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

Survey Methodology

Survey Methodology
49-1

Release date: June 30, 2023

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

June 2023

Volume 49

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A journal published by Statistics Canada

Volume 49, Number 1, June 2023

Contents

Special paper in memory of Professor Chris Skinner – Winner of the 2019 Waksberg Award

Regular papers

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 1-3
Statistics Canada, Catalogue No. 12-001-X

Tribute to Chris Skinner, a colleague and friend

Danny Pfeffermann[1]

Abstract

This brief tribute reviews Chris Skinner's main scientific contributions.

Key Words: Analysis of complex surveys; Statistical disclosure control; Official statistics.

Chris passed away about three years ago, only a few months after Fred Smith passed away. In November of last year, Tim Holt, passed away. So, within 3 years, the three legendary survey sampling statisticians from Southampton, who edited the famous Wiley 1989 book *Analysis of Complex Surveys Surveys* (Skinner, Holt and Smith, 1989), passed away. The book summarized 10 years of research at the University of Southampton and in the rest of the world devoted to this topic, paving the way for new research and applications, which continue to evolve in all kind of directions. A second Wiley book on the same topic, edited by Chambers and Skinner, *Analysis of Survey Data*, which surveyed the rapid developments in the field during the 90's, had been published in (Chambers and Skinner, 2003). Since the early 1980's, Southampton became a leading international centre in social statistics and survey sampling, led by Chris, Fred and Tim, attracting top class researchers and students in this field from all over the world.

Chris' work covers many topics related to survey sampling theory and inference, making him one of the top social and survey statisticians in the world. In what follows, I mention briefly a few of them. Starting with his PhD thesis, supervised by Tim Holt, Chris was one of the first statisticians to note that the complex sampling designs, which are in common use to collect multivariate social data, are rarely non-informative as far as statistical modelling is concerned, and that there is a need for suitable adjustments to standard inference methods to correct for this, thus avoiding possible bias and wrong inference. He continued his work in this area throughout his academic career.

Another major research area of Chris was in Statistical Disclosure Control (SDC), focusing on estimating the probabilities of re-identification of survey micro data and using them for computing inclusive disclosure risk measures. For this, Chris developed statistical models, which accounted for the type of data under risk (the key variables), the sampling method used for the sample selection and the method that might be used by the intruder to achieve disclosure. Later, Chris and Natalie Shlomo, showed that probability sampling methods as well as non-perturbative SDC methods, do not satisfy the requirement of differential privacy, a hot topic in SDC, nowadays researched jointly by statisticians and computer scientists, following among others Chris' stimulus. I refer the readers to the paper by Natalie Shlomo on Chris' very significant contributions to SDC, published in this issue of *Survey Methodology*.

1. Danny Pfeffermann, Department of Statistics, Hebrew University of Jerusalem, Israel and Southampton Statistical Sciences Research Institute (S3RI), UK. E-mail: msdanny@mail.huji.ac.il and msdanny@soton.ac.uk.

In 2013, Chris chaired an independent review of plans for the 2021 Census in the UK. The resulting parliamentary report recommended that census data should be collected online rather than following more traditional ways of census data collection. This recommendation had been implemented very successfully, with an incredible high response rate.

Throughout his academic career, Chris was heavily involved in the work of statistical agencies in the UK and internationally. He established strong research relationships with the Central Statistical Office and the Office of Population Censuses and Surveys in the UK, and later with the Office for National Statistics (ONS), when the two offices merged. Since then, the University of Southampton is the primary source of methodological advice for the ONS. During that research relationship, Chris led many high-profile projects, including variance estimation for the Labour Force Survey and the sample allocation for the Retail Prices Index. Variance estimation was one of Chris' favourite research topics. He was also instrumental in setting up the MSc program in Official Statistics at Southampton, which is training official statisticians from the UK and other countries. During the years 2000-2011, Chris was a member of Statistics Canada Statistical Advisory Committee. In 2012, Chris moved to the London School of Economics (LSE), the university where he studied for his MSc degree in 1976, before moving to Southampton. His involvement with official statistics continued after his move to the LSE.

So far for a brief review of Chris' professional achievements and seminal contributions to survey sampling inference, social statistics and SDC. Chris and I came to Southampton in 1978, I as a postdoc student and Chris as a lecturer, starting in parallel his PhD under the supervision of Tim Holt. Being two young lecturers sharing similar interest in survey sampling inference, we soon became friends, which also included our respective families. Our friendship lasted until his tragic death. However, it was only in 1998 that we published two joint articles, the first on estimation of gross flows, which was applied experimentally at the Central Bureau of Statistics in Israel, and the second on weighting in multi-level modelling. The later article has been read at a meeting of the royal Statistical Society and received a lot of attention in the literature since then. Chris authored more than 80 peer-reviewed journal papers and co-edited the two influential books on the analysis of complex survey data, which I mentioned before.

In 2019, Chris was awarded the Waksberg award, one of many awards that he received during his academic career. As part of the award ceremony, he was supposed to present his Waksberg award paper during the annual symposium of Statistics Canada, on a topic of his choice. Chris initially refused to accept the award, stating that he is not sure that he will be able to travel to Canada because of his health condition. How noble of him, showing what a wonderful person he was, leave aside his outstanding professional achievements. As the Chair of the 2019 Waksberg Award Selection Committee, I used all my convincing powers to change his mind and somehow I succeeded, and Chris started working on his presentation on New Developments in Statistical Disclosure Control.

How sad that Chris was unable to finish what he had started to prepare, and we are grateful to Natalie Shlomo for agreeing to complete and present it at the 2021 symposium. J.N.K. Rao and Jae-Kwang Kim

provided testimonies at the end of Natalie's presentation, and they have kindly agreed to put them in writing in this issue, following Natalie's paper.

Chris was a highly respected statistician and a joy as a colleague. His immense scientific contributions will continue to be applied and form the basis for new research in the future.

References

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*, New York: John Wiley & Sons, Inc.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, Chichester, UK. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 5-25
Statistics Canada, Catalogue No. 12-001-X

Statistical disclosure control and developments in formal privacy: In memoriam to Chris Skinner

Natalie Shlomo[1]

Abstract

I provide an overview of the evolution of Statistical Disclosure Control (SDC) research over the last decades and how it has evolved to handle the data revolution with more formal definitions of privacy. I emphasize the many contributions by Chris Skinner in the research areas of SDC. I review his seminal research, starting in the 1990's with his work on the release of UK Census sample microdata. This led to a wide-range of research on measuring the risk of re-identification in survey microdata through probabilistic models. I also focus on other aspects of Chris' research in SDC. Chris was the recipient of the 2019 Waksberg Award and sadly never got a chance to present hisWaksberg Lecture at the Statistics Canada International Methodology Symposium. This paper follows the outline that Chris had prepared in preparation for that lecture.

Key Words: Risk of re-identification; Data revolution; Privacy models; Differential privacy.

1. Introduction

A special memorial session was held in honour of Chris Skinner at the 2021 Statistics Canada International Methodology Symposium on October 22nd, 2021, with many moving contributions from friends and colleagues to celebrate Chris' life and achievements. Chris was the 2019 Waksberg Award recipient and was planning on attending the 2019 International Methodology Symposium to deliver his lecture. Unfortunately his illness took a turn for the worse and he sadly passed away on February 21st, 2020. In the memorial session and here in this paper, I describe the evolution of Statistical Disclosure Control (SDC) research with an emphasis on Chris' contributions to the field. The outline for the talk and for the paper was based on a set of notes that Chris had drawn up in preparation for his 2019 Waksberg Lecture provided to me by his son, Tom Skinner.

I had the great privilege of working with Chris as a PhD student developing the theory around estimating the risk of re-identification presented in Section 4.1 and later as his colleague at the University of Southampton where we continued to make progress on other topics of SDC. To read an excellent overview of Chris' personal side and his contribution to social statistics and survey methodology, see the interview that was published in the International Statistical Review (Haziza and Smith, 2019).

I discuss early SDC developments in Section 2, and move to the challenges that we are facing today due to the Data Revolution in Section 3. Section 4 describes Chris' contributions and his seminal research in SDC. Section 5 presents ongoing and future research in SDC and data privacy, and one of Chris' final contributions of embedding the Computer Science definition of Differential Privacy into the SDC tool-kit at government agencies. I close in Section 6 with some final words on the impact of Chris' research in government statistics, social statistics and survey methodology.

1. Natalie Shlomo, Social Statistics Department, University of Manchester, United Kingdom. E-mail: natalie.shlomo@manchester.ac.uk.

2. Early SDC developments and history

Public awareness around confidentiality and privacy arose in the 1960's leading to the start of public opposition to data collection, particularly for censuses within Europe following WWII. For example, there were many objections against the collection of information about the population living in the Netherlands and their last traditional census was held in 1971. This opposition led to a need by government agencies to respond to public concerns about privacy and confidentiality (Dunn, 1967) and discussed in other early work in Barabba (1975), Cox (1976), Fellegi (1972) and Dalenius (1974). Fellegi (1972, page 8) wrote: "National Statistical Institutes (NSIs) live by the good will and trust of the public so that to maintain this trust is literally a question of life or death to them". Based on the research carried out in Sweden, Dalenius (1977) was one of the first to formally define and formalize a framework for Statistical Disclosure Control (SDC) as follows: "An unauthorized party should not be able to learn something about an individual through the release of a statistic calculated from the database $D, f(D)$, that cannot be learned without access to $f(D)$".

The work by Dalenius and others provided the framework for researching and developing SDC within government agencies and the establishment of formal governance boards on the release of statistical data. The research that was carried out in the United States included, for example, the Subcommittee on Disclosure-Avoidance Techniques established in 1976 by the Federal Committee on Statistical Methodology, and sponsored by the Statistical Policy Division of Office of Management and Budget (OMB) as reported in Jabine, Michael and Mugge (1977) (see also the 1978 report and Appendix A on Statistical Disclosure Avoidance Practices in Selected Federal Agencies). In this Appendix there are five sections with recommendations: the concept of SDC; what to release; disclosure avoidance techniques; effects of disclosure on data subjects and users; and needs for research and development. There are also general rules that were put in place, for example no regional areas could be published with less than 100,000 individuals.

Further work into the 1980's placed an emphasis on SDC for outputs derived from survey data as it was originally and erroneously thought that sampling provided protection against disclosure risks (Dalenius, 1988). Paass (1988) was one of the first to estimate the fraction of identifiable records in survey microdata and took into account the sampling, the SDC method of additive noise and prior knowledge under an assumed "attack" on the data. In his paper, Paass (1988) wrote: "where there is large knowledge, the requirement for privacy protection and high-quality data perhaps may be fulfilled only if the linkage of such files with extensive additional knowledge is prevented by appropriate organizational and legal restrictions". In addition, Bethlehem, Keller and Pannekoek (1990) was one of the first papers to use probabilistic modelling to estimate the risk of re-identification in survey microdata by estimating the number of population uniques given sample uniques on a set of cross-classified quasi-identifiers. More on this methodology and the contributions of Chris in this area will be presented in Section 4.1.

Into the 1990's there was more demand for detailed outputs particularly with the availability of better technological solutions and personal computers. There were also rising concerns by users of the data on

having to work with protected, modified and perturbed outputs. This coincided with large-scale SDC developments through a scientific evolution of the methodology and the international interchange of theoretical and practical developments, for example, the International Symposium on Statistical Disclosure Avoidance held in the Netherlands in 1990 (as reported in a special issue of Statistica Neerlandica, 1992). There were also cross-collaborations within the European Union through the 4th Framework research project Statistical Disclosure Control (SDC) (1996-1998) and many other EU projects following on from this initial project, including the development of SDC software: mu-ARGUS for microdata (Hundepool, van de Wetering, Ramaswamy, Franconi, Capobianchi, de Wolf, Domingo, Torra, Brand and Giessing, 2003) and tau-ARGUS for tabular data (specifically cell suppression for magnitude tables containing business statistics) (Hundepool, van de Wetering, Ramaswamy, de Wolf, Giessing, Fischetti, Salazar, Castro and Lowthian, 2011). See https://research.cbs.nl/casc/index.htm for more details of the research projects across Europe and the book by Hundepool, Domingo-Ferrer, Franconi, Giessing, Schulte-Nordholt, Spicer, de Wolf (2012).

A special issue of the *Journal of Official Statistics*, (Vol. 14(4), 1998) titled "Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data" was particularly impactful during that time and highlighted the large-scale research undertaken in SDC in both academia and government agencies. In addition, a book and training course were developed (see: Willenborg and De Waal (1996) with contributions by Chris Skinner, and later a second edition in Willenborg and De Waal (2001)). Continuing work was happening simultaneously in the US and Canada, for example, the Federal Committee on Statistical Methodology (1994); the Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997); Disclosure Control Issues at Statistics Canada (Yeo and Robertson, 1995) including the software package CONFID that also carried out cell suppressions for magnitude tables.

Throughout the 1990's, there was growing focus on the development of access and governance arrangements and legislation, and the notion of tiered data access to provide statistical data to researchers. Data Archives and Research Data Centres were set up along with data governance approaches and frameworks for making effective use of statistical data, for example, the "5 Safes" Framework shown in Table 2.1 and put in practice at the Office for National Statistics (ONS) in the UK in 2002 (Ritchie, 2009) and later the Anonymization Decision-Making Framework (Elliot, Mackey, O'Hara and Tudor, 2016 and available at https://ukanon.net/framework/).

Table 2.1
The 5 safes framework

Safe Projects	Is this use of the data appropriate?
Safe People	Can the users be trusted to use the data in an appropriate manner?
Safe Settings	Does the access facility limit unauthorised use?
Safe Data	Is there a disclosure risk in the data itself?
Safe Outputs	Are the statistical results non-disclosive?

3. The data revolution

From around the year 2008, there has been an abundance of accessible data in the public domain, including open data and big data, leading to greater risks of breaches of privacy and confidentiality since these data sources can potentially be used to compromise released statistical data. In addition, more advanced technological tools were available that enabled better data linkages and data manipulation to increase the likelihood of re-identification in statistical data. Government agencies started to become aware that standard SDC methods may not be sufficient in protecting the confidentiality of statistical units and therefore initiated tighter restrictions and more controlled access to the data as a solution to SDC. This also manifested in changes to the legislation, particularly the 2016 European Union (EU) General Data Protection Regulation (GDPR) which provided provisions and requirements related to the processing of personal data of individuals. There was also more focus on privacy concerns in health data (El Emam, Jonker, Arbuckle and Malin, 2011) and genetic data (Homer, Szelinger, Redman, Duggan, Tembe, Muehling, Pearson, Stephan, Nelson and Craig 2008; Gymrek, McGuire, Golan, Halperin and Erlich, 2013) where the latter were shown to be of high-risk and had implications on the dissemination and sharing of DNA databases. In the commercial domain, there were many examples of breaches of privacy which were widely publicized: AOL search keywords (Barbaro and Zeller, 2006), New York City (NYC) taxi trips (Douriez, Doraiswamy, Freire and Silva, 2016), Cambridge Analytica and Facebook (Meredith, 2018), and others.

With greater technological advancements and the possibility to link data sources, this led to the development of trusted third parties to carry out data linkages and secure multi-party computing. Secure multi-party computing was originally developed in the Computer Science literature and made a cross-over to the statistical literature on how to run advanced statistical modelling under this approach (Slavkovic, Nardi and Tibbits, 2007; Snoke, Brick, Slavkovic and Hunter, 2018). In addition, collaborations between computer scientists and the statistical community grew and led to important developments on database privacy within government agencies (see Section 5 for more details). In the privacy literature, Dwork, Smith, Steinke and Ullman (2017) wrote: “beginning in the mid-2000s, the field of privacy-preserving statistical analysis of data has witnessed an influx of ideas developed some two decades earlier in the cryptography community”.

4. Contributions of Chris Skinner to SDC research

Chris’s formal research in Statistical Disclosure Control (SDC) began with his collaborations at the University of Manchester to argue for the release of sample microdata (the SARs) from the 1991 UK Census (Marsh, Skinner, Arber, Penhale, Openshaw, Hobcraft, Lievesley and Walford, 1991; Skinner, Marsh, Openshaw and Wymer, 1994; Marsh, Dale and Skinner, 1994). This led to his interest on measuring the risk of re-identification in survey microdata through probabilistic modelling first published in Skinner (1992) and described in Section 4.1. He also started his long career of advising for government

statistics and data access committees, for example: UK Census Design and Methodology Advisory Committee Statistical Disclosure Control (SDC) Subgroup (2008-2010); Understanding Society Data Access Committee (2010-2013); Expert Advisory Group on Data Access (EAGDA), Wellcome Trust, Medical Research Council (MRC), Economic and Social Research Council (ESRC) and Cancer Research UK (2012-2014).

4.1 Measuring the risk of re-identification in survey microdata and extensions

Since the 1990's, many government agencies have been releasing microdata from large-scale government surveys where the sample is drawn randomly from a finite general population and the sample fractions are small. Examples are samples from the Labour Force Survey and Family Expenditure Survey. Typically, the researcher would have to go through an application process to gain access to the data, either received on a floppy disc or via a Data Archive. Nevertheless it was soon recognized that sampling alone could not provide enough protection in the microdata and a wealth of research was generated on the subject of measuring the risk of re-identification in sample microdata and disclosure avoidance techniques.

The disclosure risk scenario for the release of sample microdata drawn from a general population is based on the following assumptions: (1) there is an "intruder" (someone with malicious intent to discredit the statistical office) who has access to the microdata and other auxiliary information about the population that allows him/her to link data sources in order to identify individuals in the sample microdata; (2) there is no "response knowledge" meaning that the intruder does not know who was drawn into the sample of the survey. The basic definition of the risk of re-identification is therefore the probability of correctly being able to make this match. Chris was among the first to develop a statistical modelling framework to estimate the probability of re-identification, conditional on the released data and assumptions about how the data is generated (knowledge of the sampling process). The model is with respect to key variables defined as a set of quasi-identifiers in both data sources and typically categorical such as age, sex, location, ethnic group. Cross-classifying the key variables leads to large contingency tables of sample counts, where many of the cells of the table have a value of zero or a value of one, and we particularly focus on the disclosure risk from the cells of size one, i.e., the sample uniques. The risk of re-identification is based on the notion of population uniqueness in the contingency table: given an observed sample unique in a cell of a table generated from cross-classifying the key variables, what is the probability that the cell is also a population unique? Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file which are useful to make informed decisions about the release of the data and the level of access via data governance boards.

The probabilistic modelling framework developed by Chris takes a simplified approach that restricts the information that would be known to intruders (Skinner and Holmes, 1998; Elamir and Skinner, 2006).

We denote F_k the population size in cell k of a table spanned by key variables having K cells, f_k the sample size in cell k, $\sum_k F_k = N$ and $\sum_k f_k = n$. The set of sample uniques, is defined: $\text{SU} = \{k : f_k = 1\}$ and these are the high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

1. Number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$.
2. Expected number of correct links if we were to match the sample uniques to the population (assuming a random assignment of the population within cell k). For example, if a sample unique matches to three individuals in the population, the match probability for that sample unique would be 1/3. Aggregating all match probabilities over the sample uniques leads us to: $\tau_2 = \sum_k I(f_k = 1) 1/F_k$.

If the population frequencies F_k are known then the global disclosure risk measures are straightforward to calculate. However, it is generally assumed that the population frequencies F_k are unknown and we need to use probabilistic modelling to estimate the disclosure risk measures as follows:

$$\hat{\tau}_1 = \sum_k I(f_k = 1) \hat{P}(F_k = 1 \mid f_k = 1) \quad \text{and} \quad \hat{\tau}_2 = \sum_k I(f_k = 1) \hat{E}(1/F_k \mid f_k = 1). \tag{4.1}$$

Given that we are modelling population counts based on a contingency table of sample counts spanned by the key variables, Chris assumed a Poisson distribution and a log-linear model to estimate disclosure risk measures in (4.1). In this model, he and his co-authors assume that F_k are realizations of independent Poisson random variables: $F_k \sim \text{Pois}(\lambda_k)$ for each cell k. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell $k : f_k \mid F_k \sim \text{Bin}(F_k, \pi_k)$. It follows that:

$$f_k \sim \text{Pois}(\pi_k \lambda_k) \quad \text{and} \quad F_k \mid f_k \sim \text{Pois}(\lambda_k (1 - \pi_k)) \tag{4.2}$$

and population cell counts F_k given the sample cell counts f_k are also realizations of independent Poisson random variables.

As typical in this type of framework, the parameters λ_k are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = \mathbf{x}_k' \boldsymbol{\beta}$ where $\mathbf{x}_k$ is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\hat{\boldsymbol{\beta}}$ are obtained by solving the score equations:

$$\sum_k (f_k - \pi_k \exp(\mathbf{x}_k' \boldsymbol{\beta})) \mathbf{x}_k = 0. \tag{4.3}$$

The fitted values are then calculated by: $\hat{\mu}_k = \exp(\mathbf{x}_k' \hat{\boldsymbol{\beta}})$ and $\hat{\lambda}_k = \hat{\mu}_k / \pi_k$. Individual disclosure risk measures for cell k under the assumption of the Poisson distribution are:

$$P(F_k = 1 \mid f_k = 1) = \exp(\lambda_k (1 - \pi_k))$$

and

$$E\left(1/F_k \mid f_k = 1\right) = \left(1 - \exp\left(\lambda_k\left(1-\pi_k\right)\right)\right) \Big/ \left(\lambda_k\left(1-\pi_k\right)\right). \quad (4.4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4.4) leads to the record-level disclosure risk measure estimates $\hat{P}\left(F_k = 1 \mid f_k = 1\right)$ and $\hat{E}\left(1/F_k \mid f_k = 1\right)$ and these are aggregated to obtain global disclosure risk measure estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ for (4.1).

As an example, assume that the statistical agency would like to release microdata from the Labor Force Survey. The agency assumes the following quasi-identifiers (key variables) that they would like to release in the microdata and on which to assess the risk of re-identification (number of categories in parentheses): sex (2), age group (10), marital status (3), ethnic groups (10), employment status (3) and occupation groups (10). When cross-classified, these key variables lead to cells $k, k = 1, \ldots, K$ where $K = 18{,}000$ cells. The agency identifies those cells of the cross-classified key variables where there is a single sample unit: $f_k = 1$ (a sample unique). For each cell that is a sample unique, we estimate the record-level disclosure risk measure according to the probability that the sample unique is also a population unique $P\left(F_k = 1 \mid f_k = 1\right)$. We also estimate the match probability for the sample unique based on (hypothetically) matching the sample unique to a population using the key variables as matching variables to obtain $E\left[1/F_k \mid f_k = 1\right]$. These record level risk measures are then aggregated to obtain estimates for the global disclosure risk measures τ_1 and τ_2 assuming that for a large $K, \left(F_k, f_k\right)$ are independent.

Skinner and Shlomo (2008) develop a method for selecting the main effects and interaction terms for the log-linear model that finds the right balance in accounting for random and structural zeros in the contingency table. The method is based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h\left(\lambda_k\right) = P\left(F_k = 1 \mid f_k = 1\right)$ for τ_1 and $h\left(\lambda_k\right) = E\left(1/F_k \mid f_k = 1\right)$ for τ_2, they consider the expression:

$$B = \sum_k E\left(I\left(f_k = 1\right)\right)\left(h\left(\hat{\lambda}_k\right) - h\left(\lambda_k\right)\right).$$

A Taylor expansion of h leads to the approximation

$$B \approx \sum_k \pi_k \lambda_k \exp\left(-\lambda_k\right)\left(h'\left(\lambda_k\right)\left(\hat{\lambda}_k - \lambda_k\right) + h''\left(\lambda_k\right)\left(\hat{\lambda}_k - \lambda_k\right)^2 \Big/ 2\right)$$

and the relations $E\left(f_k\right) = \pi_k \lambda_k$ and $E\left(\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right) = \pi_k^2 E\left(\hat{\lambda}_k - \lambda_k\right)^2$ under the hypothesis of a Poisson distribution fit lead to a further approximation of B of the form:

$$\hat{B} \approx \sum_k \hat{\lambda}_k \exp\left(-\pi_k \hat{\lambda}_k\right)\left(-h'\left(\lambda_k\right)\left(f_k - \pi_k \hat{\lambda}_k\right) + h''\left(\lambda_k\right)\left(\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right) \Big/ \left(2\pi_k\right)\right). \quad (4.5)$$

For example, for τ_1:

$$\hat{B}_1 \approx \sum_k \hat{\lambda}_k \exp\left(-\hat{\lambda}_k\right)\left(1-\pi_k\right)\left\{\left(f_k - \pi_k \hat{\lambda}_k\right) + \left(1-\pi_k\right)\left[\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right] \Big/ \left(2\pi_k\right)\right\}. \quad (4.6)$$

As can be seen, the goodness-of-fit criteria $\hat{B}$ are related to the notion of measuring over and under-dispersion as was developed in the Econometrics literature (Cameron and Trivedi, 1998). The method

selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i \Big/ \sqrt{\hat{v}_i}$ for $\hat{\tau}_i, i = 1, 2,$ where $\hat{v}_i$ are variance estimates of $\hat{B}_i$. The goodness-of-fit criteria $\hat{B}_i \Big/ \sqrt{\hat{v}_i}$ have an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}_i$ is zero.

Skinner and Shlomo (2008) also address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e., $P\left(f_k = 1 \,\middle|\, F_k\right) = F_k \pi_k \left(1 - \pi_k\right)^{F_k - 1}$, this may not be the case if sampling clusters (households). In practice, key variables typically include variables such as age, sex and occupation that tend to cut across clusters. Therefore the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities in the log-linear model. Under complex sampling, the λ_k can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas, 2003), where the estimating equation in (4.3) is modified as:

$$\sum_k \left(\hat{F}_k - \exp\left(\mathbf{x}_k' \boldsymbol{\beta}\right)\right) \mathbf{x}_k = 0 \tag{4.7}$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell k: $\hat{F}_k = \sum_{i \in k} w_i$. The resulting estimates λ_k are plugged into expressions in (4.4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k \Big/ \hat{F}_k$. The goodness-of-fit criteria $\hat{B}$ is also adapted to the pseudo-maximum likelihood method. See Skinner and Shlomo (2008) for a simulation and real application demonstrating this approach for both a survey with an equal probability design and a survey with a complex design.

The probabilistic modelling presented here and in other related work in the literature assume that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also purposely be misclassified as a means of masking the data, for example through record swapping or the post randomization method (PRAM) (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). Shlomo and Skinner (2010) adapt the estimation of the risk of re-identification based on τ_2 to take into account measurement errors. Denoting the cross-classified key variables in the population and the microdata as X and assuming that X in the microdata have undergone some misclassification or perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix M:

$$M_{kj} = P\left(\tilde{X} = k \,\middle|\, X = j\right). \tag{4.8}$$

The record-level disclosure risk measure of a match with a sample unique under measurement error is:

$$\frac{M_{kk}\left(1 - \pi_k M_{kk}\right)}{\sum_j F_j M_{kj} \Big/ \left(1 - \pi_k M_{kj}\right)} \leq \frac{1}{F_k}. \tag{4.9}$$

Under assumptions of small sampling fractions and small misclassification errors, the disclosure risk measure can be approximated by: $M_{kk} \Big/ \sum_j F_j M_{kj}$ or $M_{kk} \Big/ \tilde{F}_k$ where $\tilde{F}_k$ is the population count with $\tilde{X} = k$. Aggregating the per-record disclosure risk measures, the global risk measure τ_2 is now:

$$\tau_2 = \sum_k I\left(f_k = 1\right) M_{kk} \Big/ \tilde{F}_k . \tag{4.10}$$

Note that to calculate the measure only the diagonal of the misclassification matrix needs to be known, i.e., the probabilities of not being perturbed. Population counts are generally not known so the estimate in (4.10) can be obtained by probabilistic modelling on the misclassified sample as shown above:

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk} \hat{E}\left(1 \Big/ \tilde{F}_k \,\Big|\, \tilde{f}_k\right). \tag{4.11}$$

In more recent work with Chris and presented for the first time in Shlomo and Skinner (2022), a new direction is explored to measure the risk of re-identification for non-probability data sources. More specifically, there are registers in the public domain, where the membership of the register is not known and is sensitive. Examples of registers are of persons with a medical condition, such as Cancer or HIV, or registers that include membership to a loyalty card scheme. The approach can also be extended to the case where samples are drawn from the registers and more generally to non-probability samples, such as those arising from web-surveys. Extending the framework above, the microdata from a random sample can still be used to estimate population parameters according to the probabilistic modelling framework for estimating the risk of re-identification as above, however the complication is to also estimate the propensity of membership for the individuals in the register.

