.

The following equations are all well-known. The population proportion P is equal to $P = \sum_{h=1}^{H} \frac{N_h}{N} P_h$, while $p = \sum_{h=1}^{H} \frac{N_h}{N} p_h$ is its estimator. Assuming, as we will from now on, that N is so large that finite-population correction can be ignored, the variance of p is

$$\mathrm{Var}(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{P_h(1-P_h)}{n_h}, \tag{2.1}$$

and an unbiased estimator for this variance is

$$\mathrm{var}(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{p_h(1-p_h)}{n_h - 1}. \tag{2.2}$$

The third-central moment of p is

$$M_3(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^3 \frac{P_h(1-P_h)(1-2P_h)}{n_h^2}, \tag{2.3}$$

and an unbiased estimator for this parameter when all the $n_h > 2$ is

$$m_3(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^3 \frac{p_h(1-p_h)(1-2p_h)}{(n_h - 1)(n_h - 2)}. \tag{2.4}$$

The skewness coefficient of p is

$$\mathrm{G}(p) = M_3(p) \Big/ \left[\mathrm{Var}(p) \right]^{3/2}, \tag{2.5}$$

and a nearly unbiased estimator of this parameter (if it exists) is

$$\mathrm{g}(p) = m_3(p) \Big/ \left[\mathrm{var}(p) \right]^{3/2}. \tag{2.6}$$

The following popular *ad-hoc* measure of skewness avoids measuring the third-central moment of p:

$$\mathrm{G}^*(p) = \frac{(1-2P) \Big/ (n^+)^{1/2}}{\left[P(1-P) \right]^{1/2}} = \frac{(1-2P)}{P(1-P)} \left[\mathrm{Var}(p) \right]^{1/2}, \tag{2.7}$$

where $n^+ = \frac{P(1-P)}{\mathrm{Var}(p)}$ is the *effective sample size* of the sampling design and estimator.

A nearly unbiased estimator for $\mathrm{G}^*(p)$ is

$$\mathrm{g}^*(p) = \frac{(1-2p) \Big/ (n^*)^{1/2}}{\left[p(1-p) \right]^{1/2}} = \frac{(1-2p)}{p(1-p)} \left[\mathrm{var}(p) \right]^{1/2}, \tag{2.8}$$

where $n^* = \frac{p(1-p)}{\text{var}(p)}$ is the *estimated* effective sample size of the sampling design and estimator. Unlike $\text{g}(p)$, $\text{g}^*(p)$ can be computed when one or more $n_h = 2,$ which is often the case in practice.

Under simple random sampling and large $n,$ $n = n^+ \approx n^*,$ so that $\text{G}^*(p) = \text{G}(p),$ and $\text{g}^*(p) \approx \text{g}(p).$

3. Some simulated experiments

Under simple random sampling, a sample size of approximately 180 is needed for the $\text{G}(p)$ to be less than 0.2 when the target P is 0.1. The simulated experiments in this section were designed with this in mind.

Our goal was to evaluate two-sided 95% confidence intervals for different estimates based on stratified samples of 180 units. We considered three-stratum sampling designs with 60 sampling units in each and 90-stratum sampling designs with two sampling units in each. We considered the possibility that either the relative population sizes in every stratum was the same or that half the population was represented by a third of the sampling units, the latter in two different ways (as will be explained). Finally, we considered a homoscedastic (equal unit variance) survey-variable-assignment method where each population unit had the same probability of having the binary survey value 1 (rather than 0) and that that probability varied across the 20 numerical values 0.01, 0.02, …, 0.19, 0.20. We also considered a heteroscedastic assignment method where a third of the population had no chance of having a binary survey value 1, a third had the same chance of having survey value 1 as in the homoscedastic assignment method, and a third had double the chance of having survey value 1 as in the homoscedastic method.