More specifically, let U and U_1 denote the population and the register population, respectively, with $U_1 \subset U$. Let R_i be the register indicator variable for individual i with $R_i = 1$ if $i \in U_1$ and $R_i = 0$ otherwise. As mentioned, we suppose that membership R_i is a sensitive variable for which disclosure is undesirable.

We denote the register population frequencies in cell k by F_k^1. The most risky records are for cells with $F_k^1 = 1$ and, analogous to the derivation presented in Skinner and Shlomo (2008), a risk measure is given by

$$\tau_1^* = \sum_k P\left(F_k = 1 \,\middle|\, F_k^1 = 1\right) I\left(F_k^1 = 1\right). \tag{4.12}$$

There is no way that these measures can be estimated consistently from the register microdata alone. The microdata provide information about the F_k^1 but not about the F_k in U and distribution of X in U_1 may be quite different to that in U so the microdata carries no direct information about the F_k. Therefore, we use the random sample microdata file in which the values of X are recorded for a probability sample s from U. Let f_k denote the frequency in cell k in s. Note that the f_k and F_k^1 are observed, but the F_k are not. If the intruder has access to the sample microdata file, then it may be advantageous to restrict attention to cells with $f_k = 1,$ leading to the following risk measure

$$\tau_1 = \sum_k P\left(F_k = 1 \,\middle|\, F_k^1 = 1, f_k = 1\right) I\left(F_k^1 = 1, f_k = 1\right). \tag{4.13}$$

Following Skinner and Shlomo (2008), suppose that F_k is Poisson distributed, $F_k \sim \text{Pois}(\lambda_k)$ where the parameter λ_k obeys the log-linear model:

$$\log(\lambda_k) = \mathbf{x}_k' \boldsymbol{\beta}. \tag{4.14}$$

Suppose that within cell k the unknown membership variable R_i takes the value 1 with probability p_k, independently for each of the F_k units, so that $F_k^1 \sim \text{Pois}(\phi_k)$ where $\phi_k = \lambda_k p_k$, and the F_k^1 are binomially distributed $F_k^1 \mid F_k \sim \text{Bin}(F_k, p_k)$ conditional on the F_k. Further, we assume that p_k follows the logistic model:

$$\text{logit}(p_k) = \mathbf{x}_k' \boldsymbol{\xi}. \tag{4.15}$$

As shown in Shlomo and Skinner (forthcoming), the risk measure τ_1 is estimated by:

$$P(F_k = 1 \mid F_k^1 = 1, f_k = 1) = \frac{\exp(-(1-\pi_k)(\lambda_k - \phi_k))}{1 + (1-\pi_k)(\lambda_k - \phi_k)}$$

and to evaluate τ_1^*, we use

$$P(F_k = 1 \mid F_k^1 = 1) = P(F_k - F_k^1 = 0) = \exp(-(\lambda_k - \phi_k)) \tag{4.16}$$

since $F_k - F_k^1 \sim \text{Pois}(\lambda_k - \phi_k)$.

Therefore, the estimation of these measures requires both the estimation of $\boldsymbol{\beta}$ from $f_k \sim \text{Pois}(\pi_k \lambda_k)$, and in a second step, the estimate $\boldsymbol{\xi}$, fixing λ_k at the value implied by (4.14). We then use (4.16) and the fact that $\phi_k = \lambda_k p_k$ to write

$$\log \phi_k = \log \lambda_k + \mathbf{x}_k' \boldsymbol{\xi} - \log(1 + \exp(\mathbf{x}_k' \boldsymbol{\xi})) \tag{4.17}$$

and estimate $\boldsymbol{\xi}$ from the fact that $F_k^1 \sim \text{Pois}(\phi_k)$ using maximum likelihood estimation and treating λ_k as known. Alternative approaches of estimation are proposed in Shlomo and Skinner (2022), however, more work is needed to improve the simultaneous estimation of the model parameters $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$.

Another type of design-based estimator for measuring disclosure risk in sample microdata is called the DIS measure and was developed in Skinner and Elliot (2002) and extended in Skinner and Carter (2003) for more complex survey designs. The disclosure risk is based on a different disclosure risk scenario where an intruder draws a unit at random from the population, checks if the unit is in the sample, and if so, estimates the probability that there will be a correct match to the unit in the sample (this is known as a "fishing scenario"). Notice that this scenario is quite different than the scenario mentioned under the probabilistic modelling where the intruder has access to a unit in the released microdata and attempts to match the unit to the population. The advantage of this "fishing scenario" is that the measure can be estimated easily without the need for probabilistic modelling. The DIS measure is defined as

$$\theta = \sum_k I\left(f_k = 1\right) \Big/ \sum_k I\left(f_k = 1\right) F_k \tag{4.18}$$

and estimated by:

$$\hat{\theta} = \pi n_1 \Big/ \left[\pi n_1 + 2\left(1-\pi\right) n_2\right] \tag{4.19}$$

where n_1 are the sample uniques: $\text{SU} = \left\{k : f_k = 1\right\}$ and n_2 are the sample duplicates: $\text{SD} = \left\{k : f_k = 2\right\}$. Skinner and Shlomo (2012) extend this approach to estimate frequencies of frequencies in finite populations beyond sample uniques.

4.2 Separating disclosure risk and harm

Chris provided a conceptual framework in Skinner (2012) for separating potential disclosure risk from harm, thus linking earlier papers by Duncan and Lambert (1986) and Lambert (1993). The framework is based on decision theory where the actors are the agency, the intruder and the user and they are analysed with respect to their actions and loss functions. Chris emphasized the importance of separating out what can be measured by statistical theory (potential disclosure risk) and what aspects of decision-making requires other inputs, such as policy judgements (potential disclosure harm). This work was also motivated by the Disclosure Risk-Data Utility framework in Duncan, Keller-McNulty and Stokes (2001) and the Economics of Privacy in Abowd and Schmutte (2019).

As can be seen from these examples, Chris expanded the depth and breadth of SDC research. Other areas of research where Chris had considerable impact was on the associations between measuring disclosure risks in SDC with other related areas of research, such as record linkage (Skinner, 2009) and forensic science (Skinner, 2007). Chris' more recent work on disclosure risk and privacy will be the topic of the next section.

5. Disclosure risk and privacy

In the Computer Science privacy literature, there are formal definitions of privacy via privacy models that aim to protect against a class of attacks. The privacy models are parameterized by a threshold of disclosure risk determined *a priori*. Once the privacy model is defined, a perturbation technique is developed to guaranty protection against the attack subject to the prescribed threshold. Some examples of privacy models are k-anonymity, t-closeness, l-diversity and Differential Privacy. As mentioned, the privacy literature uses perturbative techniques resulting in a greater loss of information compared to some of the more standard techniques developed in the SDC literature. The class of attacks in the privacy literature are typically based on dealing with inferential disclosure which encompasses both identity and attribute disclosure risks, although k-anonymity (Sweeney, 2002) aims to avoid linkage attacks similar to the SDC approach described in Section 4. We note that even with increasing focus on protecting against attribute and inferential disclosure, government agencies are always obliged to protect against identity disclosure through possible linkage attacks because of legislation and codes-of-practice on protecting

statistical entities from re-identification. The SDC literature is based on matching quasi-identifying variables whilst in the privacy literature there is no distinction between identifying and sensitive variables.

It is important to point out, however, that many concepts in the privacy literature are not new to SDC. For example, the reconstruction attacks, mentioned in Garfinkel, Abowd and Martindale (2018) for motivating the use of Differential Privacy to protect 2021 US Census outputs, have the same considerations as those that motivated the development of complementary cell suppression. This approach was developed in the 1980's for protecting magnitude tables of business statistics. When selecting complementary cell suppressions, the lower and upper bounds of the suppressed cells are calculated based on information from the margins of the table and assumptions of non-negativity. Indeed, the reconstruction attack is not about linkage rather it is concerned with attribute disclosure through small cell counts, particularly on the margins. As mentioned the privacy literature mainly focuses on attribute and inferential disclosures although the SDC literature have also covered these topics, for example the predictive disclosure risk mentioned in Fuller (1993). Another privacy model in the Computer Science literature is tracing attacks where one can infer whether an individual is in a sensitive dataset, e.g., Homer et al. (2008), but the SDC literature has also focused on whether a data subject is visible in a dataset.

Since 2005, there have been four collaborative meetings between the SDC community and the Computer Science privacy community and this has led to substantial understanding of the different approaches both with respect to guarantying privacy and maintaining sufficient utility in the data. For example, in Nissim, Steinke, Wood, Altman, Bembenek, Bun, Gaboardi, O'Brien and Vadhan (2018, page 5), it is mentioned: "Privacy is a property of an informational relationship between input and output not a property of output alone", and this has led to some relaxations of the strict privacy guarantees in the Computer Science literature. An example of one relaxation can be found below in Section 5.1 formula (5.2). On the other hand, the SDC community has recognized the need to have more formal privacy guarantees, particularly with increasing demands for government agencies to allow accessing statistical data via web-based dissemination applications (for example, flexible table builders, remote access, remote analysis). Dissemination via open applications means that the agencies are relinquishing some of the strict control of what data can be released. The collaborations between the SDC community and the Computer Science privacy community have led to a journal that was initiated in 2005, titled the Journal of Privacy and Confidentiality (https://journalprivacyconfidentiality.org) of which Chris served as one of the first co-editors (Abowd, Nissim and Skinner, 2009).

In the next section we focus on the privacy model of Differential Privacy since Chris had a direct involvement in developing this approach as a possible method to be included in the SDC tool-kit at statistical agencies.

5.1 Differential privacy

One privacy model that has gained considerable traction in the SDC community is Differential Privacy (Dwork, 2006). Dwork and Naor (2010) show that the Dalenius (1977) definition of a privacy breach

introduced in Section 2 is impossible to prevent and proposed that instead of comparing information with and without the statistic $f(D)$, they compare $f(D)$ and $f(D')$ where D' is the database D without a single unit.

In Differential Privacy, a "worst case" scenario is allowed for, in which the potential intruder has complete information about all units in the database except for one unit of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases $D, D' \in A$ where A is the domain of databases and D, D' differ by one individual, and for all possible outcomes defined as subsets $S \in \text{Range}(M)$ we have:

$$p(M(D) \in S) \leq e^{\varepsilon} p(M(D') \in S). \tag{5.1}$$

A relaxation is offered by the definition of (ε, δ)-differential privacy:

$$p(M(D) \in S) \leq e^{\varepsilon} p(M(D') \in S) + \delta. \tag{5.2}$$

This means that having observed a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the intruder is unable to determine whether the output was generated from database D or D'. In other words, the ratio $p(M(D) \in S) \,/\, p(M(D') \in S)$ is bounded and the probability in the denominator cannot be zero. Thus, Differential Privacy formally bounds increased disclosure risk for an individual due to their data being in database D and that they would not have faced had their data not been part of D (Dwork and Roth, 2014). Under the (ε, δ)-differential we allow a small amount of slippage to this constraint.

The solution to guarantee Differential Privacy in the Computer Science literature is to add noise/perturbation to the outputs of the queries under specific parameterizations, for example by generating additive noise from the Laplace Distribution (or a discretized Laplace Distribution for adding noise to count data).

Shlomo and Skinner (2012) first looked at whether standard SDC methods are differentially private mechanisms according to the definition in (5.1). They found that sampling, as well as other non-perturbative SDC methods such as coarsening variables, are not differentially private. In this setting, there are two possible definitions of the database: the population database $\mathbf{x}_U = (x_1, \ldots, x_N)$ and the sample database $\mathbf{x}_s = (x_1, \ldots, x_n)$ where N is the size of the population U and n is the size of the sample s. Assume that a vector of counts of size K is released from the sample: $\mathbf{f} = (f_1, \ldots, f_K)$ where $f_k = \sum_{i \in s} I(x_i = k)$. Let $P(\mathbf{f} \mid \mathbf{x}_U)$ denote the probability of $\mathbf{f}$ with respect to the sampling and $\mathbf{x}_U$ is treated as fixed. According to this set-up, ε-differential privacy holds if:

$$\max \left| \ln \left(\frac{P(\mathbf{f} \mid \mathbf{x}_U)}{P(\mathbf{f} \mid \mathbf{x}'_U)} \right) \right| \leq \varepsilon \tag{5.3}$$

for some $\varepsilon > 0$, where the maximum is over all pairs $(\mathbf{x}_U, \mathbf{x}'_U)$ which differ in only one element and across all possible values of $\mathbf{f}$. Now in random sampling strategies there is generally a positive probability that a sample unique for a cell k can be a population unique: $f_k = F_k = 1$. For any given $\mathbf{f}$ and any

sampling scheme where f_k can equal F_k with positive probability, there exists a databases $\mathbf{x}_U$ where $f_k = F_k \geq 1$ for some k and $P\left(\mathbf{f} \mid \mathbf{x}_U\right) \neq \mathbf{0}$. If we change an element of $\mathbf{x}_U$ which takes the value k to construct $\mathbf{x}'_U$ then we obtain $F'_k = F_k - 1 < f_k$ and $P\left(\mathbf{f} \mid \mathbf{x}'_U\right) = \mathbf{0}$.

On the other hand, perturbative methods in the SDC tool-kit can be made differentially private if the perturbation mechanisms do not have zero probabilities of perturbation. As an example, SDC methods traditionally do not perturb zero cells in census tables containing whole population counts, but rather stochastically introduce more zeros through the perturbation using an approach such as random rounding. However, to make this perturbation approach differentially private, the (random) zeros of the table also need to be perturbed.

5.2 Online flexible table builders

There has been much interest by government agencies to develop online flexible table builders through bespoke web-based platforms. Users generate and download their own census tables from a set of predefined variables and categories selected through drop-down lists. Light disclosure checks are carried out on each generated table to determine whether the table can be released or not, and if so, disclosure control methods are applied to the table before release. One such application was developed at the Australian Bureau of Statistics (ABS) (see: https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder). The application uses a perturbation vector to change values of cell counts depending on the original cell value, where the perturbation mechanism has the properties of being bounded, unbiased, has maximal entropy, only allows for non-negative perturbations and zero cells are not perturbed. Shlomo and Young (2008) introduced an approach to transform the perturbation vectors in such a way that the marginal counts are preserved in expectation by introducing the property of invariance into the perturbation mechanism.

In the ABS online flexible table builder, a small random number is assigned to each individual in the census microdata. Then, when a table is requested and the individuals are aggregated into the cells of the table, the random numbers of the individuals in each cell are also aggregated. This aggregated random number is then used as the seed to determine the perturbation (Fraser and Wooton, 2005). This means that any time a same cell appears in any requested census table, it will always have the same perturbation. Therefore, there is no risk of being able to "unpick" a true cell value by averaging out independent perturbations under multiple requests of the same table. In addition, this approach ensures that the perturbation mechanism is a "non-interactive" mechanism since essentially all outcomes of perturbation on requested census tables within the online flexible table builder are pre-determined in advance.

One of Chris' last initiatives prior to his illness was to take the lead on organizing a collaborative programme between statisticians, computer scientists, social scientists and practitioners held at the Isaac Newton Institute, University of Cambridge. Together with Professor David Hand, they successfully launched the Data Linkage and Anonymization Programme (supported by the UK Engineering and

Physical Sciences Research Council (EPSRC) grant no. EP/K032208/1) held from July to December 2016. It was during this programme that a group of statisticians looked at whether Differential Privacy could be a viable solution for an online flexible table builder for generating census tables, resulting in the paper by Rinott, O'Keefe, Shlomo and Skinner (2018). The main difference with the original ABS approach was to use a differentially private perturbation mechanism (known as the Exponential Mechanism which is essentially a discretized Laplace Distribution) and to perturb the (random) zero cells. Any resulting negative perturbations are then pushed to zeros in the census tables. Assuming independent pertubations, the Exponential Mechanism is defined as follows: for a given cell count value a, choose $b \in B$ (where B is the range of b) with probability proportional to $\exp\left((\varepsilon/2)u \,/\, \Delta u\right)$ where u is the perturbation and Δu is the maximum difference of a cell count in database D versus D', which for the case of census tables of internal cells, takes the value of one. Accounting for marginals in the census tables raises the complexity of Δu and the perturbation vector (see Rinott et al., 2018 for more information about marginals). In order to ensure utility, the perturbations are capped at ± 7 thus the mechanism satisfies (ε, δ)-differential privacy and we allow for a slippage of an unbounded ratio provided that δ is very small.

Here, we also implement the methodology of "same cell-same perturbation" of Fraser and Wooton (2005) which essentially makes this a non-interactive differentially private mechanism, and hence there is one privacy budget that is needed to protect the queries for tables within the online flexible table builder. Any request for the same table will always result in the same perturbed table and there is no further privacy budget spent beyond the initial perturbation.

An example of a pertubation vector for $\varepsilon = 1.5$ and $\delta = 0.00002$ and a perturbation cap at ± 7 is in Table 5.1 where the probability of perturbation is based on a discretized Laplace Distribution: $p(u) = \frac{1}{C}\exp\left(-\varepsilon|u|\right)$ and C is a normalizing constant so that the perturbation vector sums to 1. As mentioned, any resulting negative values are pushed to zero and this does not violate the property of Differential Privacy. Examples and applications are shown in Rinott et al. (2018). The authors also show how to adjust statistical inferences when carried out on perturbed census tables since the perturbation mechanism under Differential Privacy is known and not secret. This is in sharp contrast to the SDC approach where parameters of the SDC methods are generally held secret and not released to researchers. For example, when adding noise to continuous variables, the variance of the noise distribution would not be released.

Table 5.1
Perturbation vector for differentially private mechanism $\varepsilon = 1.5$, $\delta = 0.00002$ and a cap at ± 7

u	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
p(u)	0.00002	0.00008	0.00035	0.00157	0.00706	0.03162	0.14172	0.63516	0.14172	0.03162	0.00706	0.00157	0.00035	0.00008	0.00002

Differential Privacy has more formal by-design privacy guarantees and protects against attribute and inferential disclosure risks. Therefore, Differential Privacy may provide a better solution for protecting

statistical data that is disseminated as open data or via web-based open applications where there is less control and intervention by data custodians at statistical agencies. For this reason, Differential Privacy should be included as an additional method in the SDC tool-kit. The US Census Bureau will be applying Differential Privacy in their 2021 census products (Abowd, 2018). Further research is needed on how privacy budgets are influenced when combined with other SDC approaches, such as coarsening, sampling and variable suppression. There is also ongoing research within the privacy literature to improve the utility of differentially private perturbation mechanisms, for example the bounded Differential Privacy in Kifer and Machanavajjhala (2014).

6. Final words

In summary, a key feature of Chris's approach to research on SDC, as well as other areas of research, was that it was based on finding practical solutions to real statistical problems. His research was influential because he was able to put theory to practice and solve real problems, thus advancing scientific knowledge in the social sciences, government and social statistics and survey methodology. Chris had considerable influence on other research areas besides SDC. He has edited influential books and authored many journal articles in diverse research areas in survey statistics, including missing data, measurement error, data integration, the analysis of complex survey designs, multiple frame estimation and more. Chris was a definitive voice of a generation in the research and development of social statistics and he will be missed.

References

Abowd, J.M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867. https://doi.org/10.1145/3219819.3226070.

Abowd, J.M., and Schmutte, I.M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1), 171-202.

Abowd, J.M., Nissim, K. and Skinner, C.J. (2009). First issue editorial. *Journal of Privacy and Confidentiality*, 1(1), 1-6.

Barabba, V.P. (1975). The right to privacy and the need to know. In *US Bureau of the Census: A Numerator and Denominator for Measuring Change*. Technical Paper, 37, 23-29.

Barbaro, M., and Zeller, T. Jr. (2006). A face is exposed for AOL searcher, no. 4417749. *The New York Times*.

Bethlehem, J., Keller, W. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409), 38-45.

Cameron, A.C., and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Cox, L.H. (1976). *Statistical Disclosure in Publication Hierarchies*. Report No. 14 of the research project Confidentiality in Surveys, Dept. of Statistics, University of Stockholm, available at https://hdl. handle.net/1813/111306.

Dalenius, T. (1974). The invasion of privacy problems and statistics production-an overview. *Statistisk Tidskrift*, 3, 213-225.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15, 429-444.

Dalenius, T. (1988). *Controlling Invasion of Privacy in Surveys*. Dept. of Development and Research, Statistical Research Unit, Statistics Sweden.

Douriez, M., Doraiswamy, H. Freire, J. and Silva, C.T. (2016). Anonymizing NYC Taxi Data: Does it matter? *IEEE International Conference on Data Science and Advanced Analytics (DSAA2016)*, 140-148.

Duncan, G., and Lambert, D. (1986). Disclosure-limited data dissemination (with discussion). *Journal of the American Statistical Association*, 81(393), 10-28.

Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory.

Dunn, E.S. (1967). The idea of a national data centre and the issue of personal privacy. *American Statistician*, 21, 21-27.

Dwork, C. (2006). Differential privacy. In *ICALP*, (Eds., M. Bugliesi, B. Preneel, V. Sassone and I. Wegener). Heidelberg: Springer, lecture notes in computer science 4052, 1-12.

Dwork, C., and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 93-107.

Dwork, C., and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.

Dwork, C., Smith, A., Steinke, T. and Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and its Application. Annual Review of Statistics and its Application*, 4, 61-84.

El Emam, K., Jonker, E., Arbuckle, L. and Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS ONE*, 6(12), 1-12.

Elamir, E., and Skinner, C.J. (2006). Record-level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, 22, 525-539.

Elliot, M., Mackey, E., O'Hara, K. and Tudor, C. (2016). *The Anonymisation Decision Making Framework, Manchester: UKAN*. Available at https://ukanon.net/framework/.

Fellegi, I.P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337), 7-18.

Fraser, B., and Wooton, J. (2005). *A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing*. Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 November.

Fuller, W.A. (1993). Masking procedures for micro-data disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

Garfinkel, S., Abowd, J.M. and Martindale, C. (2018). Understanding database reconstruction attacks on public data. *ACM QUEUE*, 16(5), 1-26.

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

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339(6117), 321-324.

Haziza, D., and Smith, P.A. (2019). An interview with Chris Skinner. *International Statistical Review*, 87(3), 451-470.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V, Stephan, D.A., Nelson, S.F. and Craig, D.W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4(8), 1-9.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K. and de Wolf, P.P. (2012). *Statistical Disclosure Control*. New York: John Wiley & Sons, Inc.

Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M. Salazar, J.J., Castro, J. and Lowthian, P. (2011). tau-ARGUS User's Manual (Version 3.5). BPA no: 769-02-TMO, Statistics Netherlands Project: Essnet-project, Statistics Netherlands, The Netherlands. Available at https://research.cbs.nl/casc/Software/TauManualV3.5_rev.pdf.

Hundepool, A., van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra, V., Brand, R. and Giessing, S. (2003). mu-ARGUS user's manual (Version 3.2). BPA no: 768-02-TMO, Statistics Netherlands Project: CASC-project, Statistics Netherlands, The Netherlands. Available at https://research.cbs.nl/casc/deliv/manual3.2.pdf.

Jabine, T.B., Michael, J.A. and Mugge, R.H. (1977). *Federal Agency Practices for Avoiding Stastical Disclosure: Findings and Recommendations*. Available at http://www.asasrms.org/Proceedings/y1977/Federal%20Agency%20Practices%20For%20Avolding%20Statistical%20Disclosure%20-%20Findings%20And%20Reconwnendatlons.pdf.

Kifer, D., and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1), 1-36.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9, 313-331.

Marsh, C., Dale, A. and Skinner, C.J. (1994). Safe data versus safe setting: Access to microdata from the British Census. *International Statistical Review*, 62, 35-53.

Marsh, C., Skinner, C.J., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991). A case for samples of anonymised records from the 1991 Census. *Journal of the Royal Statistical Society A*, 154, 305-340.

Meredith, S. (2018). *Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal*. CNBC.

Nissim, K., Steinke, T. Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., O'Brien, D.R. and Vadhan, S. (2018). *Differential Privacy: A Primer for a Non-technical Audience*. University of Harvard, available at https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6(4), 487-500.

Rao, J.N.K., and Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. In *Analysis of Survey Data* (Eds., R.L. Chambers and C.J. Skinner), Wiley, Chichester, UK, 85-108.

Rinott, Y., O'Keefe, C., Shlomo, N. and Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Sciences*, 33(3), 358-385.

Ritchie, F. (2009). Designing a national model for data access. *Proceedings of the Comparative Analysis of Enterprise Data Conference*, Tokyo, Japan. Available at https://gcoe.ier.hit-u.ac.jp/CAED/papers/id213_Ritchie.pdf.

Shlomo, N., and Skinner, C.J. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, 4(3), 1291-1310.

Shlomo, N., and Skinner, C.J. (2012). Privacy protection from sampling and perturbation in survey microdata. *Journal of Privacy and Confidentiality*, 4(1), 155-169.

Shlomo, N., and Skinner, C.J. (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185, 4, 1644-1662. Available at http://dx.doi.org/10.1111/rssa.12902.

Shlomo, N., and Young, C. (2008). Invariant post-tabular protection of census frequency counts. In *PSD' 2008 Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer and Y. Saygin), Springer LNCS 5261, 77-89.

Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.

Skinner, C.J. (2007). The probability of identification: Applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society, Series A*, 170, 195-212.

Skinner, C.J. (2009). Record linkage, correct match probabilities and disclosure risk assessment. In *Insights on Data Integration Methodologies: ESSnet-ISAD workshop*, Vienna, 29-30 May 2008, Eurostat Methodologies and Working papers, Luxembourg, European Communities, 11-23.

Skinner, C.J. (2012). Statistical disclosure risk: Separating potential and harm. *International Statistical Review*, 80, 349-368, with discussion and rejoinder, 379-391.

Skinner, C.J., and Carter, R.G. (2003). Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling. *Survey Methodology*, 29, 2, 177-180. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6784-eng.pdf.

Skinner, C.J., and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B*, 64, 855-867.

Skinner, C.J., and Holmes, D. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

Skinner, C.J., and Shlomo, N. (2008). Assessing identification risk in survey microdata using Log Linear models. *Journal of the American Statistical Association*, 103(483), 989-1001.

Skinner, C.J., and Shlomo, N. (2012). Estimating frequencies of frequencies in finite populations. *Statistics and Probability Letters*, 82, 2206-2212.

Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10, 31-51.

Slavkovic, A.B., Nardi, Y. and Tibbits, M.M. (2007). Secure logistic regression of horizontally and vertically partitioned distributed databases. Seventh IEEE International Conference on Data Mining Workshops, (ICDMW2007), 723-728.

Snoke, J., Brick, T. Slavkovic, A. and Hunter, M.D. (2018). Providing accurate models across private partitioned data: Secure maximum likelihood estimation. *Annals of Applied Statistics*, 12(2), 877-914.