Rather than simulating 180 sampled units drawn separately for 20 estimates in each of 12 different scenarios (2 variable-assignment methods × 2 sets of strata formations × 3 sets of relative population sizes) 100,000 times, we did the equivalent to ease the computational burden. We drew 100,000 ur-samples (i.e., original or primitive "samples"). Each ur-sample contained 180 ordered ur-sampling units. Within each of the 12 scenarios, every ur-sampling unit was assigned to 20 separate survey values – and thus 20 virtual sampling units. We call what was found about the 20 estimated confidence intervals in each scenario the result of a "simulated experiment" because we didn't really draw samples in each scenario. Nevertheless, we refer to each of the 100,000 selections of an 180-unit ur-sample and its repercussions as a "simulation".

The details of what we did follow. Each ur-sampling unit j was associated with an independent random draw d_j from the uniform distribution on the half-closed, half-open interval [0, 1) (i.e., 0 is included in the interval, but 1 is not). Letting P_v take on the 20 values 0.01, 0.02, …, 0.19, 0.20, each ur-sampling unit j had 20 binary survey variables assigned to it in one of two ways. In the homoscedastic variable-assignment method $y_{jv} = 1$ when $d_j < P_v$ and 0 otherwise. In the heteroscedastic variable-assignment method, $y_{jv} = 1$ when $d_j < P_v a_j$ and 0 otherwise, where $a_j = 0$ when $j < 61,$ $a_j = 1$ when $60 < j < 121,$ and $a_j = 2$ when $j > 120.$

The ordered ur-sampling units j were assigned to strata in the following manner. In the three-strata assignment method, the ur-sampling unit was assigned to stratum 1 when $j <$ 61; the unit was assigned to stratum 2 when $60 < j <$ 121; and the unit was assigned to stratum 3 when $j >$ 120. In the 90-strata method

when $j = 1$ or 2 the unit was assigned to stratum 1; when $j = 3$ or 4, the unit was assigned to stratum 2, and so forth.

Finally, the strata have been assigned relative population sizes (i.e, $f_h = N_h/N$) in three different ways. For the three-strata assignment method, the three ways were $f_1 = f_2 = f_3 = 1/3$; $f_1 = 1/2$, $f_2 = f_3 = 1/4$; and $f_1 = f_2 = 1/4$, $f_3 = 1/2$. For the 90-strata assignment method, the three ways were $f_h = 1/90$ for all h; $f_h = 1/60$ for $h = 1$ to 30, $f_h = 1/120$ otherwise; and $f_h = 1/120$ for $h = 1$ to 60, $f_h = 1/60$ otherwise.

For many of the 12 scenarios, the estimators' target proportion P (defined in Section 2) was the same as P_v (the straight average of the survey values in the virtual sample). The exceptions occurred with the heteroscedastic variable-assignment method when the stratum shares (i.e., relative population sizes) were not all equal. When the virtual sampling units in either the lowest stratum $(h = 1)$ under the three-strata assignment method or in the 30 lowest strata $(h = 1, \ldots, 30)$ under the 90-strata assignment method accounted for half the estimates, $P = (3/4)\,P_v$. When the samples in either the highest stratum $(h = 3)$ or 30 highest strata $(h = 61, \ldots, 90)$ accounted for half the estimates, $P = (5/4)\,P_v$.

Note that P_v (not P) appears on the x-axis in the six graphs in Figures 3.1, 3.2, and 3.3. These graphs display the average coverages across 100,000 simulations of the traditional two-sided 95% confidence intervals under the scenarios described above for P_v set at 0.01, 0.02., …, and 0.20.

The confidence intervals were computed using the traditional model-free probability-sampling estimates of $\mathrm{Var}(p)$ ignoring finite-population correction described in Section 2. That the data used in the experiments were generated from a model does not undermine the usefulness of model-free methods of inference (especially when the model generating the data is unknown, which was not the case here).

We are interested in the relationship between the coverages of the traditional two-sided 95% confidence intervals and the values of $\mathrm{G}(p)$ which (like $\mathrm{G}^*(p)$), always decrease as P_v increases at least within the range we investigated, that is, for $P_v < 0.2$ (not shown).

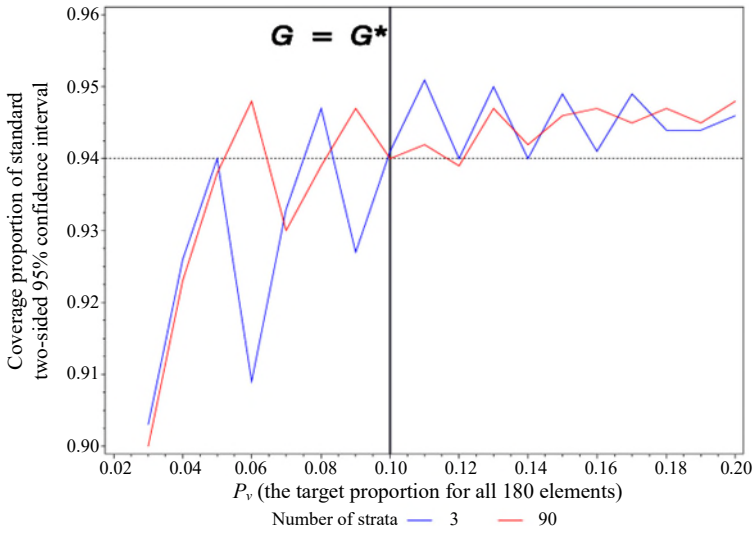
Vertical lines appear at the first value of P_v for which $\mathrm{G}(p)$ is less than 0.2. Such lines also appear for $\mathrm{G}^*(p)$ (the two lines are identical when the survey variables are homoscedastic and the stratum shares are equal). The vertical lines allow us to assess the strong version of our conjecture, namely that coverages to the right of the $\mathrm{G}(p)$ or $\mathrm{G}^*(p)$ line should *always* be at least 94%. This version fails for both $\mathrm{G}(p)$ and $\mathrm{G}^*(p)$, but a weaker crude version replacing "always" with "usually" does not.

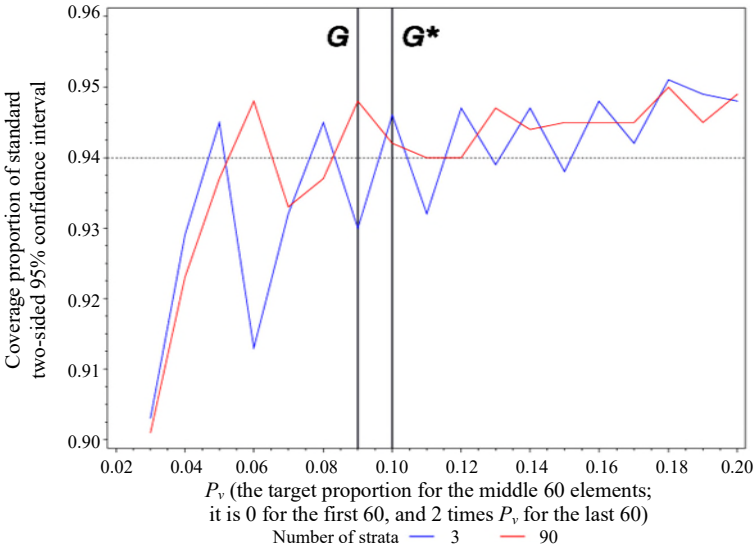
Although the virtual samples and the estimated p are the same when the variable assignment and relative stratum shares are the same, the coverages are not because the estimated variances $\mathrm{var}(p)$ (in equation (2.2)), unlike the actual variances $\mathrm{Var}(p)$ (in equation (2.1)), differ between the three-strata and 90-strata assignment methods. That is why there the red and blue lines differ in each of the six graphs displayed in the three figures, while the $\mathrm{G}(p)$ and $\mathrm{G}^*(p)$ lines (from equations (2.5) and (2.7), which are functions of the sample, $\mathrm{Var}(p)$, $\mathrm{M}_3(p)$ (from equation (2.3)), and P, are the same for both strata assignment method. This allowed us to show the results of 12 scenarios in six graphs.