Statistica Neerlandica (1992). Volume 46(1), available at https://onlinelibrary.wiley.com/toc/14679574/1992/46/1 [last accessed 14/03/2023].

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.

Willenborg, L., and De Waal, T. (1996). *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics, New York: Springer Verlag, 111.

Willenborg, L., and De Waal, T. (2001). *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, New York: Springer-Verlag, 155.

Yeo, D., and Robertson, D. (1995). *Disclosure Control Issues at Statistics Canada*. https://publications.gc.ca/collections/collection_2017/statcan/11-613/CS11-617-96-5-eng.pdf.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 27-29
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical disclosure control and developments in formal privacy: In memoriam to Chris Skinner"

J.N.K. Rao[1]

Abstract

My comments consist of three components: (1) A brief account of my professional association with Chris Skinner. (2) Observations on Skinner's contributions to statistical disclosure control, (3) Some comments on making inferences from masked survey data.

Key Words: Differential privacy; Making statistical inference from masked data; Probability of identification.

My professional association with Skinner. In 1978, my good friend Professor Fred Smith of the University of Southampton invited me to spend four months of my sabbatical leave (April through July) at his university to collaborate on a research project dealing with the analysis of complex survey data. Alastair Scott, Gad Nathan, and Tim Holt were the other members of the team. Our team's initial work stimulated a lot of research on the analysis of complex survey data, including methods for categorical data taking account of design features and regression analysis under informative sampling. Southampton group since then evolved into a leading survey research center.

Our team's joint research on the analysis of complex survey data resulted in a conference held in Southampton in 1985, and, based on the presented papers, in a Wiley book edited by Skinner, Holt and Smith (1989). Skinner finished his PhD at the University of Southampton in 1982 and became a faculty member there. I met Skinner for the first time at the 1985 conference and I was most impressed by his deep understanding of the issues underlying the analysis of complex survey data. He took a major share in editing the book, particularly Part A of the book. This important book is widely cited. A second conference on the same topic was held in Southampton in 1999 to mark Fred Smith's retirement. A second Wiley book, edited by Chambers and Skinner (2003), was based on the presentations at the conference.

Skinner visited me several times for joint research and collaboration. I have also visited him. Our collaboration led to three joint papers dealing with different topics of importance: (1) Estimation in dual frame surveys with complex designs: Skinner and Rao (1996) proposed estimators that use the same sampling weights for all variables of interest, based on the design induced by the two separate designs. (2) Jackknife variance estimation under hot-deck imputation for multivariate statistics: Skinner and Rao (2002) derived bias-adjusted estimators under common donor imputation and associated jackknife variance estimators. (3). Quasi-score test with survey data: Rao, Scott and Skinner (1998) developed analogues of customary score tests for use with survey data where the use of multi-stage sampling and variable selection probabilities cause special problems.

1. J.N.K. Rao, Carleton University. E-mail: jrao34@rogers.com.

In 2017, Skinner presented an overview paper on analysis of categorical survey data at a conference held in Kunming, China, to celebrate my 80th birthday. The paper is now published (Skinner, 2019). In the same year, he gave an invited talk at the World Statistics Congress in Marrakesh, Morocco, in a session I organized. His talk studied alternative weighting options in regression analysis of survey data. That was the last occasion I interacted with Skinner before he passed away in 2021. Skinner and I served on Statistics Canada Advisory Committee on Statistical Methods for several years.

Skinner's seminal contributions to SDC. Shlomo provides an informative account of early statistical disclosure control (SDC) developments in Section 2. She highlights major contributions of Skinner to SDC in Section 4. In Section 4.1, Shlomo gives a detailed account of Skinner's seminal work on measuring risk of reidentification in survey micro data through probabilistic models. In this connection, I found Skinner's (2007) paper, establishing correspondence between SDC and forensic statistics regarding their common use of the concept of "probability of identification", very informative and interesting. Skinner showed that one cannot ignore the search method that an intruder employs to achieve disclosure, in the sense that the probability of disclosure varies with the search method employed. He also proposed methods to handle the impact of search methods.

Differential privacy (DP) is a hot topic currently. Shlomo gives a succinct account of DP in Section 5.1. Her joint paper with Skinner (Shlomo and Skinner, 2012) studied whether SDC methods are differentially private. They showed that probability sampling and other non-perturbative methods are not differentially private. In Section 5.2, Shlomo notes that Skinner, prior to his illness, initiated a collaborative program on SDC between statisticians, computer scientists, social scientists, and practitioners. This is indeed commendable, and his untimely death is a set back to this important collaborative program.

Making inferences from masked data. Shlomo did not cover issues related to making inferences from masked data. Analysts of survey data would like to use standard methods and software on the masked data. On the other hand, it appears that specialized DP-based methods and some other masking techniques used to preserve confidentiality of micro data require knowledge of the masking techniques used to generate the masked data and specialized software tailored to those methods, as noted by Raghunathan, Reiter and Rubin (2003).

Raghunathan et al. (2003) proposed imputation methods to create multiple sets of fully synthetic data. Standard methods and software, like those used in the context of multiple imputation for missing data, are then used to make inferences. Raghunathan (2016), Section 8.5, illustrated a mass imputation method for doing a regression analysis of a dependent variable y on an independent variable x by creating multiple data sets of synthetic variables (y^*, x^*). The proposed methods typically require more synthetic data sets than those used in the context of multiple imputation for missing data. Also, the methods might retain the problems of making inferences using methods based on multiple imputation for missing data, when the data are based on a complex survey design (Fay, 1996).

References

Chambers, R.L., and Skinner, C.J. (editors) (2003). *Analysis of Survey Data.* Chichester: John Wiley & Sons, Inc.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Raghunathan, T.E. (2016). *Missing Data Analysis in Practice.* Boca Raton: CRC Press.

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

Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.

Shlomo, N., and Skinner, C.J. (2012). Privacy protection from sampling and perturbation in survey microdata. *Journal of Privacy and Confidentiality*, 4(1), 155-169.

Skinner, C.J. (2007). The probability of identification: Applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society, Series A*, 170, 195-212.

Skinner, C.J. (2019). Analysis of categorical data from complex surveys. *International Statistical Review*, 87, S64-S78.

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

Skinner, C.J., and Rao, J.N.K. (2002). Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. *Journal of Statistical Planning and Inference*, 102, 149-167.

Skinner, C.J., Holt, D. and Smith, T.M.F. (editors) (1989). *Analysis of Complex Surveys*. Chichester: John Wiley & Sons, Inc.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 31-38
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical disclosure control and developments in formal privacy: In memoriam to Chris Skinner": A note on weight smoothing in survey sampling

Jae Kwang Kim and HaiYing Wang[1]

Abstract

Weight smoothing is a useful technique in improving the efficiency of design-based estimators at the risk of bias due to model misspecification. As an extension of the work of Kim and Skinner (2013), we propose using weight smoothing to construct the conditional likelihood for efficient analytic inference under informative sampling. The Beta prime distribution can be used to build a parameter model for weights in the sample. A score test is developed to test for model misspecification in the weight model. A pretest estimator using the score test can be developed naturally. The pretest estimator is nearly unbiased and can be more efficient than the design-based estimator when the weight model is correctly specified, or the original weights are highly variable. A limited simulation study is presented to investigate the performance of the proposed methods.

Key Words: Conditional maximum likelihood method; Analytic inference; Score test; Pretest estimation.

1. Introduction

Suppose that the finite population of (x_i, y_i) is an independent and identically distributed (IID) realization of the superpopulation model with density $f(y \mid x; \theta) g(x)$, where θ is the parameter of interest and the marginal density $g(.)$ is completely unspecified. From the finite population, we obtain a probability sample A with a known first-order inclusion probability π_i. We observe (x_i, y_i) in the sample. We are interested in estimating the model parameter θ from the complex sample, which is the main problem in the area of analytic inference in survey sampling. See Korn and Graubard (1999) and Fuller (2009, Chapter 6) for comprehensive overviews of analytic inference in survey sampling.

For efficient estimation, we can construct the conditional likelihood function from the sample as follows:

$$L_c(\theta) = \prod_{i \in A} \frac{f(y_i \mid x_i; \theta) \tilde{\pi}(x_i, y_i)}{\int f(y \mid x_i; \theta) \tilde{\pi}(x_i, y) d\mu(y)} \tag{1.1}$$

where

$$\tilde{\pi}(x, y) = E(\pi \mid x, y) \tag{1.2}$$

is the conditional inclusion probability and $\mu(\cdot)$ is the dominating measure. See Section 8.2 of Kim and Shao (2021) for some details of the conditional maximum likelihood method.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. E-mail: jkim@iastate.edu; HaiYing Wang, Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, U.S.A.

To compute the conditional inclusion probability in (1.2), we can use the formula of Pfeffermann and Sverchkov (1999):

$$E\left(\pi \mid x, y\right) = \frac{1}{E_s\left(w \mid x, y\right)}, \tag{1.3}$$

where $w = \pi^{-1}$ and $E_s(\cdot)$ is the expectation with respect to the sample distribution, the conditional distribution given the sample.

The conditional inclusion probability obtained from (1.3) can be used to calculate the smoothed weight $\tilde{w}_i = \left\{\tilde{\pi}\left(x_i, y_i\right)\right\}^{-1}$. The weight smoothing can reduce the variability of the sampling weight $w_i = \pi_i^{-1}$ in estimating parameters and thus can lead to more efficient estimation, as discussed by Beaumont (2008) and Kim and Skinner (2013). To compute the conditional expectation $E_s\left(w \mid x, y\right)$, we need to build a regression model for w, which can be called a weight model.

In this article, we explore some particular parametric classes of weight models. In Section 2, a weight model using the Beta prime distribution is introduced. In Section 3, a score test for correct model specification in the weight model is proposed. In Section 4, results from a limited simulation study are presented. Some concluding remarks are made in Section 5.

2. Weight model

Because the sampling weights satisfy $w_i \geq 1 \left(i = 1, \ldots, n\right)$, it is assumed that w_i^{-1} are modeled as a Beta distribution $\text{Beta}\left(m\left(x_i, y_i\right)\phi, \left\{1 - m\left(x_i, y_i\right)\right\}\phi\right)$. Thus, the density function satisfies

$$f\left(w^{-1} \mid x, y\right) \propto \left(w^{-1}\right)^{m\phi - 1}\left(1 - w^{-1}\right)^{(1-m)\phi - 1},$$

and the conditional expectation and variance are

$$E\left(w^{-1} \mid x, y\right) = m(x, y), \quad \text{and} \quad V\left(w^{-1} \mid x, y\right) = \frac{m(x, y)\left\{1 - m(x, y)\right\}}{1 + \phi},$$

respectively, where ϕ is the precision parameter. An example of a mean function is the logistic model:

$$m(x, y; \beta) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)}. \tag{2.1}$$

This is essentially a beta regression model. Further details on beta regression can be found in Ferrari and Cribari-Neto (2004).

Unfortunately, the beta regression approach cannot be applied directly because the regression model does not necessarily hold in the sample due to informative sampling. To avoid this problem, we can derive the distribution of the sampled data. Recall that if $X \sim \text{Beta}\left(\alpha, \beta\right)$ then $1 - X$ follows $\text{Beta}\left(\beta, \alpha\right)$ and

$(1-X)/X$ follows a Beta prime distribution $\text{Beta}'(\beta,\alpha)$. Therefore, $o = w-1$ follows $\text{Beta}'\left(\{1-m(x_i,y_i)\}\phi, m(x,y)\phi\right)$, and the density function is expressed as

$$f(o \mid x,y) \propto o^{(1-m)\phi-1}(1+o)^{-\phi}.$$

Based on Bayes' theorem and $w^{-1} = (1+o)^{-1}$, the sampled distribution of o satisfies

$$f_s(o \mid x,y) \propto f(o \mid x,y)\, P(\delta=1 \mid x,y,w) = o^{(1-m)\phi-1}(1+o)^{-\phi-1}, \tag{2.2}$$

which implies $o \mid (x,y,\delta=1) \sim \text{Beta}'\left(\{1-m(x,y)\}\phi, m(x,y)\phi+1\right)$. Thus, we obtain the following.

$$E_s(w \mid x,y) = 1 + E_s(o \mid x,y) = \frac{1}{m(x,y;\beta)} \tag{2.3}$$

and

$$\begin{aligned}\text{Var}_s(w \mid x,y) &= \frac{1-m(x,y)}{m(x,y)}\,\frac{1}{m(x,y)\cdot\phi-1} \\ &\cong \frac{1-m(x,y)}{m(x,y)}\,\frac{1}{m(x,y)\cdot\phi}\end{aligned}$$

for sufficiently large ϕ. Thus, we obtain the following method of moments estimator of ϕ:

$$\hat{\phi} = \frac{1}{n}\sum_{i\in A}\frac{\{w_i \cdot m(x_i,y_i;\beta)-1\}^2}{1-m(x_i,y_i;\beta)} \tag{2.4}$$

which depends on unknown parameter β.

We can use the following iterative estimation procedure estimate model parameters.

1. Compute

$$\hat{\phi}^{(0)} = \frac{1}{n}\sum_{i\in A}\frac{(w_i/\bar{w}-1)^2}{1-1/\bar{w}}$$

as an initial estimator of ϕ, where $\bar{w} = n^{-1}\sum_{i\in S} w_i$.

2. Using $\hat{\phi}^{(t)}$, compute $\hat{\beta}^{(t)}$ by finding the maximizer of

$$\ell_c\left(\beta \mid \hat{\phi}^{(t)}\right) = \sum_{i\in S}\log f_s\left(o_i \mid x_i, y_i; \beta, \hat{\phi}^{(t)}\right)$$

with respect to β, where

$$f_s(o \mid x,y;\beta,\phi) = \frac{\Gamma(\phi+1)}{\Gamma(\phi-m\phi)\,\Gamma(m\phi+1)}\,o^{\phi-m\phi-1}(1+o)^{-\phi-1},$$

and $m = m(x,y;\beta)$.

3. Compute $\hat{\phi}^{(t+1)}$ by applying (2.4) with $\beta = \hat{\beta}^{(t)}$. Iteratively update $\hat{\phi}$ and $\hat{\beta}$ until convergence.

3. Score test for weight model specification

The weight smoothing method in Section 2 is justified under the assumption that the weight model is correctly specified. In practice, we may wish to test for the validity of the weight model before we use the model-based estimator. In this section, we consider a version of the score test for model specification.

Let $\hat{\theta}_c$ be the maximizer of the conditional likelihood function in (1.1). Let $\hat{\theta}_d$ be the design-based estimator of θ that is obtained by maximizing the pseudo log-likelihood function

$$\ell_p(\theta) = \sum_{i \in A} \frac{1}{\pi_i} \log f\left(y_i \mid \mathbf{x}_i; \theta\right). \tag{3.1}$$

The pseudo MLE has been discussed in Chambers and Skinner (2003). Thus, we can develop a test for the following null hypothesis:

$$E\left(\hat{\theta}_d\right) = E\left(\hat{\theta}_c\right). \tag{3.2}$$

However, developing a Wald-type test statistics for the null hypothesis in (3.2) can be cumbersome as the variance-covariance matrix of $\hat{\theta}_d - \hat{\theta}_c$ needs to be estimated.

Instead of testing (3.2), we can consider testing the following null hypothesis

$$H_0 : E\left\{\hat{S}_c\left(\theta_0\right)\right\} = 0, \tag{3.3}$$

where θ_0 is the true parameter and $\hat{S}_c(\theta) = n^{-1} \partial \log L_c(\theta) / \partial \theta$ is the score function obtained from the conditional log-likelihood in (1.1). That is,

$$\hat{S}_c(\theta) = \frac{1}{n} \sum_{i \in A} \left[S\left(\theta; x_i, y_i\right) - E_s\left\{S\left(\theta; x_i, Y\right) \mid x_i\right\} \right],$$

where $S\left(\theta; x, y\right) = \partial \log f\left(y \mid x; \theta\right) / \partial \theta$ and

$$E_s\left\{S\left(\theta; x, Y\right) \mid x\right\} = \frac{\int S\left(\theta; x, y\right) \tilde{\pi}\left(x, y\right) f\left(y \mid x; \theta\right) dy}{\int \tilde{\pi}\left(x, y\right) f\left(y \mid x; \theta\right) dy}.$$

Under some regularity conditions (Binder, 1983), we can establish that

$$\sqrt{n} \left[\hat{S}_c(\theta) - E\left\{\hat{S}_c(\theta)\right\} \right] \xrightarrow{\mathcal{L}} N\left[0, \mathcal{I}_c(\theta)\right], \tag{3.4}$$

as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ denotes the convergence in distribution and

$$\begin{aligned} \mathcal{I}_c(\theta) &= -E\left\{ \frac{\partial}{\partial \theta'} S_c(\theta) \right\} \\ &= n^{-1} \sum_{i=1}^{n} \left[E\left\{ S_i S_i' \tilde{\pi}_i \mid \mathbf{x}_i; \theta \right\} - \frac{\left\{ E\left(S_i \tilde{\pi}_i \mid \mathbf{x}_i; \theta \right) \right\}^{\otimes 2}}{E\left(\tilde{\pi}_i \mid \mathbf{x}_i; \theta \right)} \right]. \end{aligned} \tag{3.5}$$

The proposed test statistic is

$$T\left(\hat{\theta}_d\right) = n\hat{S}_c\left(\hat{\theta}_d\right)' \left\{\mathcal{I}_c\left(\hat{\theta}_d\right)\right\}^{-1} \hat{S}_c\left(\hat{\theta}_d\right)$$

where $\hat{\theta}_d$ is the pseudo MLE of θ_0. Note that

$$T\left(\hat{\theta}_d\right) = T\left(\theta_0\right) + o_p(1),$$

as $\hat{\theta}_d = \theta_0 + o_p(1)$, regardless of whether the weight model holds or not. Under the null hypothesis in (3.3), by (3.4), we can establish that T converges to $\chi^2(q)$ distribution where $q = dim(\theta)$. If the null hypothesis is rejected, then it implies that $\tilde{\pi}(x, y)$ in constructing the conditional likelihood in (1.1) is incorrectly specified. Otherwise, we can safely use the conditional ML estimator.

Strictly speaking, the information matrix in (3.5) ignores the uncertainty of $\hat{\beta}$ in $\tilde{\pi}_i = \tilde{\pi}\left(x_i, y_i; \hat{\beta}\right)$. To incorporate the uncertainty in $\hat{\beta}$, we can consider another information matrix for β. Ignoring the uncertainty in $\hat{\beta}$ will overestimate the variance and lead to a conservative test. See the simulation study in the next section.

4. Simulation study

To test our theory, we performed a limited simulation study. In the simulation, we generate a finite population of size $N = 10{,}000$ and use Poisson sampling to select a sample of expected size $n = 1{,}000$. We repeat this procedure independently $B = 1{,}000$ times.

In each Monte Carlo sample, we generate $\left(x_i, y_i, \pi_i\right)$ for $i = 1, \ldots, N$ where $x_i \sim (0, 2)$, $y_i = \theta_0 + \theta_1 x_i + e_i$, $(\theta_0, \theta_1) = (0.5, 0.5)$, $e_i \sim N\left(0, 0.5^2\right)$, and $\pi_i \mid x_i, y_i \sim \text{Beta}\left(m\left(x_i, y_i\right)\phi, \left\{1 - m\left(x_i, y_i\right)\right\}\phi\right)$, where

$$m(x, y; \beta) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 y\right)} \tag{4.1}$$

with $\beta_1 = 1$, $\beta_2 = 1$, and β_0 being different values for different cases to ensure that $n = 1{,}000$. We used two different values of ϕ, $\phi = 100$ versus $\phi = 1{,}000$, in the simulation study. The weight distribution is less skewed for $\phi = 1{,}000$.

We have four different sampling designs as follows:

Case 1. $\phi = 100$; weight model is specified correctly.

Case 2. $\phi = 100$; lowest 30% π_i's are multiplied by 0.25, i.e., top 30% w_i's in the full data are multiplied by 4. Thus, the weight model (4.1) is incorrectly specified.

Case 3. $\phi = 1{,}000$; weight model is correctly specified.

Case 4. $\phi = 1{,}000$; the lowest 30% π_i's are multiplied by 4, i.e., the top 30% w_i's in the full data are multiplied by 0.25. Thus, the weight model (4.1) is incorrectly specified.

We are interested in estimating θ_0 and θ_1. The following three estimators are considered.

1. PMLE: The pseudo maximum likelihood estimator $\hat{\theta}_d$ maximizing (3.1).
2. CMLE: The conditional maximum likelihood estimator $\hat{\theta}_c$ maximizing (1.1) with $\tilde{\pi}(x, y) = \{\tilde{w}(x, y)\}^{-1}$ and $\tilde{w}(x, y)$ is the smoothed weight under the specified weight model. To avoid numerical problems, we estimate σ^2 in a design-based way.
3. PreTest: The pretest estimator using the score test in Section 3. That is, the pretest estimator $\hat{\theta}_{\text{pre}}$ with $\alpha = 0.05$ is defined as

$$\hat{\theta}_{\text{pre}} = \begin{cases} \hat{\theta}_d & \text{if } T(\hat{\theta}) > q_{0.95}(\chi^2_2) \\ \hat{\theta}_c & \text{otherwise,} \end{cases}$$

where $q_{0.95}(\chi^2_2)$ is the 0.95 quantile of the $\chi^2(2)$ distribution.

Table 4.1 presents the biases, standard errors, and root mean square errors (RMSE) of the three estimators using Monte Carlo samples. The simulation results can be summarized as follows.

1. The PMLE is nearly unbiased for all cases, but it is less efficient than the other methods in Cases 1 and 3, where the weight model is correctly specified.
2. The CMLE is the most efficient but is subject to significant biases when the weight model is incorrectly specified. The efficiency gain is higher for a smaller ϕ, as the distribution of w_i's is more skewed and the advantage of weight smoothing is more significant.
3. The pretest estimator is nearly unbiased for all cases and can be more efficient than the PMLE when the weight model is correctly specified (Case 1 and Case 3), or the original weights are highly variable (Case 2).

Table 4.1
Monte Carlo biases, standard errors (SE) and root mean square errors (RMSE) of the three estimators based on 1,000 Monte Carlo samples

		θ_0			θ_1		
Case	**Method**	**SE**	**Bias**	**RMSE**	**SE**	**Bias**	**RMSE**
1	PMLE	0.0768	-0.001	0.0768	0.0799	0.001	0.0800
	CMLE	0.0608	-0.001	0.0608	0.0425	0.001	0.0425
	PreTest	0.0701	0.006	0.0704	0.0672	-0.004	0.0673
2	PMLE	0.1198	-0.000	0.1198	0.1182	0.008	0.1185
	CMLE	0.0750	0.020	0.0777	0.0375	0.066	0.0764
	PreTest	0.1198	0.001	0.1198	0.1179	0.008	0.1182
3	PMLE	0.0651	0.000	0.0651	0.0645	0.000	0.0645
	CMLE	0.0525	0.002	0.0526	0.0413	-0.002	0.0413
	PreTest	0.0561	0.003	0.0563	0.0499	-0.003	0.0500
4	PMLE	0.0455	0.001	0.0456	0.0432	0.000	0.0432
	CMLE	0.0472	0.053	0.0713	0.0432	-0.127	0.1345
	PreTest	0.0456	0.001	0.0456	0.0433	0.000	0.0433

Note: Pseudo maximum likelihood estimator (PMLE); Conditional maximun likelihood estimator (CMLE).

The rejection rates for the score test are 0.119, 0.952, 0.051, and 0.997 for the four cases, respectively, where the level of significance is $\alpha = 0.05$. The high rejection rate of 0.119 in Case 1 is due to the effect of ignoring uncertainty in weight smoothing. The effect of ignoring the uncertainty in weight smoothing is negligible in Case 3, since the effect of weight smoothing is less significant when ϕ is large. The higher rejection rate indicates that the score test is conservative in adopting the CMLE using $\tilde{w}_i$ over the PMLE.

5. Concluding remark

This article is dedicated to the memory of Professor Chris Skinner. The first author collaborated on various projects with Chris Skinner, and their first research outcome was published in Kim and Skinner (2013). When J.K. Kim visited Chris Skinner at Southampton in the summer of 2011, they first worked on analytic inference under informative sampling, studying the work of Pfeffermann and Sverchkov (1999), but they did not make a connection with weight smoothing at that time. Instead, they mainly focused on the weight smoothing method. About ten years later, we present a method connecting weight smoothing to the likelihood framework.

Weight smoothing is potentially useful, but the correct model specification is required. The pretest estimator using the score test in Section 3 can be used in practice, as it compromises the efficiency of weight smoothing and the robustness of design-based estimation. How to estimate the variance of the pretest estimator has yet to be explored in this paper and will be investigated in the future.

Acknowledgements

The authors thank the Editor and the Assistant Editor, Cynthia Bocci, for their constructive comments. The research of the first author was supported by a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. The second author's research was supported by NSF grant CCF 2105571 and UConn CLAS Research Funding in Academic Themes.

References

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 3, 539-553.

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

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Ferrari, S.L.P., and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31, 407-419.

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

Kim, J.K., and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*. CRC press, 2nd edition.

Kim, J.K., and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100, 358-398.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā, Series B*, 61, 166-186.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 39-68
Statistics Canada, Catalogue No. 12-001-X

Official Statistics based on the Dutch Health Survey during the Covid-19 Pandemic

Jan van den Brakel and Marc Smeets[1]

Abstract

The Dutch Health Survey (DHS), conducted by Statistics Netherlands, is designed to produce reliable direct estimates at an annual frequency. Data collection is based on a combination of web interviewing and face-to-face interviewing. Due to lockdown measures during the Covid-19 pandemic there was no or less face-to-face interviewing possible, which resulted in a sudden change in measurement and selection effects in the survey outcomes. Furthermore, the production of annual data about the effect of Covid-19 on health-related themes with a delay of about one year compromises the relevance of the survey. The sample size of the DHS does not allow the production of figures for shorter reference periods. Both issues are solved by developing a bivariate structural time series model (STM) to estimate quarterly figures for eight key health indicators. This model combines two series of direct estimates, a series based on complete response and a series based on web response only and provides model-based predictions for the indicators that are corrected for the loss of face-to-face interviews during the lockdown periods. The model is also used as a form of small area estimation and borrows sample information observed in previous reference periods. In this way timely and relevant statistics describing the effects of the corona crisis on the development of Dutch health are published. In this paper the method based on the bivariate STM is compared with two alternative methods. The first one uses a univariate STM where no correction for the lack of face-to-face observation is applied to the estimates. The second one uses a univariate STM that also contains an intervention variable that models the effect of the loss of face-to-face response during the lockdown.

Key Words: Small area estimation; Structural time series model; Corona crisis.

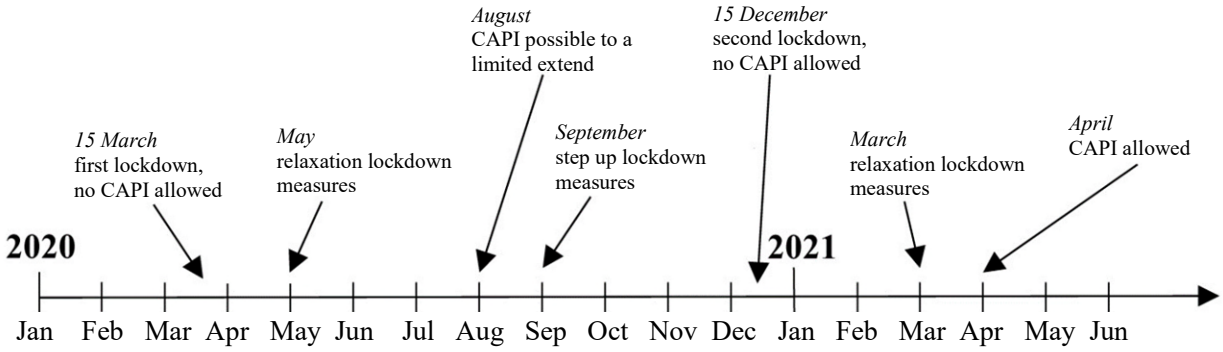
1. Introduction

The Dutch Health Survey (DHS) is a continuing survey conducted by Statistics Netherlands that measures health, healthcare use and lifestyle in the Netherlands. Data collection is based on a sequential mixed-mode design where a combination of web participation (Computer-assisted web interviewing (CAWI)) and face-to-face interviewing (Computer-assisted personal interviewing (CAPI)) is applied. Due to Dutch lockdown measures during the Covid-19 pandemic face-to-face interviewing was not allowed in parts of 2020 and 2021. Figure 1.1 displays a timeline of the lockdowns in the Netherlands and the restrictions on the CAPI mode for the DHS. In the rest of these years, there were restrictions on the normal way of data collection. This results in an abrupt change in the composition of selection effects and measurement bias and therefore results in a systematic effect on the outcomes of the DHS. A second issue is that the DHS is designed to produce reliable estimates on an annual basis, using standard direct estimators like the general regression (GREG) estimator (Särndal, Swensson and Wretman, 1992). The DHS normally publishes on an annual basis for year τ in the month of March of year $\tau+1$. The Covid-19 pandemic that started in the beginning of 2020 made clear that the release of annual data about the effect of Covid-19 on health-related themes with a delay of about one year strongly compromises the

1. Jan van den Brakel, Statistics Netherlands and Maastricht University. E-mail: ja.vandenbrakel@cbs.nl; Marc Smeets, Statistics Netherlands.

relevance of this survey. Another disadvantage of annual figures is that the period of the corona crisis is not well-delineated in the reference period of the DHS. In the second quarter of 2020, there was indeed a strong external demand for quarterly figures of the DHS, since quarterly figures are more timely and better delineate the corona period. The sample size of the DHS, however, does not allow the production of sufficiently precise direct estimates for quarterly reference periods.

Figure 1.1 Timeline of Dutch coronavirus lockdowns and restrictions on CAPI mode for DHS, January 2020 to June 2021.



To solve these issues, a bivariate structural time series model (STM) is developed for eight key variables of the DHS, defined on a quarterly frequency. This model is used to correct for the changes of measurement and selection errors due to the loss of CAPI response and is used as a form of small area estimation (Rao and Molina, 2015) since the model uses sample information observed in previous reference periods to produce sufficiently reliable model-based estimates for quarterly DHS figures. In small area estimation this is commonly called borrowing strength over time.

The models proposed in this paper can be considered as an extension of the area level model (Fay and Herriot, 1979). The extension of the area level model with a temporal component is originally proposed by Rao and Yu (1994). In this paper a time series multilevel model is applied where an AR(1) component for the domain irregular terms is assumed. Other authors who proposed time series multilevel models as an extension of the area level model are Datta, Lahiri, Maiti and Lu (1999), You, Rao and Gambino (2003), You (2008), Boonstra, van den Brakel and Das (2021) and Boonstra and van den Brakel (2022). Another class of time series models that are frequently used as a form of small area estimation are state-space models. Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Pfeffermann and Tiller (2006) and Krieg and van den Brakel (2012) use multivariate state-space models as a form of small area estimation to borrow strength over time and space. Pfeffermann (1991), Harvey and Chung (2000) and van den Brakel and Krieg (2016) propose multivariate time series models as a form of small area estimation for Labour Force surveys that are designed as a rotating panel. The basic difference of the state-space models with aforementioned time series multilevel models is that the population irregular terms are combined with the sampling error into one measurement error. Another difference is that these

models are also applied to time series at the national level to borrow strength over time only in situations where the reference period is too short to collect sufficient data to use a direct estimator even at the national level, see e.g., Pfeffermann (1991), Tiller (1992) and Harvey and Chung (2000). This paper follows the aforementioned state space approach.

Buelens and van den Brakel (2015) proposed a weighting method for sequential mixed-mode designs to stabilize the bias in period-to-period changes that arise from fluctuations in the distributions of respondents over the data collection modes in subsequent editions of a repeated survey. This method assumes a fixed distribution of the population over the different data collection modes, which is added as an additional component to the weighting model of the GREG estimators. This method cannot be considered as an alternative to compensate for the loss of CAPI during the lockdown. The method indeed increases the weights of the CAPI respondents, but will in this case increase selection bias as well because the CAPI respondents are all observed outside the lockdown period.

The net effect of the lack of CAPI is computed based on the response of previous years. This is done by removing CAPI from the response and by reweighting the remaining response. This leads to two direct estimates for one target variable: one based on the complete response (CAWI and CAPI) and one based on only web response (CAWI). In this way quarterly time series can be constructed for DHS that start in the first quarter of 2014: the complete series based on full response and the web series based on web response only. Both series are the input for the bivariate STM. The web series is available in all quarters, also during the lockdown. In quarters without CAPI there are no estimates available for the complete series and the bivariate STM then provides nowcasts for the missing figures based on the web series.

In this paper the bivariate STM is compared with two alternative and more straightforward models. The first one is a univariate STM where no correction for the lack of CAPI is applied. This method applies a univariate STM to the series of direct estimates based on all available response in every quarter. In quarters where CAPI is available the direct estimates are based on both CAWI and CAPI, so they are equal to the estimates of the complete series. In quarters where no CAPI is available the direct estimates are based on only CAWI and are thus equal to the estimates of the web series. The second one is a univariate STM that also contains an intervention variable that models the effect of the loss of CAPI during the lockdown.

The paper is organized as follows. Section 2 gives a description of the Dutch Health Survey and both the univariate and bivariate structural time series models are developed in Section 3. Section 4 explores the results and Section 5 discusses the officially published quarterly DHS figures by Statistics Netherlands. The paper ends with a discussion in Section 6.

2. Dutch Health Survey

The Dutch Health Survey is a continuing survey that measures health, healthcare use and lifestyle in the Netherlands on a yearly basis. The target population is the Dutch population living in private

households. Each month a single-stage stratified sample of approximately 1,250 persons is drawn from the Dutch Personal Records Database. The strata are defined by the municipalities.

Sampled persons are asked to participate via web interviewing (CAWI). Non-respondents are re-approached to participate in a face-to-face interview (CAPI). To reduce administration costs, the fraction of CAWI responses is increased by selecting samples from the CAWI non-respondents that are re-approached through CAPI using a target group strategy that has been used since 2018. CAWI non-respondents are first divided into so-called target groups based on age, income and migration background. From each target group only a sample is re-approached.

Until 2020 there was a yearly response of approximately 10,000 persons, of whom 6,500 responded by CAWI and 3,500 by CAPI. The response is more or less evenly divided over the months. Due to the Covid-19 pandemic that started in 2020 there was a lockdown in the Netherlands that started mid-March 2020. The first relaxations were implemented in May 2020. Due to this lockdown no face-to-face interviews were allowed from mid-March 2020 to the end of July 2020. A second lockdown started in mid-December 2020, which was gradually relaxed from March 2021. This lockdown resulted in a stop of face-to-face interviewing from mid-December 2020 until the end of March 2021. From April 2021 face-to-face interviews were possible again. In order to increase response during the pandemic, persons selected for CAPI were given the opportunity to respond via the internet. This was done by sending an invitation letter when face-to-face interviewing was not allowed and by handing over this letter otherwise. In 2020 only few people used this option and they were considered as CAWI respondents. In 2021 a substantial part of the people selected for CAPI responded via the internet. This response mode will be referred to as CAPI/CAWI response. The resulting response sizes per month and response mode are shown in Table 2.1.

Table 2.1 shows that in 2020 CAPI response is lower in the months March and December and is completely missing from April to July. The large CAWI response size in May is the result of compensation measures taken by Statistics Netherlands for the response gaps that arose due to the lockdown. In 2021 CAPI response is completely missing in the first quarter and is lower in April and May. From June CAPI response seems to recover.

Annual figures are obtained by weighting the response by means of the general regression estimator (Särndal et al., 1992). In this way it is corrected, at least partially, for selective non-response. The weighting model is given by Gender_2 × Age_16 + MaritalStatus_4 + Urbanization_5 + Region_16 + HouseholdSize_5 + Gender_2 × Age_3 × MaritalStatus_4 + Region_4 × Age_3 + Migration Background_4 + SurveySeason_4 + Income_5 + Wealth_5 + TargetGroup_12. The numbers refer to the number of categories and the times sign indicates the use of interaction terms between variables. Note that TargetGroup_12 is included since 2018.

Table 2.1
Response DHS 2020 per mode and month

		CAPI	CAPI/CAWI	CAWI	Total
2020	January	265		584	849
	February	261		586	847
	March	104		917	1,021
	April	0		455	455
	May	0		1,118	1,118
	June	0		708	708
	July	0		483	483
	August	193		527	720
	September	286		259	545
	October	149		763	912
	November	181		587	768
	December	53		286	339
	Total	1,492		7,273	8,765
2021	January	0	48	738	786
	February	0	36	546	582
	March	0	22	655	677
	April	38	77	460	575
	May	51	62	738	851
	June	109	62	283	454
	Total	198	307	3,420	3,925

Note: Dutch Health Survey (DHS); Computer-assisted personal interviewing (CAPI); Computer-assisted web interviewing (CAWI).

In consultation with the main data users of the DHS, i.e., the National Institute for Public Health and Environmental Protection, the Ministry of Health, Welfare and Sports and the Netherlands Institute for Social Research, eight DHS indicators were selected for which a model-based inference method is developed to produce quarterly figures that are corrected for the loss of CAPI during lockdown periods. These eight indicators are perceived health, fraction of people feeling mentally unhealthy, dental visit, GP consult, specialist consult, daily smoking, excessive alcohol consumption and overweight. These indicators cover the three main topics of the survey (perceived) health, healthcare use and lifestyle.

This paper only shows the results of perceived health, dental visit, daily smoking and excessive alcohol consumption. The results of mentally unhealthy are similar to perceived health and the results of the healthcare use variables GP consult and specialist consult are similar to dental visit. Overweight turns out to be a steady indicator and is hardly affected by the Covid-19 pandemic. *Perceived health* is measured for people of all ages. There are five possible answers: very good, good, fair, poor and very poor. Perceived health is the percentage of people that has given one of the positive answers very good or good. *Dental visit* measures the percentage of people of all ages that has visited a dentist in the past four weeks. *Daily smoking* concerns the percentage of people with a daily smoking habit and is measured for people aged 18 years or older. *Excessive alcohol consumption* is measured for the population aged 18 years or older and measures the percentage of people that report a consumption of 21 or more units per week for men or a consumption of 14 or more units per week for women.

3. Structural time series method

3.1 Univariate models

Two univariate STMs are considered. Let $\hat{y}_t^{\mathrm{A}}$ denote the GREG estimate in quarter t for the unknown population parameter based on all the available response. The first univariate STM ignores the loss of CAPI and starts with a measurement error model that states that the sample estimates is the result of the true population parameter, say ϑ_t, for quarter t and a sampling error, say $\varepsilon_t^{\mathrm{A}}$. This leads to the following measurement error model: $\hat{y}_t^{\mathrm{A}} = \vartheta_t + \varepsilon_t^{\mathrm{A}}$. In a next step the population parameter is modelled with a trend that describes the low frequency variation in the series, say L_t, a seasonal component for seasonal fluctuations, say S_t, and a population white noise for the unexplained variation of the population parameter, say I_t. This implies the following so-called basic STM for the population parameter: $\vartheta_t = L_t + S_t + I_t$. Inserting the STM for the population parameter into the measurement error model gives the first univariate STM:

$$\hat{y}_t^{\mathrm{A}} = L_t + S_t + I_t + \varepsilon_t^{\mathrm{A}} \equiv L_t + S_t + e_t^{\mathrm{A}}. \tag{3.1}$$

Note that in (3.1) the population white noise and sampling error are conveniently combined into one measurement error, i.e., $e_t^{\mathrm{A}} = I_t + \varepsilon_t^{\mathrm{A}}$. The trend L_t is modelled by a smooth trend model (Durbin and Koopman, 2012, Chapter 3), given by

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1} \\ R_t &= R_{t-1} + \eta_t^{\mathrm{R}}, \end{aligned} \tag{3.2}$$

where

$$\eta_t^{\mathrm{R}} \sim N\left(0, f_t \sigma_{\mathrm{R}}^2\right), \mathrm{Cov}\left(\eta_t^{\mathrm{R}}, \eta_{t'}^{\mathrm{R}}\right) = 0, \quad \text{for} \quad t \neq t', \quad \text{and} \quad f_t \geq 1.$$

The trend model consists of a level L_t and a slope R_t with a slope disturbance term η_t^{R}. In a standard smooth trend model, the variance of the slope disturbance terms are time invariant, i.e., $f_t = 1$ for all t. The variance of the slope disturbance terms σ_{R}^2, which are estimated by maximum likelihood (see Subsection 3.4), determines the flexibility of trend model (3.2). For some variables the Covid-19 pandemic causes a sudden strong increase in the quarter-to-quarter changes of the direct estimates. Particularly at the start of the Covid-19 pandemic, the maximum likelihood estimates for σ_{R}^2 are based on the period-to-period changes observed in the past. A sudden increase in the period-to-period changes of the input series therefore results in a temporarily miss-specification of the STM. Or to phrase it differently, for some variables the assumption that the volatility of the period-to-period changes is not affected by the Covid-19 pandemic is violated. To avoid temporal miss-specification of the STM model at the start of the Covid-19 pandemic, the flexibility of the trend model is increased by defining a time-dependent variance for the slope disturbance terms. This is achieved by multiplying the maximum

likelihood estimate for σ_{R}^2 with a factor $f_t \geq 1$. As a result, the variance of the slope disturbance terms is equal to $f_t \sigma_{\mathrm{R}}^2$. Values for f_t are determined outside the model, as explained in Section 4. This approach is initially proposed by van den Brakel, Souren and Krieg (2022) and is compared with alternative approaches to account for sudden shocks in the input series of an STM due to the Covid-19 pandemic.

Increasing the variance of the slope disturbance terms through factors f_t has the following interpretation. As the variance of the slope disturbance terms increases, the influence of more distant observations on the level of the trend becomes smaller. The proposed approach implies that the filtered estimates attach less weight to the prediction based on observations from the past and more weight to the direct estimates obtained in the last month. This seems reasonable in periods where the world suddenly changes and becomes incomparable with the past, as was the case with the COVID-19 pandemic.

The seasonal component S_t is modelled by a trigonometric seasonal model (Durbin and Koopman, 2012, Chapter 3), given by

$$S_t = \gamma_{1,t} + \cdots + \gamma_{J/2,t}, \tag{3.3}$$

where

$$\gamma_{j,t} = \gamma_{j,t-1}\cos\left(\frac{\pi j}{J/2}\right) + \gamma_{j,t-1}^{*}\sin\left(\frac{\pi j}{J/2}\right) + \omega_{j,t}$$

$$\gamma_{j,t}^{*} = \gamma_{j,t-1}^{*}\cos\left(\frac{\pi j}{J/2}\right) - \gamma_{j,t-1}\sin\left(\frac{\pi j}{J/2}\right) + \omega_{j,t}^{*} \quad \text{for} \quad j = 1, \ldots, J/2.$$

For quarters $J = 4$, it holds that

$$S_t = \gamma_{1,t} + \gamma_{2,t}, \tag{3.4}$$

with harmonics

$$\gamma_{1,t} = \gamma_{1,t-1}^{*} + \omega_{1,t},$$

$$\gamma_{1,t}^{*} = -\gamma_{1,t-1} + \omega_{1,t}^{*},$$

$$\gamma_{2,t} = -\gamma_{2,t-1} + \omega_{2,t}.$$

Note that the last component defined by (3.3) equals $\gamma_{2,t}^{*} = \gamma_{2,t-1}^{*} + \omega_{2,t}^{*}$ and can be left out since $\gamma_{2,t}^{*}$ is not used in the previous three harmonics and also does not play a role in the measurement equation. The following assumptions for the seasonal disturbance terms,

$$\omega_{1,t} \sim N\left(0, \sigma_{\omega}^2\right), \omega_{1,t}^{*} \sim N\left(0, \sigma_{\omega}^2\right), \omega_{2,t} \sim N\left(0, \sigma_{\omega}^2\right),$$

and

$$\text{Cov}\left(\omega_{j,t}, \omega_{j,t'}\right) = 0, \quad \text{for } t \neq t' \text{ and } j = 1, 2$$

$$\text{Cov}\left(\omega^*_{1,t}, \omega^*_{1,t'}\right) = 0, \quad \text{for } t \neq t'$$

$$\text{Cov}\left(\omega_{j,t}, \omega^*_{1,t}\right) = 0, \quad \text{for all } t \text{ and } j = 1, 2$$

$$\text{Cov}\left(\omega_{1,t}, \omega_{2,t}\right) = 0, \quad \text{for all } t.$$

The Covid-19 pandemic may influence both the trend and the seasonal pattern. Since it is not possible to estimate a structural change in the seasonal pattern due to the Covid-19 pandemic, with less than one year of observations during the Covid-19 pandemic it is assumed that there is only an effect on the development of the trend. The seasonal component S_t is therefore modelled by a trigonometric seasonal model with a time-independent variance. In this way the seasonal pattern is modelled dynamically and therefore has the flexibility to accommodate effects of the Covid-19 pandemic on the seasonal pattern.

To accommodate heteroscedasticity caused by e.g., changes in response size and the sample design, the measurement error e_t^{A} is scaled with the standard error of the input series of $\hat{y}_t^{\text{A}}$ (Binder and Dick, 1990):

$$e_t^{\text{A}} = \sqrt{\hat{\text{V}}\left(\hat{y}_t^{\text{A}}\right)}\, \tilde{e}_t^{\text{A}}, \tag{3.5}$$

$$\tilde{e}_t^{\text{A}} \sim N\left(0, \sigma^2_{\text{e, A}}\right),$$

$$\text{Cov}\left(\tilde{e}_t^{\text{A}}, \tilde{e}_{t'}^{\text{A}}\right) = 0, \quad \text{for } t \neq t',$$

and with $\hat{\text{V}}\left(\hat{y}_t^{\text{A}}\right)$ the variance estimate of $\hat{y}_t^{\text{A}}$. It is understood that $\hat{\text{V}}\left(\hat{y}_t^{\text{A}}\right)$ is estimated outside the STM from the sample data and that these estimates are used as *a priori* known values in the STM. Note that in (3.5) a multiplicative model is chosen for the variance structure of the measurement error. As an alternative an additive structure of the form could be considered. Note that $\hat{\text{V}}\left(\hat{y}_t^{\text{A}}\right)$ in (3.5) is not the real population variance but an estimate of the variance that is subject to uncertainty and can over or under estimate the real variance. The advantage of a multiplicative model is that it scales the variance of the GREG estimator and has the flexibility to reduce the variance if $\hat{\text{V}}\left(\hat{y}_t^{\text{A}}\right)$ over-estimates the real variance. Similar variance structures are used by e.g., Binder and Dick (1990), van den Brakel and Krieg (2015), Elliot and Zong (2019) and Gonçalves, Hidalgo, Silva and van den Brakel (2022).

Model (3.1) borrows strength from the past through both the trend L_t and the seasonal pattern S_t in order to improve the accuracy of the direct estimates. Model (3.1) also accounts for a sudden increase of the volatility of the population parameter by making the trend temporarily more flexible. To account for sudden changes in measurement and selection errors due to the loss of CAPI during the lockdown, model (3.1) is extended with an intervention variable. This gives rise to the second univariate model:

$$\hat{y}_t^{\mathrm{A}} = L_t + S_t + \beta \frac{x_t}{3} + e_t^{\mathrm{A}}. \tag{3.6}$$

Here x_t is the number of months in quarter t without CAPI response and β a regression coefficient that can be interpreted as the net effect of the change in measurement and selection bias due to the loss of CAPI. In a quarter with full CAPI response, $x_t/3=0$ and β is switched off. In a quarter without any CAPI respondents, $x_t/3=1$ and β absorbs the effect of the loss of CAPI and avoids that the model estimates for the population parameter ϑ_t are affected, at least partially. If a quarter only contains one or two months without CAPI, then $x_t/3=1/3$ or $x_t/3=2/3$ respectively and the correction of β contributes proportionally to the number of months without CAPI in that quarter. The trend, seasonal component and measurement error are defined in (3.2), (3.3), and (3.5), respectively.

Compared to model (3.1) it is expected that model (3.6) better accommodates for the loss of CAPI during the lockdown. Model (3.6), however, assumes no structural change in the evolution of the population parameter ϑ_t. If the lockdown results in e.g., strong turning points in the population parameter, it can be expected that this is partially and incorrectly absorbed in the regression coefficient of the intervention variable. To accommodate for this risk, the bivariate model, proposed in the next section is developed.

3.2 Bivariate model

The input series for the bivariate model are the quarterly direct estimates based on the complete response, denoted $\hat{y}_t^{\mathrm{C}}$ (*complete series*) and the quarterly direct estimates based on the web response only, denoted $\hat{y}_t^{\mathrm{W}}$ (*web series*). The systematic difference between both series observed during the years before the start of the Covid-19 pandemic is used in a bivariate STM to make model-based estimates for the population parameter that correct for the loss of CAPI during the lockdown. The bivariate STM given by:

$$\begin{pmatrix} \hat{y}_t^{\mathrm{C}} \\ \hat{y}_t^{\mathrm{W}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(L_t + S_t \right) + \begin{pmatrix} 0 \\ \lambda_t \end{pmatrix} + \begin{pmatrix} e_t^{\mathrm{C}} \\ e_t^{\mathrm{W}} \end{pmatrix}. \tag{3.7}$$

The first component states that $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{W}}$ are two estimates for the unknown population parameter that is decomposed in a trend and a seasonal component. The population irregular term I_t is combined with the sampling errors, similar to the univariate models. The trend L_t is modelled by the smooth trend model were the variance of the slope disturbance terms is made time varying, as defined by equation (3.2) and the seasonal component S_t by the trigonometric model given by equation (3.4). The second component of (3.7), i.e., λ_t, models the systematic difference between the regular series and the web series as a random walk, given by

$$\lambda_t = \lambda_{t-1} + \eta_{\lambda,t}, \tag{3.8}$$

where

$$\eta_{\lambda,t} \sim N\left(0, \sigma_\lambda^2\right)$$

$$\mathrm{Cov}\left(\eta_{\lambda,t}, \eta_{\lambda,t'}\right) = 0, \quad \text{for } t \neq t'.$$

Because a random walk is assumed, the model accommodates gradual changing differences between $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{W}}$. The third component of (3.7) contains the measurement error. They contain the sampling error of $\hat{y}_t^k$ and the population irregular term, i.e., $e_t^k = I_t + \varepsilon_t^k$ for $k \in \{\mathrm{C}, \mathrm{W}\}$. The measurement error component accommodates heteroscedasticity by scaling the measurement error with the sampling error of the input series and accounts for the positive correlation between $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{W}}$ that arises because both estimates use the same web respondents. This is achieved with the following measurement error model:

$$e_t^k = \sqrt{\hat{\mathrm{V}}\left(\hat{y}_t^k\right)}\, \tilde{e}_t^k, \quad \text{with } \hat{\mathrm{V}}\left(\hat{y}_t^k\right) \text{ the variance estimate of } \hat{y}_t^k \tag{3.9}$$

and

$$\tilde{e}_t^k \sim N\left(0, \sigma_{\mathrm{e},k}^2\right)$$

$$\mathrm{Cov}\left(\hat{y}_t^{\mathrm{C}}, \hat{y}_t^{\mathrm{W}}\right) = \frac{\sqrt{n_t^{\mathrm{W}}}}{\sqrt{n_t^{\mathrm{C}}}} \sqrt{\hat{\mathrm{V}}\left(\hat{y}_t^{\mathrm{C}}\right)} \sqrt{\hat{\mathrm{V}}\left(\hat{y}_t^{\mathrm{W}}\right)}$$

$$\mathrm{Cov}\left(\tilde{e}_t^k, \tilde{e}_{t'}^k\right) = 0, \quad \text{for } t \neq t'.$$

The covariance between the measurement errors is obtained as follows. Following Kish (1965), the correlation between two variables observed in two partial overlapping samples is given by

$$\mathrm{Cor}\left(z_1, z_2\right) = \rho \frac{n_{1\cap 2}}{\sqrt{n_1}\sqrt{n_2}},$$

where

- z_1 the variable observed in sample s_1 of size n_1,
- z_2 the variable observed in sample s_2 of size n_2,
- $n_{1\cap 2}$ the size of the sample overlap between s_1 and s_2,
- ρ the correlation between z_1 and z_2 based on the $n_{1\cap 2}$ respondents that are included in s_1 and s_2.

In this application, sample s_1 is the sample with complete response and s_2 the sample with CAWI respondents. Suppose that $z_1 = \hat{y}_t^{\mathrm{C}}$, $z_2 = \hat{y}_t^{\mathrm{W}}$, $n_1 = n_t^{\mathrm{C}}$, and $n_2 = n_t^{\mathrm{W}}$, with n_t^{C} is the size of the complete response in quarter t and n_t^{W} the size of the web response in quarter t. In this case the sample overlap is also the sample with CAWI respondents. Therefore we have $n_{1\cap 2} = n_t^{\mathrm{W}}$ and $\rho = 1$. From this it follows that

$$\mathrm{Cor}\left(\hat{y}_t^{\mathrm{C}}, \hat{y}_t^{\mathrm{W}}\right) = \rho \frac{n_t^{\mathrm{W}}}{\sqrt{n_t^{\mathrm{C}}}\sqrt{n_t^{\mathrm{W}}}} = \frac{\sqrt{n_t^{\mathrm{W}}}}{\sqrt{n_t^{\mathrm{C}}}}$$

and

$$\text{Cov}\left(\hat{y}_t^{\text{C}}, \hat{y}_t^{\text{W}}\right) = \frac{\sqrt{n_t^{\text{W}}}}{\sqrt{n_t^{\text{C}}}} \sqrt{\hat{\text{V}}\left(\hat{y}_t^{\text{C}}\right)} \sqrt{\hat{\text{V}}\left(\hat{y}_t^{\text{W}}\right)}.$$

As a result, the covariance matrix for the measurement errors in (3.7) is given by

$$\begin{pmatrix} e_t^{\text{C}} \\ e_t^{\text{W}} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{\text{V}}\left(\hat{y}_t^{\text{C}}\right)\sigma_{\text{e,C}}^2 & \text{Cov}\left(\hat{y}_t^{\text{C}}, \hat{y}_t^{\text{W}}\right) \\ \text{Cov}\left(\hat{y}_t^{\text{C}}, \hat{y}_t^{\text{W}}\right) & \hat{\text{V}}\left(\hat{y}_t^{\text{W}}\right)\sigma_{\text{e,W}}^2 \end{pmatrix}\right]. \tag{3.10}$$

Similar to the univariate models, $\hat{\text{V}}\left(\hat{y}_t^{\text{C}}\right)$, $\hat{\text{V}}\left(\hat{y}_t^{\text{W}}\right)$, and $\text{Cov}\left(\hat{y}_t^{\text{C}}, \hat{y}_t^{\text{W}}\right)$ are estimated outside the STM from the sample data. These estimates are used as *a priori* known values in STM (3.7).