For these six graphs, we used the standard 95% confidence intervals produced by SAS's PROC SURVEYMEANS (2020); that is, $\mathrm{CI}_d(P) = p \pm t_{0.05}(d)\sqrt{\mathrm{var}(p)}$, where $t_{0.95}(d) =$ 1.9754 for the three-stratum

assignment $(d = 180 - 30 = 150$ being the nominal degrees of freedom of $\text{var}(p))$ and $t_{0.95}(d) =$ 1.9872 for the 90-stratum assignment $(d = 180 - 90 = 90)$. As noted above, the p are the same for both strata-assignment methods, but the $\text{var}(p)$ are not. The jaggedness of coverages has been noted in pervious empirical work (e.g., Brown, Cai and Dasgupta (2001) and Dean and Pagano (2015)) and is attributed to the discrete nature of the determination of coverage (the interval either covers P or it does not). It is a bit surprising that the 90-strata (red) lines appear less jagged than the 3-strata (blue) ones even though the latter have more degrees of freedom. An investigation into why that is the case must wait for another time.

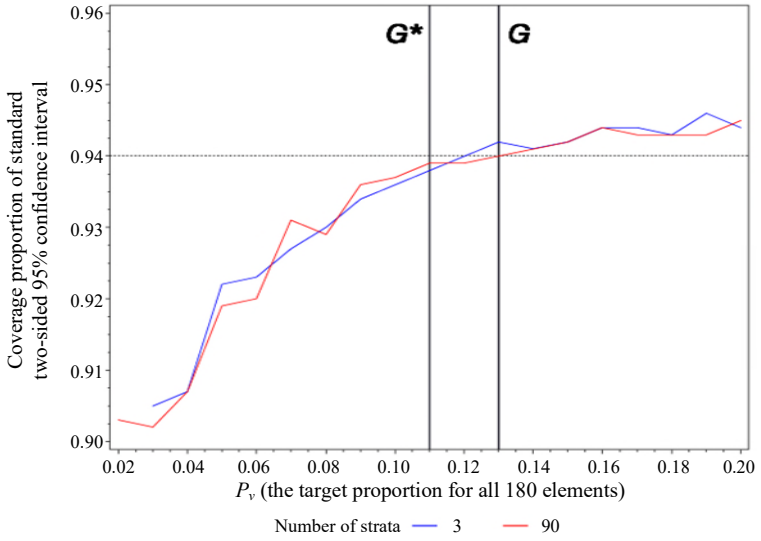
Figure 3.1 Coverages when all stratum shares (f_h) are equal.

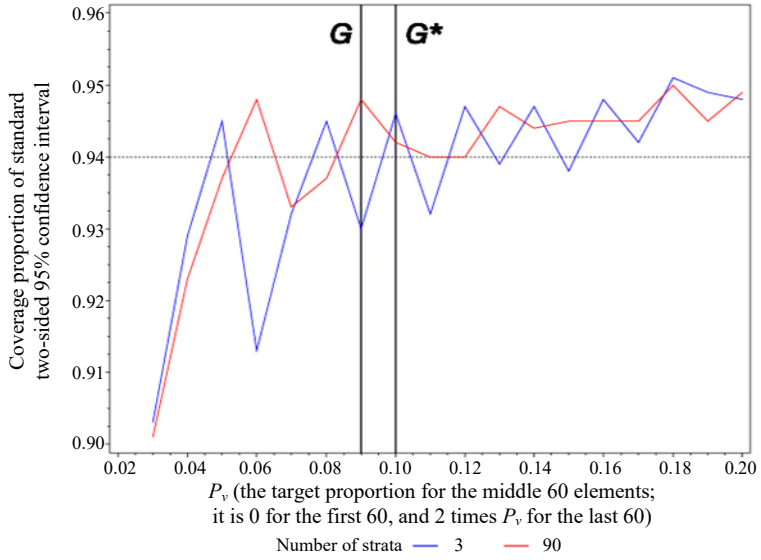




Note: G(G*) is first less than 0.2 at its vertical line, decreasing further as P_v increases.