During the lockdown, $\hat{y}_t^{\text{C}}$ is missing but $\hat{y}_t^{\text{W}}$ is observed. With bivariate STM (3.7) it is possible to obtain estimates for the trend $\left(L_t\right)$ and the signal $\left(L_t + S_t\right)$ of the population parameters of interest. These estimates are corrected for the bias due to the loss of CAPI, because the model accounts for the systematic difference between $\hat{y}_t^{\text{C}}$ and $\hat{y}_t^{\text{W}}$ through the second model component $\left(\lambda_t\right)$. This correction relies on the assumption that the systematic difference between $\hat{y}_t^{\text{C}}$ and $\hat{y}_t^{\text{W}}$ as observed before the start of the Covid-19 pandemic does not change during the lockdown.

3.3 Direct estimates for time series models

For the DHS direct quarterly estimates can be computed starting in the first quarter of 2014. From the first quarter of 2014 up to the last quarter of 2019 these direct estimates are based on the weighted annual DHS response obtained by applying the GREG estimator. Quarterly estimates $\hat{y}_t^{\text{C}}$ for the complete series are obtained by computing the domain estimator based on the GREG estimator with quarter t as domain. Quarterly estimates $\hat{y}_t^{\text{W}}$ are obtainedby recalculating the GREG estimator using the CAWI response only and subsequently computing the domain estimator based on the GREG estimator with quarter t as the domain. In the quarters before 2020 there was no loss of CAPI and the direct estimates $\hat{y}_t^{\text{A}}$ are equal to $\hat{y}_t^{\text{C}}$. Standard errors are computed in R (R Core Team, 2015) with the package "survey" (Lumley, 2014). For the estimation of the standard errors the sample design of the DHS is taken into account, where the stratification is based on the cross-classification of months and provinces. Here provinces are used, because the subdivision into municipalities leads to strata with too little response.

Since the decision to publish quarterly figures was made in June 2020, the direct estimates for the first two quarters of 2020 are based on the weighted response based on the GREG estimator available from January to June 2020. Estimates for the third quarter of 2020 are based on the weighted response available from January to September 2020 and the fourth quarter is based on the weighted annual response of 2020. For all quarters the same weighting model and the same population totals of the covariates are used. Direct estimates for the first quarter of 2021 are computed in a similar way and are based on the response from

January to March 2021. Estimates for the second quarter of 2021 are based on the response from January to June 2021. In this way the estimates $\hat{y}_t^{\mathrm{W}}$ for the web series are obtained for all quarters of 2020 and for the first two quarters of 2021.

Adding quarterly samples during an ongoing year results in a progressively larger annual data set. The main advantage of this approach is that all available data are used for the weighting scheme of the GREG estimator. Note that this will only slightly increases the heterogeneity between the quarterly direct estimates, since the variance of the quarterly direct estimates is of the order of the quarterly sample size, not the total sample size. There might be a minor effect since the fluctuation of the GREG weights decreases if the sample used for weighting increases. This is, however, not an issue since the variance of the measurement errors is taken proportional to the variance of the GREG estimates used in the input series, as can be seen from formula (3.9) and (3.10). This approach also does not create additional dependency between the quarterly estimates, since there is no sample overlap between the quarterly estimates and the variance of the GREG estimates are based on the GREG residuals, which are assumed to be independent.

For the complete series $\hat{y}_t^{\mathrm{C}}$ in 2020 the second quarter is missing and the other quarters are based on response where CAPI is partially missing (Table 2.1). In the first quarter of 2020 CAPI is only missing in the last two weeks of March and for this quarter it is assumed that sufficient CAPI response is available to obtain plausible estimates. So in the first quarter of 2020 the estimates $\hat{y}_t^{\mathrm{C}} = \hat{y}_t^{\mathrm{A}}$ are based on the available CAWI and CAPI response and in the second quarter of 2020 $\hat{y}_t^{\mathrm{C}}$ is missing and $\hat{y}_t^{\mathrm{A}} = \hat{y}_t^{\mathrm{W}}$. In the third quarter of 2020 CAPI response is only available in August and September. Here a correction is applied to $\hat{y}_t^{\mathrm{C}}$ based on the bivariate model (3.7). The direct estimate $\hat{y}_t^{\mathrm{C}}$ for the third quarter of 2020 is obtained by computing the domain estimator of the GREG applied to the available response in August and September minus $1/3$ of the difference $\hat{\lambda}_t$ estimated by model (3.7) in the second quarter. No correction is applied to the corresponding standard errors. The direct estimate $\hat{y}_t^{\mathrm{A}}$ in the third quarter is equal to the uncorrected weighted mean of the available response in August and September. In the fourth quarter of 2020 CAPI is also missing for only two weeks and it is assumed that there is enough CAPI response available to obtain plausible estimates, so $\hat{y}_t^{\mathrm{C}} = \hat{y}_t^{\mathrm{A}}$.

In 2021 there is besides CAPI and CAWI also CAPI/CAWI response (Section 2). To find out how to use the CAPI/CAWI response in the best possible way, two scenarios were elaborated. In the first scenario quarterly figures are computed where CAPI/CAWI response is considered as CAPI and in the second scenario CAPI/CAWI response is considered as CAWI. Since there were no major differences in the results of both scenarios the CAPI/CAWI response is considered as CAWI. Results of this comparison are not shown in this paper. In the first quarter of 2021 $\hat{y}_t^{\mathrm{C}}$ is missing and $\hat{y}_t^{\mathrm{A}} = \hat{y}_t^{\mathrm{W}}$ and in the second quarter of 2021 CAPI is available and so $\hat{y}_t^{\mathrm{C}} = \hat{y}_t^{\mathrm{A}}$.

In this way input series for models (3.1), (3.6) and (3.7) are obtained. The series run from the first quarter in 2014 up to the second quarter in 2021. The series $\hat{y}_t^{\mathrm{A}}$ and $\hat{y}_t^{\mathrm{W}}$ are available for all quarters and for the series $\hat{y}_t^{\mathrm{C}}$ estimates are missing in the second quarter of 2020 and in the first quarter of 2021.

3.4 Model-based estimates

Given the series of direct estimates $\hat{y}_t^{\mathrm{A}}$, $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{W}}$, model-based estimates based on one of the models (3.1), (3.6) or (3.7) can be produced. To this end the three models are expressed in state space representation, where after the Kalman filter is applied to obtain optimal estimates for the state variables, i.e., the variables that define the trend $\left(L_t, R_t\right)$, the seasonal component $\left(\gamma_{1,t}, \gamma_{1,t}^{*}, \gamma_{2,t}\right)$, and the bias parameter $\left(\lambda_t\right)$. The Kalman filter assumes that values for the hyperparameters, i.e., the variances of the measurement errors and state disturbance terms $\left(\sigma_{\mathrm{R}}^2, \sigma_{\omega}^2, \sigma_{\lambda}^2, \sigma_{\mathrm{e,A}}^2, \sigma_{\mathrm{e,C}}^2, \sigma_{\mathrm{e,W}}^2\right)$, are known. Estimates for these hyperparameters are obtained with maximum likelihood. To this end a likelihood function, obtained by the one-step-ahead error decomposition, is maximized using numerical optimization algorithm MaxBFGS. The Kalman filter is a recursive algorithm that runs from $t=1$ to the last observation of the series and gives optimal estimates with their standard errors for the state variables and the signal for each period t based on the observed series until period t. These are the so-called filtered estimates. The filtered estimates of past state vectors can be updated if new data become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. In this application, interest is mainly focused on the filtered estimates, since they are based on the complete set of information that would be available in the regular production process to produce a model-based estimate for quarter t. The state variables in the Kalman filter are initialized with a diffuse initialization, which means that the starting values for the state variables are equal to zero with a very large standard error. After a few iterations, the filtered estimates for the states converge to a proper distribution. For this reason the filtered estimates for the states of the first d periods of the series are ignored in the analysis, where d equals the number of state variables with a diffuse initialization. See Durbin and Koopman (2012) for more details of the state space representation of the STMs, the Kalman filter and the maximum likelihood estimation procedure for the hyperparameters. The computations are conducted with Ssfpack 3.0 (Koopman, Shephard and Doornik, 2008) in combination with Ox (Doornik, 2009).

The Kalman filter provides optimal estimates for the state variables. For this application the trend $\left(L_t\right)$ and the signal $\left(L_t + S_t\right)$ of the population parameter are of particular interest, since these are the variables that are published as official quarterly health indicators. Standard errors of these estimates are obtained from the Kalman filter recursion. These standard errors do not account for the additional uncertainty that arises since the values of the hyperparameters are replaced by their maximum likelihood estimates in the Kalman filter recursions. This is the standard approach in state space applications, but it will result in over-optimistic estimates for the standard errors. Note that Pfeffermann and Tiller (2005) propose a bootstrap that accounts for the additional uncertainty of the maximum likelihood estimates of the hyperparameters in the Kalman filter.

Model selection is based on likelihood-based model diagnostics such as the AIC and BIC (Durbin and Koopman, (2012, Chapter 7)). The normality assumptions of the state disturbance terms in the STMs presented in Subsections 3.1 and 3.2 imply that the standardized innovations or one-step-ahead predictions are standard normally distributed. For all three models it is evaluated whether they meet these underlying

assumptions by testing to which extent the standardized innovations are standard normally and independently distributed. This is done by testing the standardized innovations on normality using Bowman-Shenton normality test, drawing QQ-plots and histograms of the standardized innovations. Sample autocorrelograms and the Durbin Watson test are applied to test for serial correlation in the standardized innovations. An F-test for heteroscedasticity is applied to test for equal variance of the standardized innovations. Finally, time series plots of the standardized innovations are drawn to check for outliers. For more details on these tests it is referred to Durbin and Koopman (2012, Chapter 2). These model diagnostics indicate that the underlying model assumptions of the finally selected models are not seriously violated.

In quarters where CAPI is missing, additional assumptions for the three STMs are required. For the univariate STM (3.1) it is assumed that there are no mode effects between CAPI and CAWI. For the univariate STM (3.6) it is assumed that the trend and the seasonal component correctly describe the evolution of the population parameter and that sudden strong changes in the true values of the population parameter, such as turning points, are not partially absorbed in the level intervention component. These assumptions are evaluated in Section 4. For the bivariate STM (3.7) it is assumed that the difference between CAWI and CAPI response does not change due to the Covid-19 pandemic. This implies that the composition of the web response does not change during the Covid-19 pandemic. It is not possible to verify whether is assumption is met. A response analysis showed that no structural change in the CAWI response and non-response distributions before and after the start of the corona crisis is observed. There were also no structural difference between the answer categories under the CAPI and the CAWI response before the first lockdown and the third and fourth quarter of 2020 where CAPI was started up again. See also the results for the bias parameter λ_t in the bottom-right panels of Figures 4.5-4.8 in Section 4.

4. Results time series models

The three models are fitted to the series of direct estimates as described in Subsection 3.3. Due to the Covid-19 pandemic some DHS variables show a strong increase in the quarter-to-quarter changes, especially at the beginning of the two lockdown periods. In these periods, the smooth trend model is not flexible enough to follow the increased period-to-period movements of the input series. This can be expected since the flexibility of the trend, which is determined by the variance of the slope disturbance terms of the trend model, is based on the quarter-to-quarter movements observed in the period before the Covid-19 crisis. A sudden increase in the dynamics of the population parameter results in temporary miss-specification of the STM, which becomes visible in large values for the standardized innovations in these periods. To accommodate in the STM for the suddenly increased volatility of the population parameters, the flexibility of the smooth trend is temporarily increased by multiplying the variance of the slope disturbance terms $\left(\sigma_{\mathrm{R}}^2\right)$ in (3.2) by a time-dependent factor $f_t \geq 1,$ as explained in Subsection 3.1.

The values for f_t are chosen in such a way that the standardized innovations in the period during the start of the Covid-19 pandemic have values within or just outside the admissible range of 1.96 in absolute

terms. In this way, the value of the factor $f_t > 1$ is kept as small as possible, so that the model can still borrow strength from the past. Note that adjusting the variance σ_R^2 in quarter t influences the slope disturbance term from quarter $t+1$ and the trend only from quarter $t+2$. So there is a lag of two quarters in the effect of the outcomes after adjustment of σ_R^2. Thus to increase the flexibility of the slope in Q2 of 2020, the value of σ_R^2 must be increased at the latest in Q4 of 2019. For several variables it was necessary to increase the variance already in Q3 2019. To avoid a large sudden change in the variance of the slope disturbance terms, the values of f_t are slightly increased in the quarters preceding Q3 2019. In the quarters after the first lockdown in Q2 2020, the values of f_t are reduced to 1 as soon as possible.

From the analysis of the standardized innovations it follows that for most variables it is necessary to make the slope more flexible during the pandemic. Table 4.1 shows the values of the factors $f_t > 1$ for models (3.1), (3.6) and (3.7). In quarters where $f_t = 1$ no values are shown. Variables for which it was not necessary to make the slope more flexible are not shown in the tables either. For a correct interpretation, the values for f_t must be compared with the maximum likelihood estimates for σ_R^2 in Table 4.2. For perceived health and dental visit a flexible slope is applied in the quarters before the first lockdown, i.e., the second quarter of 2020. For daily smoking, $f_t > 1$ only for the univariate STM without intervention (3.1) and only before the second lockdown. For excessive alcohol assumption it is not necessary to make the trend more flexible. The factors in Table 4.1 are relatively large compared to the values $\hat{\sigma}_R$ of in Table 4.2. Because the variances of the slope disturbance terms are generally small, large values for f_t are required to give the trend component sufficient flexibility to follow the strong period-to-period changes at the start of the corona crisis. Note this is an empirical result that differs between applications.

Table 4.1
Values of flexibility parameter f_t in quarters where $f_t > 1$. In quarters and for variables where no value is displayed, $f_t = 1$. In the first two quarters of 2021 and in the quarters before the third quarter of 2018, $f_t = 1$ for all variables

		2018 Q4	2019 Q1	2019 Q2	2019 Q3	2019 Q4	2020 Q1	2020 Q2	2020 Q3	2020 Q4
Univariate STM without intervention	Perceived health			10	100	100	100	10		
	Dental visit				10	100	100	10		
	Daily smoking								10	50
Univariate STM with intervention	Perceived health		10	100	200	100	100	10		
	Dental visit	10	100	1,000	5,000	8,000	100	10		
Bivariate STM	Perceived health			10	100	100	100	10		
	Dental visit				10	100	100	10		

Note: Structural time series model (STM).

Figures 4.1-4.4 display the standardized innovations for perceived health estimated by the three models (3.1), (3.6) and (3.7). For all series the innovations, estimated by the model where the variance of the slope disturbance terms is not temporarily increased (black dashed line), exceed the interval of (-1.96, 1.96) implying that the model is miss-specified at the start of the first lockdown. By making the

slope more flexible the standardized innovations (red solid line) get admissible values. The standardized innovations for the other variables are not shown here. After setting the values for f_t, the underlying model assumptions are evaluated by testing whether the standardized innovations are standard normally and independently distributed. For all three models the performed tests (Section 3.4) show some small violations of these assumptions for some of the variables. Alternative model formulations did not further improve the model diagnostics.

Figure 4.1 Standardized innovations for perceived health estimated by univariate STM without intervention (3.1).

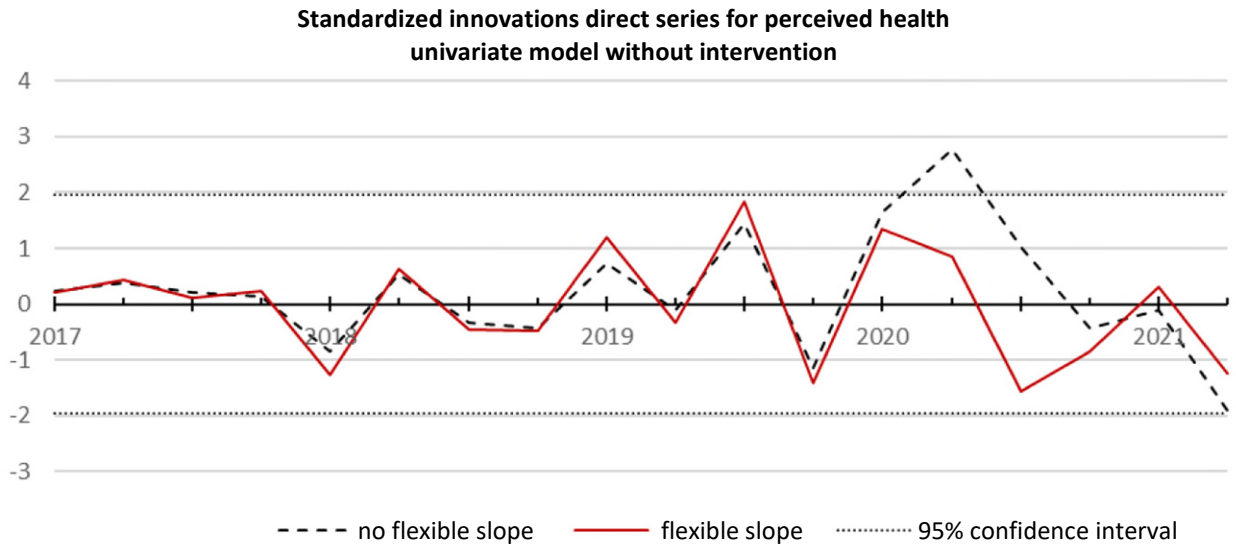


Figure 4.2 Standardized innovations for perceived health estimated by univariate STM with intervention (3.6).

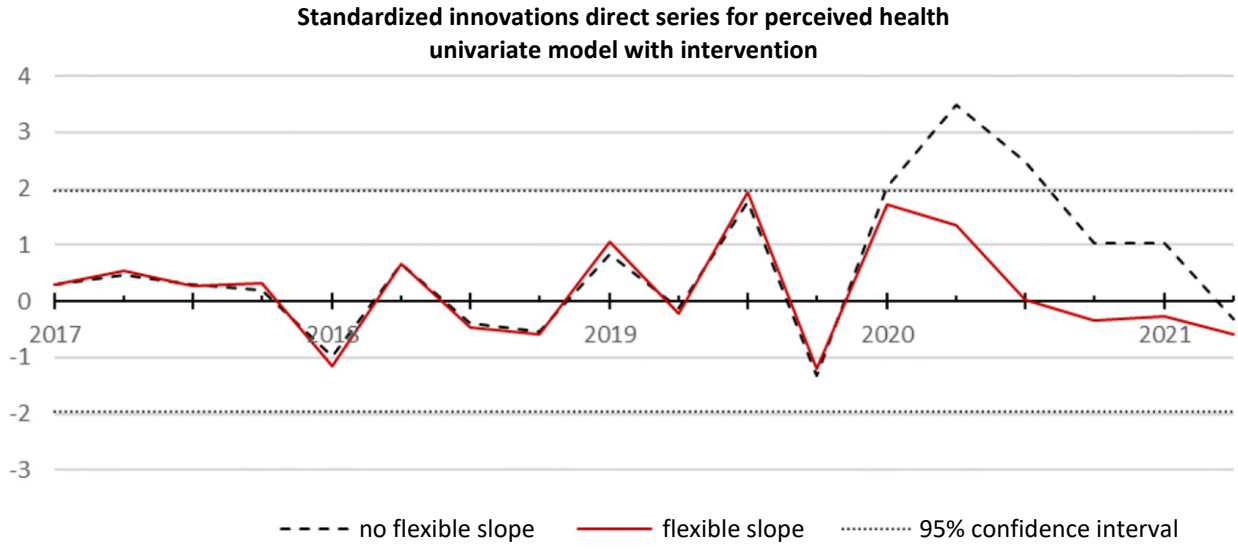


Figure 4.3 Standardized innovations complete series for perceived health estimated by bivariate STM, given by (3.7).

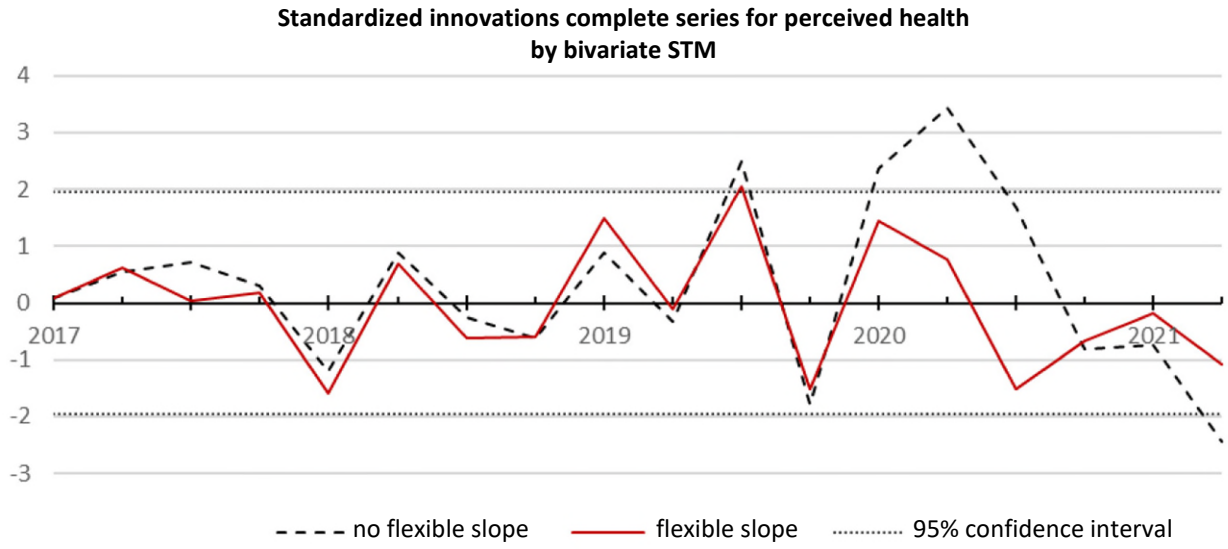


Figure 4.4 Standardized innovations web series for perceived health estimated by bivariate STM, given by (3.7).

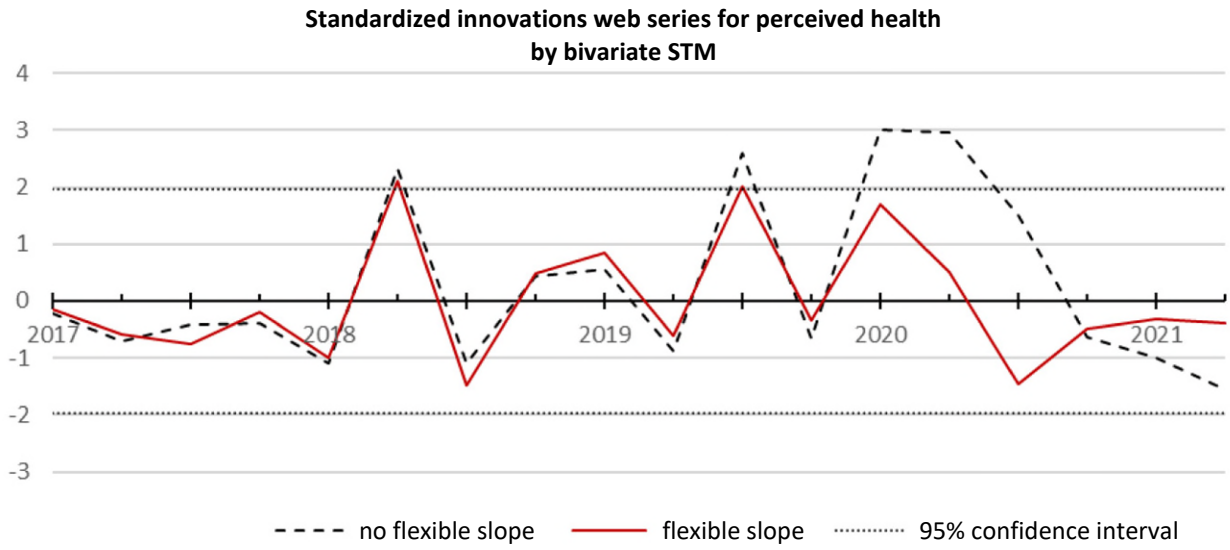


Table 4.2 gives the real-time or concurrent maximum likelihood estimates of the hyperparameters of the three STMs. This means that the maximum likelihood estimates are based on the series observed until the particular quarter in the table. In order to show the values of the hyperparameters before the pandemic, the estimates are also displayed for the second quarter of 2019. Even though the variance σ_{R}^2 is multiplied by a factor f_t in the model, it can be seen that in many cases the (square root of the) variance estimate $\hat{\sigma}_{\mathrm{R}}$

increases. The largest increases occur for dental visit before the first lockdown. For daily smoking the variance estimate $\hat{\sigma}_{\mathrm{R}}$ increases in the second quarter of 2021 for the univariate model.

The estimates of some of the variance components in Table 4.2 are very small. This is the case with $\hat{\sigma}_{\mathrm{R}}$ for perceived health, daily smoking and excessive alcohol consumption and $\hat{\sigma}_{\omega}$ for daily smoking. These hyperparameters could, on the one hand, be removed from the model and it can therefore be assumed that the trend and seasonal components are time invariant. The slope disturbance terms, however, cannot be removed from the model because the flexibility of the trends needed to be increased during the corona crisis by increasing the variance of the slope disturbance terms. Also the variance of the seasonal disturbance terms are kept to make the models more robust for changes in the seasonal pattern during the corona crisis. In a similar way the $\hat{\sigma}_{\lambda}$ for dental visits could be set to zero, but that would make the assumption that the difference between CAPI and CAWI after the start of the corona crisis did not change even stronger.

Figures 4.5-4.8 show the results of the estimates for the variables under the three models. The displayed series start in the first quarter of 2017. Since a diffuse initialisation of the Kalman filter is applied, the model predictions for the first three years obtained with the STM are ignored. For all variables, four graphs are displayed. The first one compares the direct estimates $\hat{y}_t^{\mathrm{C}}$ (*dir compl*) and $\hat{y}_t^{\mathrm{W}}$ (*dir web*) with the model-based estimates $\hat{L}_t + \hat{S}_t$ based on the bivariate STM (*STM biv*), the univariate model without intervention (*STM univ*) and the univariate model with intervention (*STM univ with int*). The second graph shows the estimated standard errors of the quarterly estimates of the point estimates presented in the first graph. The graphs in the bottom-left panel shows the intervention coefficient β of the univariate model (*intervention STM univ*) of STM (3.6). The graph in the bottom-right panel shows the systematic difference λ_t (syst. diff. web and compl. resp.) of STM (3.7) together with the 95% confidence intervals.