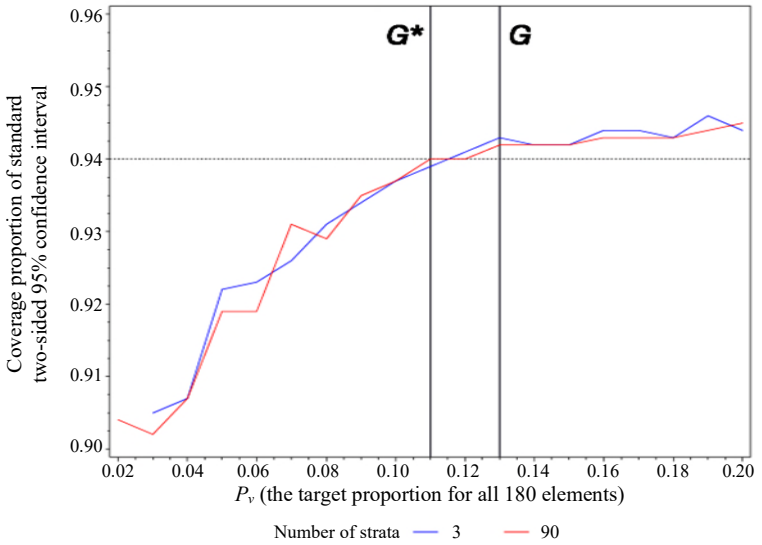
Figure 3.2 Coverages when first third of strata have half the shares.

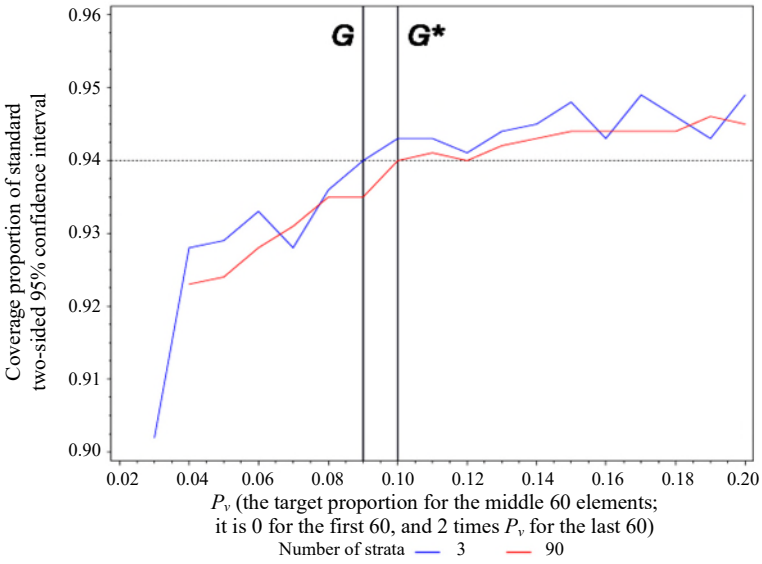




Note: G(G*) is first less than 0.2 at its vertical line, decreasing further as P_v increases.

Figure 3.3 Coverages when last third of the strata have half the share.





Note: G(G*) is first less than 0.2 at its vertical line, decreasing further as P_v increases.

A little thought reveals why the bottom graphs on Figures 3.1 and 3.2 coincide exactly, while the top graphs on Figures 3.2 and 3.3 are nearly identical (it is only the finite nature of the simulations that causes them to differ slightly).

4. A discussion

The use of a standard two-sided 95% confidence interval for an estimated proportion relies on the estimate being asymptotically normal. Among other things, the skewness coefficient of a normally distributed estimator is 0. An estimated proportion based on a finite sample has a non-zero skewness coefficient. The skewness coefficient tends to decrease as the size of the sample on which the estimates is based increases. We developed our version of Cochran's rule to be able to determine when the size of a complex sample is large enough for the reasonable use of a standard two-sided 95% confidence interval.

Our crude version of Cochran's rule, namely that the standard two-sided 95% confidence interval for an estimated proportion p based on a complex sample is reasonable when the absolute value of its skewness coefficient $\mathrm{G}(p)$ is less than 0.2, often cannot be used directly because $\mathrm{G}(p)$ is unknown. As it happens, its unbiased estimator $\mathrm{g}(p)$ tends to have a slight upward bias due to the random nature of its denominator (in equation (2.6)) when $\mathrm{G}(p)$ itself is positive. Consequently, it appears (from our limited simulation experiments) usually safe to replace the crude rule:

> the standard two-sided 95% confidence interval for an estimated proportion p (based on a complex sample) is reasonable when the absolute value of its skewness coefficient $\mathrm{G}(p)$ is less than 0.2,

with the more operational rule:

> The standard two-sided 95% confidence interval for an estimated proportion p is reasonable when the absolute value of its estimated skewness coefficient $\mathrm{g}(p)$ is less than 0.2,

or even

> The standard two-sided 95% confidence interval for an estimated proportion p is reasonable when the absolute value of the alternative estimated skewness coefficient $\mathrm{g}^*(p)$ is less than 0.2 when $\mathrm{g}(p)$ cannot be computed.

These operational versions of Cochran's rule for a proportion estimated from a complex sample are even more likely to be reasonable when the standard two-sided 95% confidence interval for p is expressed as $\mathrm{CI}_{60}(P) = p \pm 2\sqrt{\mathrm{var}(p)}$, and $\mathrm{var}(p)$ has at least 60 nominal degrees of freedom.

$\mathrm{CI}_{60}(P) = p \pm 2\sqrt{\mathrm{var}(p)}$ is the version of the standard two-sided 95% confidence interval for P that many sophisticated users internally calculate when provided only an estimated proportion p and its estimated standard error $\sqrt{\mathrm{var}(p)}$. That suggests the following suppression rule for an estimated proportion:

Suppress p when the absolute value of its estimated skewness coefficient $g(p)$ or its alternative estimated skewness coefficient $g^*(p)$ is greater than 0.2,

because it is at that point that the widely used confidence interval $CI_{60}(P) = p \pm 2\sqrt{var(p)}$ may no longer be reasonable.

Although we have not looked at clustered sampling *per se*, the 90-strata experiments can be viewed as representing a two (or more) stage sample design with perfect correlation within each of the 180 primary sampling units (PSUs); that is, every element in each PSU has the same survey value (0 or 1) as every other element in the PSU. Dean and Pagano (2015) and their supplementary material (available from the authors) show that for a particular version of a two-stage sample (30 PSUs selected using probability proportional to size sampling and seven elements drawn with equal probability from within each PSU) the standard two-sided 95% interval does a poorer job covering a small P as the correlation within the PSUs increase (their analysis stopped when the intracluster correlation reached a high of 0.5). Even the "exact" Clopper-Pearson interval covered poorly when the intracluster correlations were high and P was small.

References

Brown, L., Cai, T. and Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* 16, 101-133.

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

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.

Evans, M., Hastings, N. and Peacock, B. (2000). *Statistical Distributions* (3rd edition), New York: John Wiley & Sons, Inc.

SAS (2020). *SAS/STAT 15.2 User's Guide.* SAS Institute Inc., Cary, NC.

Sugden, R.A., Smith, T.M.F. and Jones, R.P. (2000). Cochran's rule for simple random sampling. *Journal of the Royal Statistical Society* (B), 62, 4, 787-793.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2024.