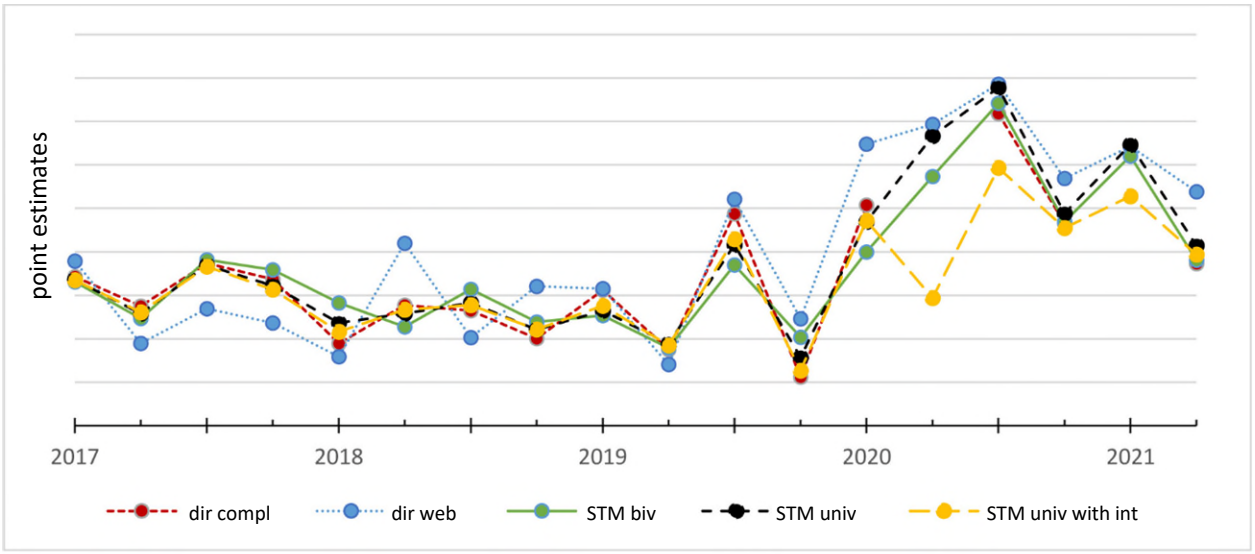
By comparing the series of the direct estimates based on the complete response and the web response and by analysing the estimates of the systematic difference (λ_t) it follows for most variables that there is a clear mode effect between the CAPI and CAWI response. This is picked up by the λ_t parameter of the bivariate model. For perceived health the differences between the series with and without CAPI are relatively small (Figure 4.5, top panel). For dental visit, CAWI respondents score higher than CAPI respondents and the systematic difference λ_t varies between 1.5% and 2% (Figure 4.6, top panel). For daily smoking and excessive alcohol consumption it is just the other way around (Figures 4.7 and 4.8, top panel). For these variables CAPI scores are higher than CAWI and for daily smoking the difference is the largest with a systematic difference, measured by λ_t, of around -4%. This illustrates that ignoring the effect of the loss of CAPI during the lockdown, results in a substantial bias in the direct estimates. Combining direct quarterly estimates that are based on CAWI only for the lockdown periods with estimates based the complete response obtained in forgoing or preceding periods of the lockdown in one time series, would result in misleading period-to-period changes during the Covid-19 period. See e.g., the top panel of Figure 4.7 for daily smoking.

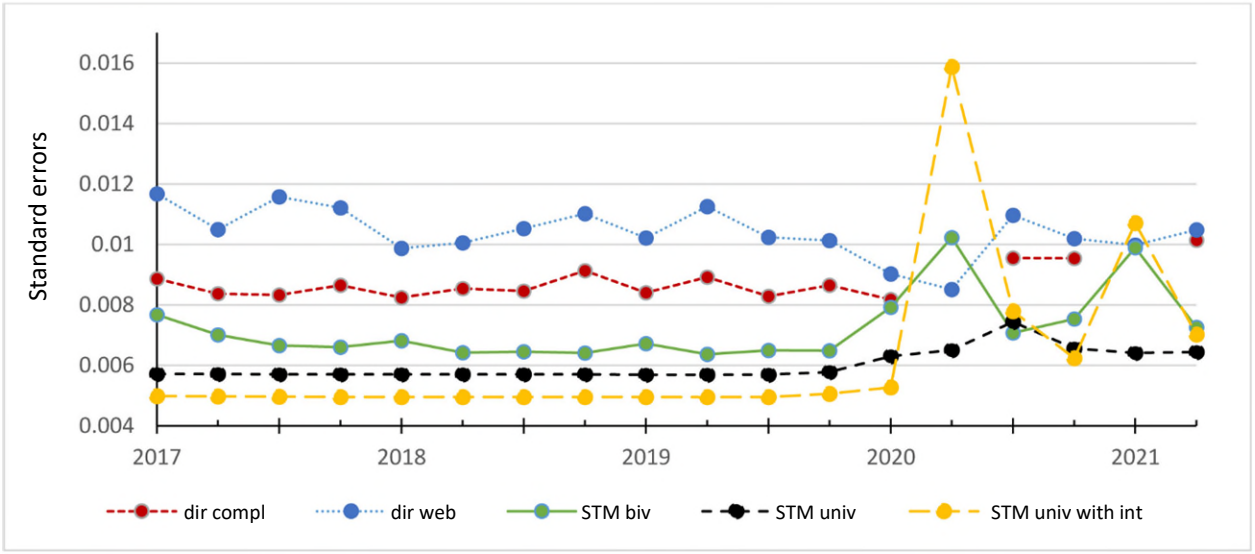
Table 4.2
Concurrent maximum likelihood estimates hyperparameters STM

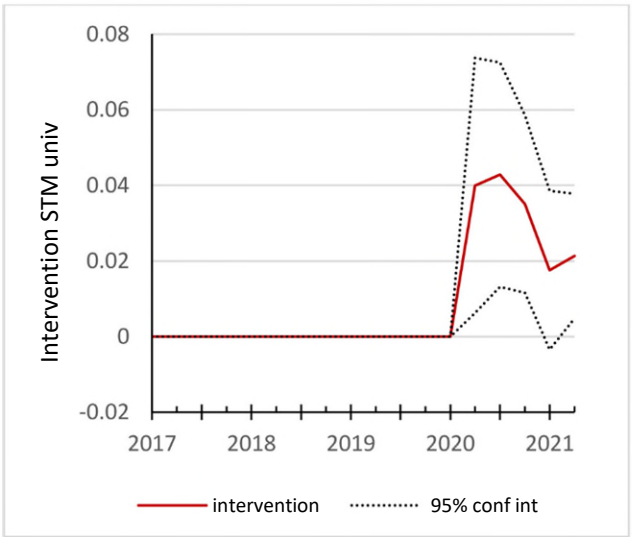
	Perceived health						**Daily smoking**					
	2019 Q2	**2020 Q2**	**2020 Q3**	**2020 Q4**	**2021 Q1**	**2021 Q2**	**2019 Q2**	**2020 Q2**	**2020 Q3**	**2020 Q4**	**2021 Q1**	**2021 Q2**
	Univariate STM without intervention						Univariate STM without intervention					
$\hat{\sigma}_{R}$	<0.001	0.001	<0.001	0.001	0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.002
$\hat{\sigma}_{\omega}$	<0.001	0.002	0.001	0.002	0.002	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{e,A}$	0.007	0.007	0.008	0.007	0.007	0.007	0.010	0.011	0.010	0.01	0.010	0.010
	Univariate STM with intervention						Univariate STM with intervention					
$\hat{\sigma}_{R}$	<0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\omega}$	<0.001	0.002	0.002	0.002	0.002	0.001	<0.001	<0.001	<0.001	0.001	<0.001	0.002
$\hat{\sigma}_{e,A}$	0.007	0.006	0.006	0.007	0.008	0.008	0.010	0.009	0.010	0.009	0.009	0.009
	Bivariate STM						Bivariate STM					
$\hat{\sigma}_{R}$	<0.001	0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\omega}$	<0.001	<0.001	0.002	0.002	0.002	0.002	<0.001	<0.001	<0.001	0.001	0.003	0.003
$\hat{\sigma}_{\lambda}$	<0.001	<0.001	0.002	<0.001	<0.001	0.004	0.002	0.003	0.007	0.01	0.002	0.002
$\hat{\sigma}_{e,C}$	0.957	1.120	0.979	0.934	0.928	0.862	1.310	1.220	1.200	1.250	1.710	1.710
$\hat{\sigma}_{e,W}$	1.310	1.340	1.180	1.240	1.240	1.040	1.280	1.210	1.100	0.653	0.718	0.760
	Dental visit						**Excessive alcohol consumption**					
	Univariate STM without intervention						Univariate STM without intervention					
$\hat{\sigma}_{R}$	<0.001	0.003	0.005	<0.001	0.002	0.003	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\omega}$	<0.001	<0.001	<0.001	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	<0.001
$\hat{\sigma}_{e,A}$	0.009	0.008	0.008	0.011	0.009	0.010	0.008	0.010	0.009	0.009	0.009	0.010
	Univariate STM with intervention						Univariate STM with intervention					
$\hat{\sigma}_{R}$	<0.001	0.003	0.005	<0.001	0.002	0.003	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\omega}$	<0.001	<0.001	<0.001	0.004	0.003	0.002	0.002	<0.001	<0.001	0.001	0.001	<0.001
$\hat{\sigma}_{e,A}$	0.009	0.008	0.008	0.011	0.009	0.010	0.008	0.010	0.010	0.010	0.009	0.009
	Bivariate STM						Bivariate STM					
$\hat{\sigma}_{R}$	<0.001	0.003	0.006	0.004	0.003	0.006	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\omega}$	<0.001	<0.001	<0.001	0.004	0.004	0.001	0.002	0.004	<0.001	<0.001	<0.001	<0.001
$\hat{\sigma}_{\lambda}$	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.003	0.002	0.002	0.002	0.003	0.002
$\hat{\sigma}_{e,C}$	1.080	1.130	1.100	0.713	0.624	0.954	1.360	1.460	1.450	1.480	1.490	1.470
$\hat{\sigma}_{e,W}$	1.070	1.080	1.140	1.050	1.010	1.190	0.822	1.070	1.050	1.040	1.040	1.050

Note: Structural time series model (STM).

Figure 4.5 Results STM for perceived health.







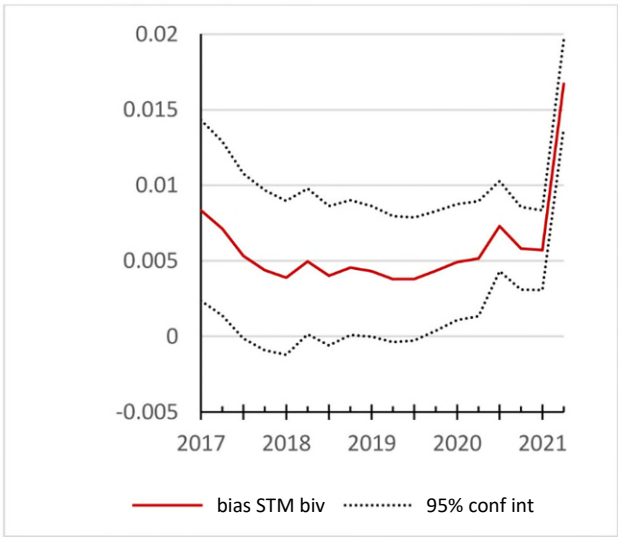


Figure 4.6 Results STM for dental visit.

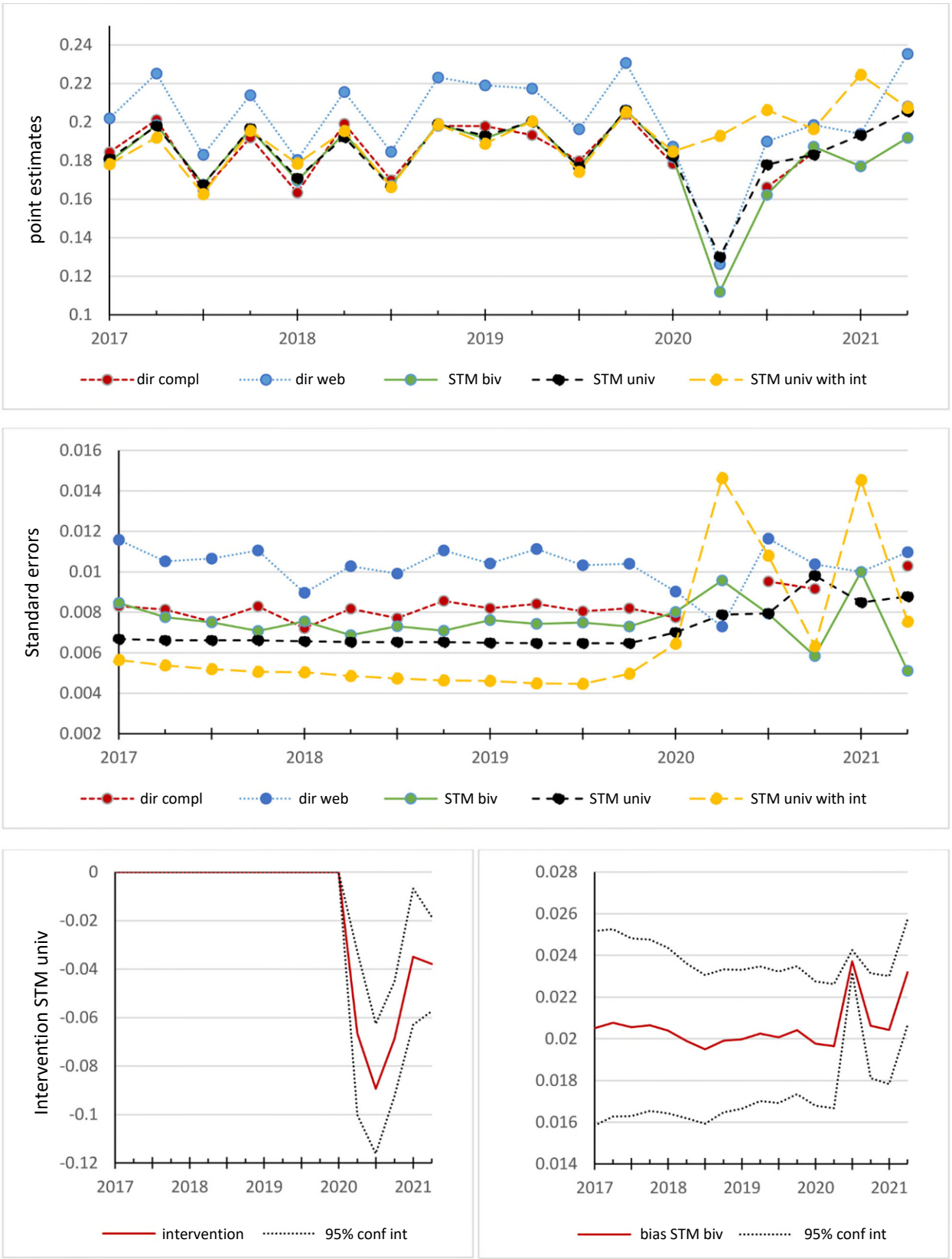


Figure 4.7 Results STM for daily smoking.

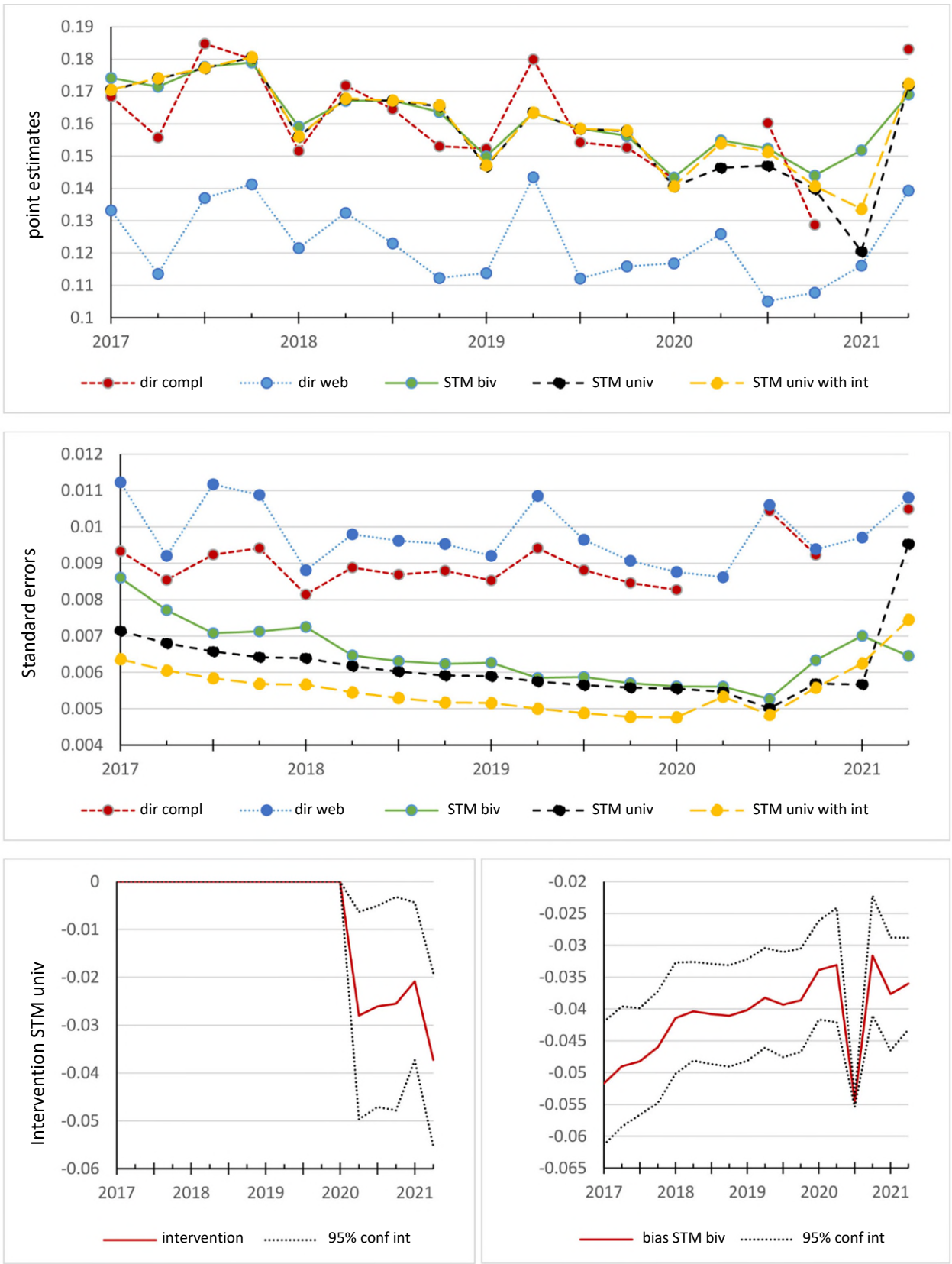
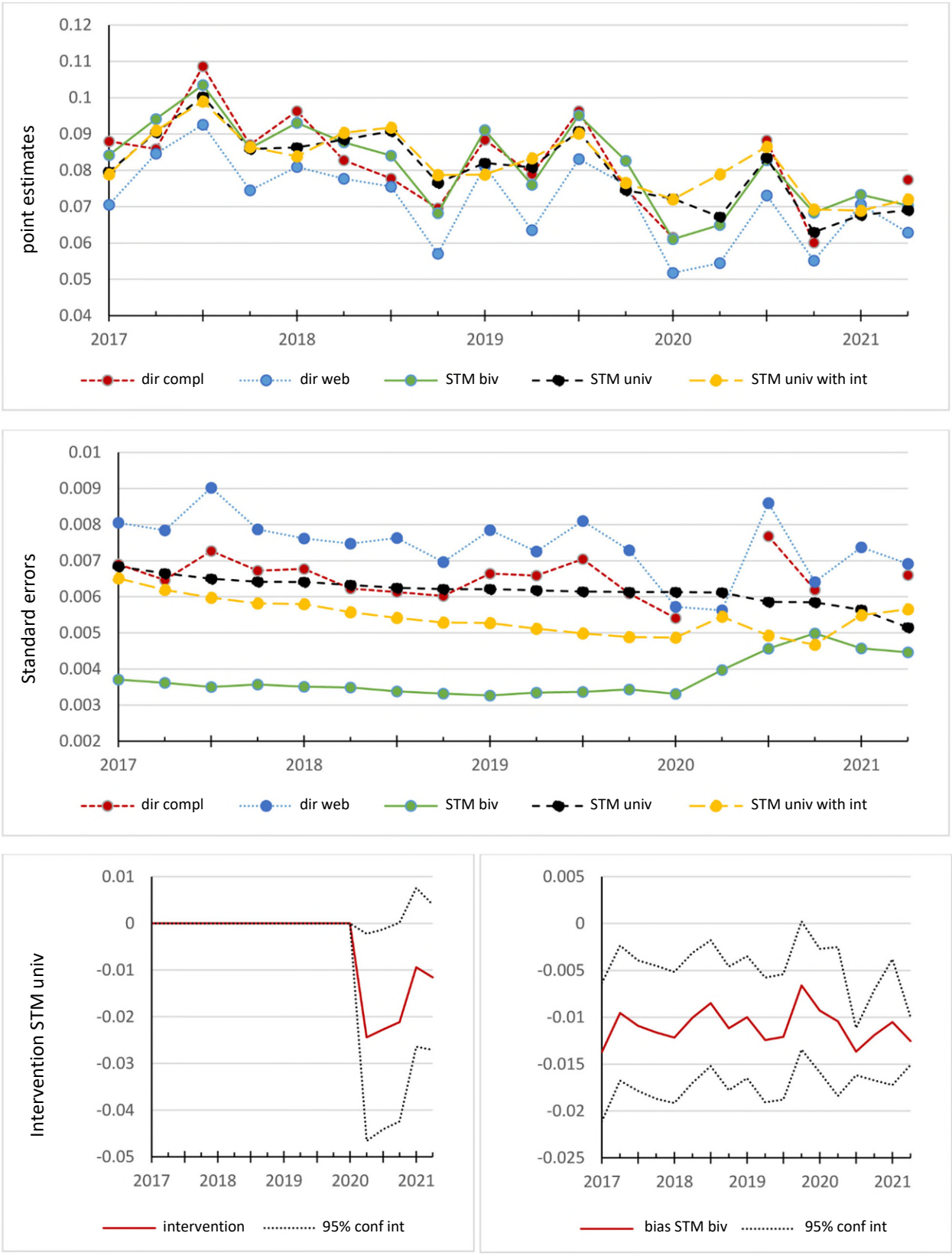


Figure 4.8 Results STM for excessive alcohol consumption.



Until 2020 there was no loss of CAPI and the STM estimates based on the univariate and bivariate models are very similar. During the Covid-19 pandemic that started in 2020 there are more clear differences between the STM estimates. Especially in quarters where CAPI is missing and for variables with a clear mode effect the univariate STM without intervention produces estimates at the level of the web series while the STM estimates by the bivariate model are at the level of the complete series. That is for example the case in the first quarters of 2020 for perceived health (Figure 4.5, top panel) and dental visit (Figure 4.6, top panel) and in the first quarter of 2021 for daily smoking (Figure 4.7, top panel). For excessive alcohol consumption similar effects are found in 2020 and 2021 (Figure 4.8, top panel), but to a lesser extent. The univariate STM without intervention produces, as expected, biased estimates during the Covid-19 pandemic in quarters where CAPI is partially or completely missing.

The univariate STM with intervention also leads to biased estimates in quarters where CAPI is partially or completely missing during one of the lockdowns. This is because the model incorrectly interprets a part of the sudden changes in the real quarterly developments as differences in measurement bias and selection effects. This can result in a large estimate for the intervention coefficient β. The effect can be seen for all variables, but is the largest for dental visit (Figure 4.6, bottom-left panel). For dental visit the resulting bias is the largest in the second quarter of 2020, when dentists in the Netherlands were only open for emergency treatments.

The bivariate STM avoids that sudden changes in the developments of the population parameter are interpreted as differences in measurement and selection bias, because nowcasts are obtained for the missing estimates based on the complete response by means of the systematic difference λ_t in the model observed in the period before the lockdown. Estimates based on the bivariate STM are at the level of the complete series and are therefore used as the official quarterly DHS figures, since they provide the most plausible correction for the loss of the CAPI respondents.

For most variables the standard errors of the STM estimators are smaller than those of the direct estimators and the standard errors of the estimates based on the univariate models are generally smaller than those based on the bivariate model. At first sight it might come as a surprise that the standard errors under the bivariate model are larger than those of the univariate models. It should be understood that the series based on CAWI is based on the same respondents that are also used in the series of the complete response. Therefore the CAWI series does not provide new sample information to the time series model. This is reflected in the covariance structure of the measurement errors (3.10). From that perspective the univariate models are more parsimonious resulting in smaller standard errors for the parameter estimates of interest. In quarters where the flexibility parameter $f_t > 1,$ the models assign more weight to the direct estimates and less strength is borrowed from the past. This results in larger standard errors that sometimes exceed the standard errors of the direct estimates. For the univariate STM with intervention this effect is large in the second quarter of 2020 (see e.g., the middle panel of Figure 4.5).

5. Official publications based on the DHS

Official quarterly figures have been published for the eight selected DHS variables (Section 2) based on the bivariate STM (3.7). The first quarterly series were published in August 2020. These series ran from the first quarter in 2017 up to the second quarter of 2020. Subsequently, new estimates were published every quarter. The quarterly figures are computed in real time and will not be revised after publication. Based on the quarterly figures also quarterly and annual developments are published. Quarterly developments are defined as the difference between two consecutive quarters and the annual developments as the difference between the same quarters in two consecutive years. The developments can be directly derived from the published quarterly figures. Standard errors for the quarterly developments are obtained by calculating the linear combination $\Delta_t^{\mathrm{Q}} = L_{t|t} - L_{t-1|t} + S_{t|t} - S_{t-1|t}$ via the Kalman filter recursion in (3.10) and (3.11). For the annual developments the standard errors are computed by calculating the linear combination of trends $\Delta_t^{\mathrm{A}} = L_{t|t} - L_{t-4|t}$. Here the linear combination of signals $L_{t|t} - L_{t-4|t} + S_{t|t} - S_{t-4|t}$ has not been used, because in that case many extra state variables should be kept in the state vector in order to compute the seasonal components $S_{t-4|t}$. This may lead to unstable estimates.

The annual DHS figures for 2020 and 2021 have been benchmarked with the quarterly figures by extending the regular weighting model described in Section 2 with the quarterly STM estimates for the eight variables for which STMs are developed. For each variable a component is constructed with eight categories that is added to the weighting model. Each target variable specifies the distribution over two categories, i.e., the fraction of people that meet the characteristic of that variable (e.g., daily smoker) and a rest category (e.g., not being a daily smoker). The components in the weighting model specify the distribution of the population over these two categories on a quarterly basis. The numbers per quarter are divided by four, such that the sum over the eight categories is equal to the size of the target population. In this way numerical consistency is achieved between the annual and quarterly publications. There is also a correction for the loss of CAPI for more detailed breakdowns of the eight variables. And finally a best possible correction is realized for the loss of CAPI for other related variables for which no model-based quarterly estimates are developed. Quarterly and annual publications for 2017, 2018 and 2019 have not been made consistent with each other, since revisions are undesired and since the size of the revision is small because there was no loss of CAPI response during this period.

The extension of the weighting model with the quarterly STM estimates resulted in a slight increase of the dispersion of the regression weights. Table 5.1 shows some results of the annual DHS figures for 2020, including the variables cancer (ever had) and bronchitis (past 12 months). The estimates in the table are percentages and the corresponding standard errors are given in parentheses. The corrections to the annual figures for the variables for which quarterly figures have been estimated are in line with the previous results discussed in Section 4. For perceived health and dental visit there is a negative correction for the loss of CAPI in 2020, while CAWI respondents score higher than CAPI respondents (Section 4). For daily smoking, and excessive alcohol consumption the correction is positive, while CAWI scores

lower than CAPI. One would expect that cancer is related to the lifestyle variables, but this variable is negatively corrected from 6.47 (regular weighting) to 6.44 (extended weighting). At first sight it appears that for this variable the correction by means of the model-based quarterly figures does not work very well. On the other hand, this variable concerns all types of cancer and the relationship may be less strong and it can be anticipated that the majority of people that faced cancer in the past gave up smoking afterwards. For bronchitis, where a strong relation is expected with daily smoking, the correction is indeed in the same direction as for daily smoking.

Table 5.1
Results annual figures DHS 2020. Estimates are in percentages and standard errors in parentheses

Variable	Regular weighting	Extended weighting
Perceived health	81.70 (0.45)	81.46 (0.46)
Dental visit	16.83 (0.42)	16.08 (0.42)
Daily smoking	13.61 (0.45)	14.87 (0.49)
Excessive alcohol consumption	6.43 (0.30)	6.93 (0.33)
Cancer	6.47 (0.26)	6.44 (0.26)
Bronchitis	4.28 (0.23)	4.33 (0.23)

Note: Dutch Health Survey (DHS).