- C. Adrijo, *US Food and Drug Administration White Oak Campus*
- P. Andersson, Stockholm University
- N. Bates, *U.S. Census Bureau* (retired)
- B. Bell, *U.S. Census Bureau*
- E. Berg, *Iowa State University*
- C. Bocci, *Statistics Canada*
- H.-J. Boonstra, *Statistics Netherlands Heerlen*
- C. Boulet, *Statistics Canada*
- J. Breidt, *NORC at the University of Chicago*
- J.M. Brick, *Westat Inc.*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- L. Chen, *National Institute of Statistical Sciences*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- R. Clark, *Australian National University*
- M. Cohen, *American Institutes for Research*
- T. Crossley, *European University Institute*
- M. Dagdoug, *Laboratoire de Mathématiques de Besançon*
- S. Das, *Maastricht University*
- M. DeBell, *Stanford University*
- M. del Mar Rueda, *University of Granada*
- J. Drechsler, *Institute for Employment Research*
- P. Duchesne, *Université de Montréal*
- J.L. Eltinge, *U.S. Census Bureau*
- N. English, *National Opinion Research Center*
- A. Erciulescu, *Westat Inc.*
- W.A. Fuller, *Iowa State University*
- J. Gambino, *Statistics Canada*
- G. Goh, *Kyungpook National University*
- D. Haziza, *University of Ottawa*
- D. Hedlin, *Stockholm University*
- M.A. Hidiroglou, *Statistics Canada*
- J. Jiang, *University of California*
- D. Judkins, *ABT Associates Inc Bethesda*
- Y. Kawakubo, *Chiba University*
- B. Kim, *University of Maryland*
- J.K. Kim, *Iowa State University*
- P.S. Kott, *RTI International*
- K. Larbi, *L'Institut national de la statistique et des études économiques*
- P. Lahiri, *University of Maryland*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- T. Lewis, *George Mason University*
- A. Manda, *University of Georgia*
- A. Matei, *Université de Neuchâtel*
- S. Matthews, *Government of Canada*
- K. McConville, *Reed College*
- M. McRoy, *NORC at the University of Chicago*
- E. Médous, *Nantes Université*
- T. Merly-Alpa, *Institut national d'études démographiques*
- I. Molina, *Universidad Complutense de Madrid*
- J. Moore, *University of Essex*
- R. Münnich, *University of Trier*
- J. Opsomer, *Westat Inc.*
- P. Parker, *University of California*
- J. Pascale, *U.S. Census Bureau*
- D. Pfeffermann, *University of Southampton*
- J.N.K. Rao, *Carleton University*
- P. Righi, *Italian National Institute of Statistics*
- L.-P. Rivest, *Université Laval Faculté des sciences et de génie*
- A. Ruiz-Gazen, *Toulouse School of Economics*
- F.J. Scheuren, *National Opinion Research Center*
- A. Sen, *University of Maryland at College Park*
- N. Shlomo, *The University of Manchester*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- E. Slud, *University of Maryland at College Park*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- S. Sugasawa, *Keio University*
- H. Sun, *West Tennessee Healthcare*
- X. Tang, *The University of Arizona*
- M. Templ, *University of Applied Sciences and Arts Northwestern Switzerland*
- Y. Tillé, *University of Neuchatel Faculty of Sciences*
- R. Tiller, *U.S. Bureau of Labor Statistics*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- J. van den Brakel, *Statistics Netherlands*
- A. Veiga, *Brazilian Institute of Geography and Statistics*
- F. Verret, *Statistics Canada*
- G. Walejko, *U.S. Census Bureau*
- L. Wang, *University of Virginia*
- M. Williams, *RTI International*
- B. West, *University of Michigan*
- C. Wu, *University of Waterloo*
- D. Yang, *US Bureau of Labor Statistics*
- Y. You, *Statistics Canada*
- W. Yung, *Statistics Canada*
- L.-C. Zhang, *University of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2024 issues: Céline Ethier of Economic Statistics Methods Division; Patrick O'Leary of Social Statistics Methods Division; Julie Bélanger and Catherine Pelletier of Official Release, Publishing and Creative Services Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Isabelle Gravelle, Kathy Charbonneau, Ashley Perry, Travis Robinson and Karo-Lynn Audy as well as our partners in the Communications Division.

ANNOUNCEMENTS

Nominations Sought for the 2026 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 2026 Waksberg Invited Address at the Statistics Canada Symposium, expected to be held in the autumn of 2026. The paper will be published in an upcoming issue of *Survey Methodology* (targeted for December 2026).

The author of the 2026 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, 2025 to the chair of the committee, Jae Kwang Kim (jkim@iastate.edu). Nominations should include a CV and a nomination letter.** Nominations will remain active for 5 years.

Members of the Waksberg Paper Selection Committee (2024-2025)

Jae-Kwang Kim, *Iowa State University* (Chair)
Paul Smith, *University of Southampton*
Alina Matei, *Université de Neuchatel*
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)
Maria Giovanna Ranalli (2022-2023)
Denise Silva (2023-2024)

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 40, No. 2, June 2024

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 40, No. 3, September 2024

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 52, No. 1, March/mars 2024

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 52, No. 2, June/juin 2024

Research Article

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.