6. Discussion

Based on the Dutch Health Survey (DHS), until 2020 only annual figures on health, healthcare use and lifestyle were published by Statistics Netherlands. As a result of the Covid-19 pandemic and the associated lockdown it was decided in June 2020 to publish a series of quarterly figures based on a structural time series model (STM) for a selection of eight DHS key variables. This serves multiple purposes. Firstly, with quarterly figures the period of the corona crisis can be better delineated, so that possible effects of the crisis on the health figures is portrayed more clearly. Secondly, quarterly figures are more timely available, namely already during the statistical year and not only after the end of the reference year. This clearly increases the relevance of the health figures. Because the sample size of the DHS is too small to produce sufficiently precise quarterly figures with a direct estimator, structural time series models are used as a form of small area estimation to improve the precision of the quarterly figures with sample information from preceding reference periods. And finally, the bivariate time series model corrects for the bias that is a result of the loss of face-to-face observation during the lockdown.

The bivariate STM combines two series of direct estimates, a series based on complete response and a series based on web response. The differences between the complete series and the series based on web response are modelled dynamically in a separate component as a random walk. In quarters where face-to-face response is missing, there are no estimates available based on the complete response. For these periods, the bivariate model provides nowcasts for the population parameter of interest that are not affected by the sudden change in measurement and selection effects that are the result of the loss of CAPI

because the model accommodates this difference in the aforementioned component. This approach is based on the assumption that the observed differences between the two input series in the period before the lockdown, do not change during the lockdown. The validity of this assumption is difficult to evaluate, but it has been established through a response analysis that the composition of the web response did not change during the corona crisis.

Two univariate STMs are considered as an alternative. The univariate model without an intervention component to model the shock in the input series that is the result of the loss of CAPI response, assumes that there are no mode effects between web response and face-to-face response. For the selected DHS variables there are clearly mode effects implying that this univariate STM produces biased estimates in quarters during the lockdown when there is no or less face-to-face observation possible. The second univariate STM attempts to model the change in measurement and selection bias with a level intervention variable. This is also a less optimal solution, since the lockdown also has a strong effect on the population parameters. A part of the real evolution of the population parameters is incorrectly absorbed in the level intervention, resulting in biased model predictions for the population parameters of interest. For these reasons the univariate models are unsuitable for estimating quarterly figures during the Covid-19 pandemic. Based on the bivariate STM official quarterly figures are published for the eight selected DHS variables.

The corrections for the loss of face-to-face interviewing have been incorporated in the annual figures of 2020 and 2021 by including in the weighting model of the annual response a table with the corrected model-based quarterly figures for the eight selected DHS variables. This provides numerical consistency between quarterly and annual figures. In this way a correction is also realized for the loss of face-to-face response for more detailed breakdowns of the annual figures of these eight variables and to some extent also for other related variables for which no model-based quarterly estimates are developed.

An essential advantage of using the STM is that model-based estimates are more accurate than direct estimates. In particular, period-by-period developments can be estimated much more accurately thanks to the positive correlation between trend estimates and consecutive periods.

For some variables the pandemic has had a major effect on the development. In order to account for the sudden increase in the dynamics of these figures in the time series model, it is necessary to make the trend component more flexible during the pandemic. This has been done by increasing the variance of the disturbance terms of the trend component during the pandemic. A consequence is that the standard errors of the model-based estimates increase for these quarters and are in some cases larger than the standard errors of the direct estimates.

The Covid-19 crisis increased the awareness that variance is not the only quality concept for official statistics, but that other quality dimensions such as timeliness and comparability over time are at least as important. As a result of this, Statistics Netherlands extended the traditional design-based inference approach for the annual publications of the DHS, with a model-based inference method as a form of small area estimation to produce more timely figures. At the same time, the proposed method compensates for

the bias that occurs as a result of the temporal loss of CAPI responses to maintain comparability over time and avoid a sudden increased MSE.

Acknowledgements

The authors wish to thank two anonymous referees and the Associate Editor for careful reading of the first draft of this paper and providing constructive comments. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

References

Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 2, 239-253. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14533-eng.pdf.

Boonstra, H.J., and van den Brakel, J.A. (2022). Multilevel time series models for small area estimation at different frequencies and domain levels. *Annals of Applied Statistics*. Accepted for publication.

Boonstra, H.J, van den Brakel, J.A. and Das, S. (2021). Multilevel time series modeling of mobility trends. *Journal of the Royal Statistical Society A series*, 184, 985-1007.

Buelens, B., and van den Brakel, J.A. (2015). Measurement error calibration in mixed-mode surveys. *Sociological Methods & Research*, 44, 391-426.

Datta, G., Lahiri, P., Maiti, T. and Lu, K. (1999). Hierarchical bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Doornik, J.A. (2009). *An Object-oriented Matrix Programming Language Ox 6*. London: Timberlake Consultants Press.

Durbin, J., and Koopman, S. (2012). *Time Series Analysis by State Space Methods (second edition)*. Oxford University Press, Oxford.

Elliot, D., and Zong, P. (2019). Improving timeliness and accuracy of estimates from the UK labour force survey. *Statistical Theory and Related Fields*, 3, 186-198.

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

Gonçalves, C., Hidalgo, L., Silva, D. and van den Brakel, J.A. (2022). Model-based single-month unemployment rate estimates for the Brazilian Labour Force Survey. *Journal of the Royal Statistical Society*, 185, 1707-1732.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.

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

Koopman, S.J., Shephard, N. and Doornik, J.A. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form.* London: Timberlake Consultants Press.

Krieg, S., and van den Brakel, J.A. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics and Data Analysis*, 56, 2918-2933.

Lumley, T. (2014). 'survey': Analysis of Complex Survey Samples. R package version 3.30.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.

Pfeffermann, D., and Bleuer, S.R. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 2, 149-163. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14458-eng.pdf.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26(6), 893-916.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, 2nd edition, New York: John Wiley & Sons, Inc.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 22(4), 511-528.

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

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

van den Brakel, J.A., and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, 2, 267-296. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14231-eng.pdf.

van den Brakel, J.A., and Krieg, S. (2016). Small area estimation with statespace common factor models for rotating panels. *Journal of the Royal Statistical Society A series*, 179, 763-791.

van den Brakel, J.A., Souren, M. and Krieg, S. (2022). Estimating monthly Labour Force Figures during the COVID-19 pandemic in the Netherlands. *Journal of the Royal Statistical Society, Series A*, 185, 1560-1583.

You, Y. (2008). Small area estimation using area level models with model checking and applications. SSC annual meeting, Proceedings of the Survey Methods Section.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 69-87
Statistics Canada, Catalogue No. 12-001-X

Combining data from surveys and related sources

Dexter Cahoy and Joseph Sedransk[1]

Abstract

To improve the precision of inferences and reduce costs there is considerable interest in combining data from several sources such as sample surveys and administrative data. Appropriate methodology is required to ensure satisfactory inferences since the target populations and methods for acquiring data may be quite different. To provide improved inferences we use methodology that has a more general structure than the ones in current practice. We start with the case where the analyst has only summary statistics from each of the sources. In our primary method, uncertain pooling, it is assumed that the analyst can regard one source, survey *r*, as the single best choice for inference. This method starts with the data from survey *r* and adds data from those other sources that are shown to form clusters that include survey *r*. We also consider Dirichlet process mixtures, one of the most popular nonparametric Bayesian methods. We use analytical expressions and the results from numerical studies to show properties of the methodology.

Key Words: Administrative data; Bayesian methods; Clustering; Dirichlet process mixture; Pooling data; Survey sampling.

1. Introduction

With substantially reduced response rates and limited budgets there has been an increased emphasis on efficient use of all of the information available to the survey analyst. Specifically, one may be able to improve inferences by using results from several sample surveys and related sources such as administrative records. The methodology that we use to combine information has more structure than the methods currently used in survey sampling, so should lead to better inferences. Starting with the data from the survey that is the best choice for inference, these data are augmented with other, concordant, data. We use analytical expressions and the results from numerical studies to show properties of the methodology.

This research was motivated by a study (Ha and Sedransk, 2019) of health insurance coverage in Florida's counties where the authors noted very different estimates from three surveys. Correspondingly, we could have estimates from a well established probability survey and two non-probability surveys. In either case there is the question about how to make better inferences.

In the sequel we refer to the collection of studies as "surveys", recognizing that these may be probability surveys, non-probability surveys, administrative records and other sources. We consider the case where the analyst has only a point estimate and associated standard error from each survey. This is common, as noted in Section 7 of the review paper, Lohr and Raghunathan (2017). In the motivating example, and in many other cases, there are not any covariates that can be used to improve inferences. Our methodology extends to cases where the inferential objectives and models are more complex.

With survey estimates, $\left\{\hat{Y}_i : i = 1, \ldots, L\right\}$, it is commonly assumed that the $\hat{Y}_i$ are independent

$$\hat{Y}_i \sim N\left(\mu_i, V_i\right) \tag{1.1}$$

1. Dexter Cahoy, University of Houston-Downtown; Joseph Sedransk, University of Maryland. E-mail: jxs123@case.edu.

where the V_i are assumed to be known.

A common prior distribution expressing similarity among $\{\mu_1, \ldots, \mu_L\}$ is

$$\mu_i \,|\, \nu, \delta^2 \sim N(\nu, \delta^2) \tag{1.2}$$

independently for each i, and ν and δ are assigned locally uniform prior distributions.

The resulting posterior expected value of μ_i is a convex combination of the estimate, $\hat{Y}_i$, for that survey, and a weighted average of $\{\hat{Y}_1, \ldots, \hat{Y}_L\}$. The weakness of this approach is that the prior distribution in (1.2) assumes independent sampling of the μ_i from a common distribution. The posterior mean is

$$E\left(\mu_i \,|\, \hat{Y}_1, \ldots, \hat{Y}_L\right) = \lambda_i \hat{Y}_i + (1 - \lambda_i)\hat{Y}_w \tag{1.3}$$

where $\lambda_i = \delta^2 / (\delta^2 + V_i)$ and $\hat{Y}_w = \sum_{i=1}^{L} \lambda_i \hat{Y}_i / \sum_{i=1}^{L} \lambda_i$. This may lead to unsatisfactory inferences when, e.g., $\mu_1, \ldots, \mu_b$ are each close to μ^* while $\mu_{b+1}, \ldots, \mu_L$ are each close to μ^{**} and $\mu^* << \mu^{**}$. Here, estimation of μ_1 would include, perhaps inappropriately, a large contribution from $\hat{Y}_{b+1}, \ldots, \hat{Y}_L$. The difficulty is that the prior distribution is not sufficiently flexible. Here we use more flexible prior distributions that permit the amount and nature of the pooling to be determined by the sample data.

The specification in (1.1) and (1.2) is common in meta-analyses and in situations where inferences for small subpopulations and geographical areas are desired. For example, the U.S. Census Bureau uses such models (augmented by terms to accommodate covariates) to make inference for U.S. county level poverty rates: see example 6.1.2 in Rao and Molina (2015). However, as just noted, the assumptions in (1.1) and (1.2) of full exchangeability may not be appropriate, especially for combining the information from L surveys.

The theory for our principal method, uncertain pooling, was developed by Malec and Sedransk (1992) and Evans and Sedransk (2001) with further work in Evans and Sedransk (1999) and Yan and Sedransk (2011). To our knowledge this methodology has not been used in a survey sampling application: the comprehensive review paper, Lohr and Raghunathan (2017), makes no reference to any technique similar to ours. We also modify (1.1) and (1.2) by using a Dirichlet process mixture (DPM) as, e.g., employed by Polettini (2017) for small area inference.

In this paper we describe the methodology for both uncertain pooling and DPM, and use them to analyze data from Ha and Sedransk (2019). Then we modified these data to exhibit properties of the methods. Finally, there is a simulation study to establish sampling properties. Clearly, the results from this evaluation will apply equally if the three sources were, e.g., a well established probability survey and two non-probability surveys or other choices.

We assume that the sample variances $\{V_1, \ldots, V_L\}$ are specified. In our context none of the alternatives given in the literature for making inferences for the $\{V_1, \ldots, V_L\}$, all based on inferences for small areas, is fully satisfactory. In Section 4 we discuss this challenging problem of making inferences for the sample variances.

Both uncertain pooling and DPM have a more general structure than that in the common specification, (1.1) and (1.2). This should lead to improved inferences. As seen in Section 2 the uncertain pooling model is the natural extension of (1.1) and (1.2); i.e., the model in (1.1) and (1.2) is a special case. Moreover, the output from uncertain pooling includes the posterior probabilities associated with the possible clustering of the L surveys.

Finally, note that we address only one of the many aspects of "combining survey data", well summarized by Lohr and Raghunathan (2017). Their Section 7, "Hierarchical models for combining data sources", gives additional examples where our methodology may be useful.

The methodology that we use is outlined in Section 2, and the results from our numerical studies are summarized in Section 3. A brief summary and discussion are in Section 4.

2. Methodology

As in Section 1 assume that there are L survey estimates, $\hat{Y}_1, \ldots, \hat{Y}_L$, with

$$\hat{Y}_i \overset{\text{ind}}{\sim} N\left(\mu_i, V_i\right) \tag{2.1}$$

where the V_i are assumed to be known.

2.1 Uncertain pooling

The uncertain pooling method is based on Malec and Sedransk (1992) and Evans and Sedransk (2001). They showed that a prior for $\mu = \left(\mu_1, \mu_2, \ldots, \mu_L\right)^t$ can be selected to reflect the beliefs that there are subsets of μ such that the μ_i in each subset are "similar", and that there is uncertainty about the composition of such subsets of μ. Let G be the total number of partitions of the set $\mathcal{L} = \{1, \ldots, L\}$, g be a particular partition $\left(g = 1, \ldots, G\right)$, $d\left(g\right)$ be the number of subsets of $\mathcal{L}$ in the g^{th} partition $\left(1 \le d\left(g\right) \le L\right)$, and $S_k\left(g\right)$ be the set of survey labels in subset $k\left(k = 1, \ldots, d\left(g\right)\right)$. For example, for $L = 3$, there are $G = 5$ partitions: $\{g = 1\} \sim \{(123)\}$, $\{g = 2\} \sim \{(13), (2)\}$, $\{g = 3\} \sim \{(12), (3)\}$, $\{g = 4\} \sim \{(23), (1)\}$, $\{g = 5\} \sim \{(1), (2), (3)\}$. Then, $S_1\left(g = 2\right) = \{(13)\}$, $S_2\left(g = 2\right) = \{(2)\}$, $d_1\left(g = 2\right) = 2$ and $d_2\left(g = 2\right) = 1$.

To specify a prior for μ, first condition on g. Malec and Sedransk (1992) and Evans and Sedransk (2001) assume that there is independence between subsets, and within $S_k\left(g\right)$ the μ_i are independent with

$$\mu_i \,\big|\, \nu_k\left(g\right) \sim N\left(\nu_k\left(g\right), \delta_k^2\left(g\right)\right), \qquad i \in S_k\left(g\right). \tag{2.2}$$

Also, the $\nu_k\left(g\right)$ are mutually independent with

$$\nu_k\left(g\right) \big|\, \theta_k\left(g\right) \sim N\left(\theta_k\left(g\right), \gamma_k^2\left(g\right)\right) \tag{2.3}$$

where the $\theta_k\left(g\right)$ and the $\gamma_k^2\left(g\right)$ are hyperparameters. The definition in (2.3) is the first step in obtaining a reference prior for the $\nu_k\left(g\right)$, i.e., one that is dominated by the likelihood. This will include letting the

$\gamma_k^2(g) \to \infty$, but is considerably more complicated, as described below. The $\delta_k^2(g)$ are also hyperparameters, to be assigned a prior distribution.

The formal definition in (2.3) is included as the first step in obtaining a reference prior for the $\nu_k(g)$, i.e., one that is dominated by the likelihood, as described below in the evaluation of $f(g, \Delta^2 \mid y)$. Conditioning on the $\delta_k^2(g)$ and $\gamma_k^2(g)$ (but suppressing them in our notation), and letting $\gamma_k^2(g) \to \infty$ leads to the following expected results for the posterior moments conditional on the partition g. As discussed below, additional care is needed to obtain the posterior distribution of g.

Defining $y = (\hat{Y}_1, \ldots, \hat{Y}_L)^t$, letting $\Delta^2 = \{\delta_k^2(g) : k = 1, \ldots, d(g); g = 1, \ldots, G\}$ and writing $\hat{\mu}_i = \hat{Y}_i$

$$E(\mu_i \mid y, g, \Delta^2) = \{\lambda_i(g)\}\, \hat{\mu}_i + \{1 - \lambda_i(g)\}\, \hat{\mu}_k(g), \qquad i \in S_k(g) \tag{2.4}$$

and

$$\operatorname{cov}(\mu_i, \mu_j \mid y, g, \Delta^2) = \begin{cases} \delta_k^2(g)\{1-\lambda_i(g)\} + \dfrac{\{1-\lambda_i(g)\}^2 \delta_k^2(g)}{\sum_{i \in S_k(g)} \lambda_i(g)}, & i = j; i, j \in S_k(g)) \\ \dfrac{\{1-\lambda_i(g)\}\{1-\lambda_j(g)\}\delta_k^2(g)}{\sum_{i \in S_k(g)} \lambda_i(g)}, & i \neq j; i, j \in S_k(g) \\ 0, & i \in S_{k_1}(g), j \in S_{k_2}(g), k_1 \neq k_2, \end{cases} \tag{2.5}$$

where

$$\lambda_i(g) = \frac{\delta_k^2(g)}{\delta_k^2(g) + V_i}, \hat{\mu}_k(g) = \frac{\sum_{j \in S_k(g)} \lambda_j(g) \hat{\mu}_j}{\sum_{j \in S_k(g)} \lambda_j(g)}. \tag{2.6}$$

Note that $E(\mu_i \mid y, g, \Delta^2)$ has the familiar form of a weighted average of $\hat{\mu}_i$ and $\hat{\mu}_k(g)$, but, here, $\hat{\mu}_k(g)$ is restricted to the surveys in $S_k(g)$.

Assuming the basic model in (1.1) and (1.2) corresponds, here, to the "pool-all" partition, $\{g = 1\}$, where all of the L surveys comprise a single cluster. Thus, for $\{g = 1\}$ the moments in (2.4), (2.5) and (2.6) are those that would be obtained by an analysis using (1.1) and (1.2). An analysis based on (1.1) and (1.2) is a special case of an analysis based on the uncertain pooling specification.

Inference about μ includes uncertainty about the value of g, i.e.,

$$f(\mu \mid y) = \int f(\mu \mid y, g, \Delta^2)\, f(g, \Delta^2 \mid y)\, dg d\Delta^2 \tag{2.7}$$

where the notation is simplified by using integration rather than summation for g. Using the "most likely" partition g^* (i.e., $p(g^* \mid y) \geq p(g \mid y) : g = 1, \ldots, G$) to make inference would understate the overall precision.

To evaluate (2.7) we need $f(g, \Delta^2 \mid y)$. However, when evaluating $f(g \mid \Delta^2, y)$ one must be careful about specifying the rate at which the $\gamma_k^2(g) \to \infty$: a natural choice leads to an expression for

$f\left(g\middle|\,\Delta^2, y\right)$ that is not invariant to changes in the scale of Y; see Section 4 of Malec and Sedransk (1992). Malec and Sedransk (1992) provided a solution by using an empirical Bayes argument. Here we use a fully Bayesian alternative, described in Section 5 of Evans and Sedransk (2001). It postulates little prior, relative to sample, information about the $\nu_k(g)$, and is invariant to changes in the scale of Y. Let $\nu(g) = \left(\nu_1(g),\ldots,\nu_{d(g)}(g)\right)^t$ and $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g)\middle|\, y\right)\right\}$ be the Kullback-Leibler information about $\nu(g)$. With prior $f\left(g, \Delta^2\right) = f(g)f\left(\Delta^2\right)$ and letting the $\gamma_k^2(g) \to \infty$ subject to $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g)\middle|\, y\right)\right\} = \text{constant}$,

$$f\left(g, \Delta^2\middle|\, y\right) \propto f\left(\Delta^2\right) f(g) \exp\left\{\frac{-d(g)}{2}\right\} \prod_{k=1}^{d(g)} \prod_{i \in S_k(g)} \left\{1-\lambda_i(g)\right\}^{1/2} \times \exp\left[-\frac{1}{2}\sum_{k=1}^{d(g)} \sum_{i \in S_k(g)} \left\{\frac{\lambda_i(g)}{\delta_k^2(g)}\right\} \left\{\hat{\mu}_i - \hat{\mu}_k(g)\right\}^2\right]. \tag{2.8}$$

The term in the exponent,

$$Q\left\{d(g)\right\} = \sum_{k=1}^{d(g)} \sum_{i \in S_k(g)} \left\{\frac{\lambda_i(g)}{\delta_k^2(g)}\right\} \left\{\hat{\mu}_i - \hat{\mu}_k(g)\right\}^2, \tag{2.9}$$

is likely to decrease as $d(g)$ increases, for example for a new partition of $\bigcup_{k=1}^{d(g)} S_k(g)$ obtained by creating subsets of the existing $\left\{S_k(g)\right\}$. Since $f\left(g, \Delta^2\middle|\, y\right)$ increases as $Q\left\{d(g)\right\}$ decreases, it is helpful to have the second term, $\exp\left\{-d(g)/2\right\}$, that penalizes partitions with larger values of $d(g)$.

For our analysis we take $\delta_k^2(g) = \delta^2$ and write $\lambda_i(g) = \delta^2 \big/ \left(\delta^2 + V_i\right)$. Inference for μ is made using (2.7) and (2.8) with

$$\mu\middle|\, y, g, \delta^2 \sim N\left(E\left(\mu\middle|\, y, g, \delta^2\right), V\left(\mu\middle|\, y, g, \delta^2\right)\right) \tag{2.10}$$

where the conditional posterior moments of μ are given in (2.4) and (2.5).

We assume that $f(g)$ is constant, i.e., that all partitions are equally likely, *a priori*, and take the Inverse Beta prior for δ^2, i.e.,

$$f\left(\delta^2\right) \propto 1 \big/ \left(1+\delta^2\right)\sqrt{\delta^2}, \quad 0 < \delta^2 < \infty. \tag{2.11}$$

Inference for μ is made using (2.7). To start, evaluate the right side of (2.8) for

$$\left\{g : g = 1,\ldots,G; \quad R \text{ grid points for } \delta^2\right\}, \tag{2.12}$$

then standardize by dividing the individual terms in the grid by their sum. This provides an approximation for $f\left(g, \delta^2\middle|\, y\right)$. Then select a random sample of size B from the RG normalized values of $f\left(g, \delta^2\middle|\, y\right)$. For each selection, $\left(g_*, \delta_*^2\right)$, sample μ from $f\left(\mu\middle|\, y, g_*, \delta_*^2\right)$. Here, we generated $B =$ 5,000 values of μ. Finally, note that approximations for the marginal posterior distributions, i.e., $f\left(g\middle|\, y\right)$ and $f\left(\delta^2\middle|\, y\right)$, can be obtained directly from the grid approximation of $f\left(g, \delta^2\middle|\, y\right)$.

Assuming that survey r is the single best choice for inference we can consider the posterior distribution corresponding to survey r to be the object of inference. In a common contemporary application there will be data from a well established probability survey (the single best choice) and data from other sources such as non-probability surveys, administrative records, etc. In other settings it is likely that there will be a preference for one of the surveys.

Then, using the posterior expected value for illustration,

$$E\left(\mu_r \mid y\right) = E_{g,\delta^2 \mid y} E\left(\mu_r \mid y, g, \delta^2\right) \tag{2.13}$$

where $E\left(\mu_r \mid y, g, \delta^2\right)$ is defined in (2.4). Thus, inference for μ_r is a function of $\hat{\mu}_r$ together with data from the other $L-1$ studies as determined by the form of (2.4), and, critically, by the likelihood associated with the set of subsets, $S_k(g)$, containing study r. See Evans and Sedransk (2001) for additional details and an application to a notable study of the effect of using aspirin by patients following a myocardial infarction.

The model given by Chakraborty, Datta and Mandal (2014) has a superficial resemblance to the one in (2.1), (2.2), and (2.3). Taking x_i to be the scalar with value 1, the model in (2.1) of Chakraborty et al. (2014) is

$$\hat{Y}_i = \mu_i + e_i, \quad i = 1, \ldots, L \tag{2.14}$$

where $\mu_i = \xi + \left(1-\delta_i\right)\nu_{1i} + \delta_i \nu_{2i} + e_i$ with $e_i, \delta_i, \nu_{1i}, \nu_{2i}$ independent, $p\left(\delta_i = 1 \mid p\right) = 1 - p, \nu_{1i} \sim N\left(0, A_1\right)$ and $\nu_{2i} \sim N\left(0, A_2\right)$. Finally, $e_i \sim N\left(0, V_i\right)$ with V_i known. Thus, unlike the uncertain pooling method, there is only a single focal point, ξ. This permits appropriate treatment of outliers, but does not take advantage of possible clustering of the μ_i. This can also be seen in (2.4) of Chakraborty et al. (2014) where

$$E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) = \hat{Y}_i - \kappa_i\left(\hat{Y}_i - \xi\right) \tag{2.15}$$

and κ_i is a function of $V_i, A_1, A_2, p\left(\delta_i = 0 \mid \xi, A_1, A_2, p, y\right)$ with $y = \left(\hat{Y}_1, \ldots, \hat{Y}_L\right)^t$. Chakraborty et al. (2014) show that if survey i is an outlier, $E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) \approx \hat{Y}_i$, as desired. Now suppose that surveys $\{1, \ldots, b\}$ and $\{b+1, \ldots, L\}$ form two distinct clusters with a very large separation between them. Then inference for μ_1, say, will not, in general, use the data in an appropriate manner. In (2.15) there should be two values of ξ, i.e., corresponding to the two subsets. And appropriate use of information about subsets is the essence of the uncertain pooling method.

2.2 Dirichlet process mixture

An alternative to the uncertain pooling method is to use a Dirichlet process mixture (DPM), one of the most popular nonparametric Bayesian methods. This methodology is presented in detail in Sections 2.1 and 2.2 of Muller, Quintana, Jara and Hanson (2015). For our analyses we have used the R function

DPmeta from the package DPpackage: see Jara, Hanson, Quintana, Muller and Rosner (2011) for details. The model in DPmeta is

$$y_i \mid \theta_i \overset{\text{iid}}{\sim} f_{\theta_i} \tag{2.16}$$

and

$$\theta_i \mid H \overset{\text{iid}}{\sim} H \tag{2.17}$$

with $H \sim \text{DP}(M, H_0)$.

In (2.16) and (2.17) $y_i = \hat{Y}_i, \theta_i = \mu_i, f_{\theta_i}$ is the pdf of a $N(\mu_i, V_i)$ random variable with V_i fixed, and $H_0 = N(\eta, \tau^2)$.

The (independent) hyperparameters are

$$M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$$

$$\eta \mid \eta_b, S_b \sim N(\eta_b, S_b)$$

$$\tau^{-2} \mid \phi_1, \phi_2 \sim \text{Gamma}(\phi_1/2, \phi_2/2). \tag{2.18}$$

Polettini (2017) has proposed using a DPM of this nature for inference about small area parameters. As in Section 2.1 of our paper Polettini (2017) indicates the value of extending the typical random effects model (e.g., the well known Fay-Herriot model) that assumes full exchangeability of the set of small area parameters.

The uncertain pooling method requires only that one specify a prior distribution for g and δ^2. By contrast DPmeta requires substantial prior input, i.e., values for $a_0, b_0, \eta_b, S_b, \phi_1$ and ϕ_2. Without strong prior information we can't make proper inferences for these quantities with only $L = 3$ surveys. So, we have omitted the specification $M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$ and made inference for a selected set of values of M as suggested by Escobar (1994). Also, we replaced η_b, S_b, ϕ_1 and ϕ_2 with their maximum *a posteriori* probability estimates.

3. Results

Ha and Sedransk (2019) made inference for the proportion of adults without health insurance in each of the 67 Florida counties, and compared these estimates with those from two other sources. Some of the differences were striking, motivating us to consider methodology to make appropriate inferences in such cases. We use these data, and modifications of these data, to show the benefits of using the methodology outlined in Section 2. One source is the Small Area Health Insurance Estimates Program (SAHIE, hereafter survey 1). The SAHIE program uses point estimates from the American Community Survey (ACS) together with administrative data such as Federal income tax returns and Medicaid/Children's

Health Insurance Program (CHIP) participation rates. There is detailed area level modelling. The principal ones are models of ACS estimates of the proportions in income groups and the proportions insured. There are additional models such as ones modelling the number of persons enrolled in Medicaid or CHIP, Supplementary Nutrition Assistance Program (SNAP) participation, and Internal Revenue Service (IRS) tax exemptions. For a full understanding of this program see the twenty-two page technical report, Bauder, Luery and Szelepka (2018). We have added a non-technical summary in the Appendix.

The analyses using both survey 2, denoted by HS, based on Ha and Sedransk (2019), and survey 3 denoted by CDC (Centers for Disease Control and Prevention) use unit level models based on 2010 data from the Behavioral Risk Factor Surveillance System (BRFSS), obtained through telephone interviews. While the sample designs differ somewhat over states, the one in Florida was typical, i.e., a disproportionate stratified sample design. In Florida, the set of telephone numbers were divided into two strata (high and medium density) that were sampled separately. In addition there was a stratification by area codes, i.e., three geographic strata and a fourth stratum consisting of area codes with large estimated Hispanic populations. For additional, general information see http://www.cdc.gov/brfss/annual_data/annual_2010.htm while for technical details see Pierannunzi, Xu, Wallace, Garvin, Greenlund, Bartoli, Ford, Eke and Town (2016) and Ha and Sedransk (2019). Both use, essentially, the same covariates but the modeling in HS is more detailed. Moreover, Pierannunzi et al. (2016) give only point estimates, noting that standard errors were being developed. A further complication is that the CDC analysis is frequentist while the SAHIE and HS analyses are Bayesian. Thus, we have an (empirical) Bayes posterior standard deviation for SAHIE, none for CDC and an estimated SE for HS obtained by taking the 95% credible interval for a county proportion and dividing by 3.92. While the limitations just noted preclude definitive conclusions from these data they illustrate the methodology. Moreover, conditions where standard errors are missing or unreliable are, at least, fairly common for non-probability samples, a focus of this paper.

The first set of analyses is based on the observed data. To show other properties of the methodology a second set of analyses is based on modifications of these data. Finally, to show sampling properties there is a simulation study. Note that each of our analyses is based *only* on data from a single county. Additional research is needed to permit inference using data from all of the sources and counties. See Section 4 for discussion.

3.1 Data-based analyses

Using the uncertain pooling methodology a summary of the results for Dixie county is presented in Table 3.1. These are typical of most of the county-based analyses that we have done. Throughout, SE denotes the sample standard error. There are three panels, corresponding to choices of the CDC SE, taken equal to 0.5, 1.0, 2.0 times the HS SE. For each panel the column headings are the observed proportion, posterior mean of the county proportion, estimated SE of the observed proportion, posterior standard deviation and lower and upper bounds of the 95% credible interval for the county proportion. At the bottom of each panel there are the values of $p(g \mid y)$ with $p(g \mid y) \geq 0.001$ where $\{g = 1\} \sim \{(123)\}$,

$\{g=2\}\sim\{(13),(2)\}$, $\{g=3\}\sim\{(12),(3)\}$, $\{g=4\}\sim\{(2,3),(1)\}$, $\{g=5\}\sim\{(1),(2),(3)\}$, and summaries corresponding to $\{g=1\}$, labelled "pool-all".

We first analyze these data using the uncertain pooling methodology, then compare them with those from DPmeta.

A common way to summarize a set of sample proportions is to assume that the corresponding set of true proportions come from a common source, i.e., $p(g=1)=1$. However, for each of the three cases in Table 3.1, $p(g=1\mid y)\le 0.001$. Thus, there is very little support for pooling all of the data from the three surveys. For further investigation of the effect of assuming a common source, assume $g=1$. Then, as in (1.1) and (1.2),

$$\hat{Y}_i \overset{\text{ind}}{\sim} N(\mu_i, V_i)$$

$$\mu_i \overset{\text{iid}}{\sim} N(\nu, \delta^2), \quad i=1,\ldots,L. \tag{3.1}$$

With a locally uniform prior on ν and the Inverse Beta prior on δ^2 in (2.11)

$$f(\nu\mid y)=\int f(\nu\mid \delta^2, y) f(\delta^2\mid y)\, d\delta^2 \tag{3.2}$$

where the posterior distribution of ν given δ^2 is normal with $E(\nu\mid \delta^2, y)=\frac{\sum_{i=1}^{L}\hat{Y}_i/(V_i+\delta^2)}{\sum_{i=1}^{L}1/(V_i+\delta^2)}$ and $\text{Var}(\nu\mid \delta^2, y)=\left(\sum_{i=1}^{L}(V_i+\delta^2)^{-1}\right)^{-1}$.

For panel 1 in Table 3.1, $E(\nu\mid y)=0.313$, $\text{SD}(\nu\mid y)=0.017$ and the 95% credible interval is (0.290, 0.340). Inferences based on the posterior distribution of ν are not consistent with the notion that any one of the three surveys is the nominal "gold standard". If, for example, survey 1 is taken as the "gold standard", the posterior mean of μ_1, 0.254, is substantially smaller than the posterior mean of ν, 0.313. Moreover, 0.254 is not included in the 95% interval for ν, (0.290, 0.340). The conclusions from panels 2 and 3 are essentially the same. Finally, recall that $p(g=1\mid y)\le 0.001$, indicating very little support for pooling all of the data.

In the following assume, for illustration, that one prefers the HS methodology. Then there may be substantial gains in precision (measured by the posterior standard deviation) by using the uncertain pooling methodology. The gain in precision is measured by comparing the posterior standard deviation from the uncertain pooling methodology with that obtained by using only the data from the specific survey, here HS (survey 2). For the latter and a locally uniform prior for μ_2, the posterior distribution of μ_2 is normal with posterior mean equal to the observed proportion and posterior standard deviation equal to the estimated SE. If we take CDC SE $= k$ (HS SE) for $k=0.5, 1, 2$, then the reductions in the posterior standard deviation for HS (survey 2), and corresponding to $k=0.5, 1, 2$, are 29, 18 and 7%. (For example, from panel 1 of Table 3.1, i.e., $k=0.5$, the percent reduction in the posterior SD for HS is $100\,(0.028-0.020)\,/\,0.028\%=29\%$.) Note that the relatively small SEs for each of the surveys means that the "all singletons" partition, i.e., $\{g=5\}$, has a relatively large posterior probability (about 0.38).

The corresponding reductions in the posterior standard deviation (uncertain pooling vs. no pooling) for CDC (survey 3) are 7, 18 and 14%.

Table 3.1
Observed proportions, standard errors and posterior summaries from Dixie County, Florida using uncertain pooling

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Int
CDC SE = 0.5 × HS SE	1	0.254	0.254	0.014	0.014	(0.225, 0.283)
	2	0.361	0.360	0.028	0.020	(0.317, 0.403)
	3	0.359	0.359	0.014	0.013	(0.333, 0.385)
	pool-all		0.313		0.017	(0.290, 0.340)
	$P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.621;\ P(g=5\mid y)=0.377.$					
CDC SE = HS SE	1	0.254	0.254	0.014	0.014	(0.225, 0.283)
	2	0.361	0.360	0.028	0.023	(0.313, 0.406)
	3	0.359	0.359	0.028	0.023	(0.312, 0.404)
	pool-all		0.290		0.011	(0.268, 0.312)
	$P(g=2\mid y)=0.002;\ P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.619;\ P(g=5\mid y)=0.376.$					
CDC SE = 2 × HS SE	1	0.254	0.254	0.014	0.014	(0.226, 0.284)
	2	0.361	0.360	0.028	0.026	(0.307, 0.412)
	3	0.359	0.349	0.056	0.048	(0.256, 0.303)
	pool-all		0.279		0.012	(0.255, 0.305)
	$P(g=1\mid y)=0.001;\ P(g=2\mid y)=0.107;\ P(g=3\mid y)=0.002;\ P(g=4\mid y)=0.554;\ P(g=5\mid y)=0.336.$					

Note: Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

As noted in Section 2, a complete specification of DPmeta requires specifying the values of many hyperparameters, and we have no prior information to make informed choices. So, we have replaced η_b, S_b, ϕ_1 and ϕ_2 with their maximum *a posteriori* probability (MAP) estimates. We have followed Escobar (1994) by considering $M \in \{L^{-1}, L^0, L^1, L^2\} = \{1/3, 1, 3, 9\}$.

From (2.10) in Muller et al. (2015) the prior probability of k clusters is a function of M. Let $p_M = (p_{1M}, p_{2M}, p_{3M})$ where p_{kM} is the *prior* probability of k clusters with precision M. Then p_{kM} can be calculated using the probability associated with any partition, i.e.,

$$\frac{M^{k-1}\prod_{j=1}^{k}\Gamma(L_j)}{(M+1)(M+2)\cdots(M+L-1)} \tag{3.3}$$

where L_j is the number of surveys in cluster j with $\sum_{j=1}^{k} L_j = L$. Then $p_{1/3} = (18/28, 9/28, 1/28)$, $p_1 = (2/6, 3/6, 1/6)$, $p_3 = (2/20, 9/20, 9/20)$ and $p_9 = (2/110, 27/110, 81/110)$. Since $p_{1/3}$ and p_9 are too extreme we have emphasized $M = 1$ and $M = 3$. The results corresponding to $M = 1$ and $M = 3$ are very close, so only the latter are presented in Table 3.2, which has the same format as Table 3.1.

Comparing the results from the uncertain pooling method with those from DPmeta it is apparent that, in general, there is close agreement. For the posterior mean they are similar except in panel 3 where there is greater shrinkage for surveys 2 and 3. The results for the posterior SD are also close except for a larger

value for survey 2 in panel 3. There are only small differences in the intervals except for panel 3 where the DPmeta intervals for surveys 2 and 3 are wider.

Table 3.2
Observed proportions, standard errors and posterior summaries from Dixie County, Florida using DPmeta

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Interval
CDC SE = 0.5 × HS SE	1	0.254	0.254	0.014	0.014	(0.227, 0.282)
	2	0.361	0.359	0.028	0.013	(0.334, 0.384)
	3	0.359	0.360	0.014	0.012	(0.335, 0.384)
CDC SE = HS SE	1	0.254	0.256	0.014	0.016	(0.227, 0.290)
	2	0.361	0.357	0.028	0.024	(0.291, 0.399)
	3	0.359	0.357	0.028	0.025	(0.290, 0.399)
CDC SE = 2 × HS SE	1	0.254	0.264	0.014	0.018	(0.230, 0.287)
	2	0.361	0.332	0.028	0.044	(0.261, 0.406)
	3	0.359	0.321	0.056	0.048	(0.249, 0.402)

Note: Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

The small value of the SAHIE SE seen in almost all counties limits the scope of our evaluation. Thus, we have used modified data sets based on the original data. Here, as before, we take the CDC SE to be 0.5, 1 and 2 times the HS SE, but also take the SAHIE SE to be 2, 5 and 10 times the HS SE. Table 3.3, with the same format as Table 3.1, shows, for uncertain pooling, the results for Orange county with the CDC SE = 0.5(HS SE). These results are typical of our analyses for a sizeable number of FL counties. Recall, though, that each analysis is based only on the data from the specific county. For panel 1 in Table 3.3, $E\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0.199$, $\mathrm{SD}\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0.008$ and the 95% credible interval is (0.184, 0.215). Inferences based on the posterior distribution of ν are inappropriate if one regards any one of the three surveys as the "gold standard". For example, the posterior mean of μ_1, 0.278, is outside the 95% credible interval for ν. As in Table 3.1, $p\left(g = 1 \middle| y\right) \leq 0.001$, indicating very little support for pooling all of the data.

The percent reductions in the posterior standard deviation of μ_1, i.e., for SAHIE (survey 1), are 11, 26 and 44%, corresponding to the three panels in Table 3.3. As the value of the SAHIE SE is increased, there is, as expected, additional pooling of the SAHIE observed proportion with the CDC observed proportion. Noting that the observed proportion for SAHIE is 0.294, the posterior means for μ_1 decrease from 0.278 (panel 1) to 0.240 (panel 3). One reason for this can be seen by comparing the posterior distributions of g, i.e., $\left\{g, p\left(g \middle| y\right) : g = 1, \ldots, 5\right\}$, given at the bottom of each panel. For example, $p\left(g = 2 \middle| y\right)$ increases from 0.006 to 0.339 while $p\left(g = 5 \middle| y\right)$ decreases from 0.479 to 0.253. These results show that the uncertain pooling methodology is taking proper account of the increased variability associated with the SAHIE estimates, i.e., increasing the likelihood of pooling the data from surveys 1 and 3.

Comparing the results from the uncertain pooling method in Table 3.3 with those from DPmeta in Table 3.4 it is apparent that there are greater differences than those seen in Tables 3.1 and 3.2. For the posterior means it is notable that for survey 1 the posterior mean from DPmeta is somewhat smaller than that from uncertain pooling. This reflects greater pooling of the data from survey 1 with that from survey

3. For surveys 2 and 3 the two sets of posterior means are similar. The most notable difference is for the posterior SDs where, for survey 1 (panels 2 and 3) the values are very much smaller from DPmeta than from uncertain pooling. For the other seven cases, the two sets of posterior SDs are similar. Correspondingly, for survey 1 the intervals from DPmeta are much shorter than those from uncertain pooling while those for surveys 2 and 3 are only a little wider. From Table 3.4 and for survey 1 the percent reductions in the posterior SD (relative to the ObsSEs) are (42%, 55%, 75%). There are increases, though, for survey 2 in panels 2 and 3.

Table 3.3
Observed proportions, standard deviations and posterior summaries from Orange County, Florida, where CDC SE = 0.5 × HS SE using uncertain pooling

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Int
SAHIE SE = 2 × HS SE	1	0.294	0.278	0.036	0.032	(0.227, 0.352)
	2	0.257	0.261	0.018	0.017	(0.226, 0.294)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.199		0.008	(0.184, 0.215)
	$P(g=2\mid y)=$ 0.006; $P(g=3\mid y)=$ 0.514; $P(g=5\mid y)=$ 0.479.					
SAHIE SE = 5 × HS SE	1	0.294	0.251	0.089	0.066	(0.162, 0.417)
	2	0.257	0.258	0.018	0.018	(0.223, 0.293)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.195		0.008	(0.180, 0.211)
	$P(g=2\mid y)=$ 0.224; $P(g=3\mid y)=$ 0.468; $P(g=5\mid y)=$ 0.308.					
SAHIE SE = 10 × HS SE	1	0.294	0.240	0.179	0.101	(0.059, 0.520)
	2	0.257	0.257	0.018	0.018	(0.222, 0.292)
	3	0.179	0.179	0.009	0.009	(0.162, 0.197)
	pool-all		0.195		0.008	(0.179, 0.211)
	$P(g=2\mid y)=$ 0.339; $P(g=3\mid y)=$ 0.408; $P(g=5\mid y)=$ 0.253.					

Note: Small Area Health Insurance Estimates (SAHIE); Centers for Disease Control and Prevention (CDC); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

Table 3.4
Observed proportions, standard errors and posterior summaries from Orange County, Florida using DPmeta

	Survey	ObsProp	PostMean	ObsSE	PostSD	95% Cred Interval
SAHIE SE = 2 × HS SE	1	0.294	0.262	0.036	0.021	(0.202, 0.297)
	2	0.257	0.263	0.018	0.018	(0.222, 0.296)
	3	0.179	0.180	0.009	0.009	(0.162, 0.199)
SAHIE SE = 5 × HS SE	1	0.294	0.226	0.089	0.040	(0.168, 0.290)
	2	0.257	0.246	0.018	0.023	(0.186, 0.291)
	3	0.179	0.182	0.009	0.011	(0.163, 0.205)
SAHIE SE = 10 × HS SE	1	0.294	0.217	0.179	0.044	(0.166, 0.288)
	2	0.257	0.243	0.018	0.031	(0.185, 0.290)
	3	0.179	0.183	0.009	0.011	(0.162, 0.205)

Note: Small Area Health Insurance Estimates (SAHIE); Standard error (SE); Ha and Sedransk (HS); Standard deviation (SD).

3.2 Results from simulation study

To evaluate properties such as bias and coverage of the credible interval we have carried out a simulation study based on several modifications of the Orange county data. Specifically, we generate $\{\hat{Y}_i : i = 1, 2, 3\}$ from

$$\hat{Y}_1 \sim N(\psi_1, V_1)$$

$$\hat{Y}_2 \sim N(\psi_1, V_2)$$

$$\hat{Y}_3 \sim N(\psi_2 + \Delta, V_2) \tag{3.4}$$

where ψ_1 and ψ_2 are from the third panel of Table 3.3; ψ_1 is the average of the observed proportions from surveys 1 and 2 while ψ_2 is the observed proportion from survey 3 (CDC). Also, we took V_1 to be much larger than V_2. These choices were made to represent a common situation where survey 1 is a probability sample, with relatively large sample variance while surveys 2 and 3 are non-probability samples with much smaller sample variances. Finally, $\Delta \in \{0, 4(0.0193) = 0.0772, 8(0.0193) = 0.1544\}$.

Table 3.5 gives the values of ψ_1, ψ_2, V_1 and V_2 in the footnote. There are three rows, corresponding to $\Delta = 0, 0.0772, 0.1544$. In each row there are the medians over 500 replications of $\{p(g \mid y) : g = 1, \ldots, 5\}, \{E(\mu_i \mid y) : i = 1, 2, 3\}$ and $\{\text{SD}(\mu_i \mid y) : i = 1, 2, 3\}$, together with the estimated coverages.

Table 3.5
Simulation results from 500 replications of (3.4)

Inc Size	$P(g \mid \text{data})$					Coverage			PostMean			PostSD		
	g : 1	2	3	4	5	i : 1	2	3	i : 1	2	3	i : 1	2	3
$\Delta = 0$	0	0.148	0.462	0	0.332	0.973	0.958	0.960	0.262	0.275	0.179	0.050	0.006	0.006
$\Delta = 0.0772$	0.032	0.292	0.303	0.033	0.249	0.984	0.941	0.939	0.269	0.275	0.257	0.039	0.006	0.006
$\Delta = 0.1544$	0	0.280	0.401	0	0.300	0.971	0.958	0.952	0.292	0.275	0.333	0.044	0.006	0.006

$\psi_1 = 0.276$, $\psi_2 = 0.179$, $V_1 = 0.06^2$, $V_2 = 0.006^2$.

Note: Standard deviation (SD).

The principal findings are: (a) the medians of the posterior means are close to the values used to generate the data, i.e., ψ_1 and ψ_2, (b) the coverages are close to the nominal 95%, and (c) there are significant reductions in the posterior standard deviation for survey 1 (SAHIE), i.e., 16.7%, 35.0% and 26.7% corresponding to $\Delta = 0, 0.0772$ and 0.1544. There are no reductions in the posterior standard deviations for surveys 2 and 3.

For $\Delta = 0.1544$ note that $p(g = 2 \mid y) = 0.280$ while $p(g = 4 \mid y) = 0$. That is, we pool data from surveys 1(SAHIE) and 3(CDC), $\{g = 2\}$, because of the relatively large SE for survey 1(SAHIE). However, we do not pool data from surveys 2(HS) and 3(CDC), $\{g = 4\}$, because of the relatively small SEs for survey 2(HS) and survey 3 (CDC). Of course, we pool data from surveys 1 and 2, $\{g = 3\}$, because they have the same mean, ψ_1.

4. Discussion and summary

With reduced response rates and diminished resources there is considerable interest in combining data from several sources such as sample surveys and administrative data. Currently there is special interest

when the sources include non-probability surveys. Appropriate methodology is required to ensure satisfactory inferences since the target populations and data acquisition methods may be quite different.

There are many situations where it may be beneficial to combine such data, as shown in the review paper by Lohr and Raghunathan (2017). Here, we have investigated the case where the analyst has only summary statistics from each of the sources, and where one can think of one source, r, as the single best source for inference. While it is often beneficial to use the data from related sources to improve inferences from r, it is essential that the data that are combined be concordant with the data from r. The methodology in this paper can also be used in settings where the data are not limited to summary statistics and inferential objectives and models are more complex. As seen in this paper, failure to consider biases due to pooling "unlike" data may lead to poor inference. Using analytical expressions and examples we have shown that both the uncertain pooling and DPM methods provide appropriate inferences. However, our analyses based on uncertain pooling are fully Bayes while those from DPmeta are empirical Bayes – due to the need to specify values for many hyperparameters. Moreover, the uncertain pooling method provides additional information in the form of the posterior probabilities for the partitions, g.

The methods can be implemented. For DPmeta there is an R package, DPpackage (Jara et al., 2011) while an R package is being developed for the uncertain pooling method. When completed it will be submitted to The Comprehensive R Archive Network. It contains functions that allow Bayesian analyses of the type described in this paper (a) with user-supplied point estimates and associated variances, or (b) with binomial data, cases (y) and total counts (n). Case (b) provides an analysis based on the logit transformation of the sample proportion. We have implemented the latter when there are eleven surveys.

Making inference for the sample variances, $V_1, \ldots, V_L$ is a very challenging problem. Polettini (2017) provides an extensive discussion of methods that have been proposed. Of particular interest are solutions proposed by You and Chapman (2006), Sugasawa, Tamae and Kubokawa (2017) and Polettini (2017). However, these solutions are posed in the context of small area inference, not when the objective is combining data from surveys and related sources.

While the discussion below is in the context of extending the DPM method (Section 2.2), the ideas are also relevant for the uncertain pooling method (Section 2.1). Polettini (2017) augments the DPM model in Section 2.2 with

$$\delta_i S_i^2 \overset{\text{ind}}{\sim} V_i \chi^2_{\delta_i}, \quad i = 1, \ldots, L \tag{4.1}$$

and

$$V_i^{-1} \overset{\text{iid}}{\sim} \text{Gamma}(a_1, b_1) \tag{4.2}$$

where S_i^2 is the sampling variance and δ_i is a measure of the degrees of freedom.

As noted by Polettini (2017) the assumption of the χ^2 distribution in (4.1) is questionable, surely so when there is a complex survey design. With only a single sample one cannot verify the *sampling distribution* of S_i^2, a point also made by Polettini (2017) on page 731. Moreover, in survey sampling the

form of S_i^2 is likely to be a complex function of the values of the variable of interest, Y, and survey weights. Thus, it is unlikely that the distribution of the observed Y's can be used to infer a reasonable approximation for the distribution of S_i^2.

The assumption of constant population parameters in (4.2) is problematic for our case, i.e., combining data. We expect considerable differences among the surveys, e.g., for a collection of probability and non-probability samples. You and Chapman (2006) generalize by replacing (4.2) with

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\right). \tag{4.3}$$

This requires values for $\left(a_i, b_i\right)$, a difficult choice without prior information. Moreover, Gelman (2006) shows that selecting both a_i and b_i very small, a natural choice (and one made by You and Chapman (2006)), may lead to poor inferences. Sugasawa et al. (2017) provide an alternative to You and Chapman (2006) by assuming

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\gamma\right), \tag{4.4}$$

with a prior on γ, but this, too, requires specifying values for a_i and b_i. Clearly, making better inference for the sample variances is an important topic for future research.

There has been an increased interest in making inference for small subpopulations, i.e., "small area" inference, when there are several data sources; see, e.g., Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011) and Nandram, Berg and Barboza (2014). While further research is needed to extend the uncertain pooling methodology to this case the approach is clear. Let j denote a small area, e.g., a US county, and i denote a data source where $j=1,\ldots,J$ and $i=1,\ldots,L$. As above g denotes a generic partition with generic subset $S_k\left(g\right)$ for $k=1,\ldots,d\left(g\right)$. Define $\mathcal{G}=\left\{\left(ij\right): j=1,\ldots,J; i=1,\ldots,L\right\}$. Then for fixed g, $S_k\left(g\right)$ is a subset of $\mathcal{G}$ with $S_k\left(g\right)\bigcap S_m\left(g\right)=\varnothing$ for $k\neq m$ and $\bigcup_{k=1}^{d(g)} S_k\left(g\right)=\mathcal{G}$. For example, let $J=2$ and $L=3$. Then each partition is a collection of the disjoint subsets of $\mathcal{G}=\left\{(11),(12),(21),(22),(31),(32)\right\}$ whose union is $\mathcal{G}$. By analogy with the discussion in Section 2, there would be a single best source for each small area, identified as $\left(j, i\left(j\right)\right)$ for some i in small area j.

Then the following model, analogous to the one in Section 2, is

$$\hat{Y}_{ij} \overset{\text{ind}}{\sim} N\left(\mu_{ij}, V_{ij}\right). \tag{4.5}$$

By analogy with (2.2)

$$\mu_{ij} \overset{\text{ind}}{\sim} N\left(\nu_k\left(g\right), \delta_k^2\left(g\right)\right), \qquad ij\in S_k\left(g\right). \tag{4.6}$$

If the same (limit) assumptions are made about the $\gamma_k^2\left(g\right)$ and $\delta_k^2\left(g\right)=\delta^2$, the expressions for posterior inference for the μ_{ij} will be the same as in Section 2. However, the assumption of constant δ^2 may not be reasonable. Since there will be a very large number of partitions computation will be challenging, especially since it is expected that many $p\left(g\,|\,y\right)$ will be very small.

The premise of our work is that one should include the possibility that the parameters associated with different surveys may not be exchangeable. (With a probability sample and several non-probability

samples this may be especially important.) Similarly, it is natural to generalize so that the parameters associated with the small areas are not assumed to be exchangeable. However, if exchangeability across both surveys and small areas can be assumed, the model in Section 2.1 of Kim, Park and Kim (2015) (possibly modified to accommodate Bayesian inference) should be easier to implement.

Although the very large number of partitions of $\mathcal{G}$ may pose an obstacle to implementation, one may be able to apply DPmeta when there are data from a set of small areas and several data sources. One problem is the specification in DPmeta of a common distribution for the μ_{ij}, i.e., over small areas and surveys, which is unlikely to be appropriate. The possibilities include an ANOVA model (Section 4.4.2) or nested model (Section 7.3.1) of Muller et al. (2015), although the ANOVA model has no interaction terms and our model is a cross-classified one.

Future research should include making inference for the sample variances, as noted above. Also, we need improved methodology to handle the extension to small area inference when there are data from several surveys. In some cases one may be able to simplify the model for the μ_{ij}. Using a grid-based method for sampling g and δ^2 is difficult to implement when G is extremely large. So, using a standard MCMC approach, possibly with an informative prior on g, may be a better way to make inference. For example, see Dahl, Day and Tsai (2017).

Other approaches could also be explored. For example, Park, Kim and Stukel (2017) suggest a different approach for combining data from two surveys. Here, there are covariates, X, observed in each survey while Y_1, the study variable of interest, is observed only in survey 1 and Y_2 is observed only in survey 2. Inference for the population mean of Y_1 is desired, given data from both surveys. The densities that they use are $f_1\left(Y_1 \mid X, \theta_1\right), f_2\left(Y_2 \mid X, Y_1, \theta_2\right)$ and, for identifiability, it is assumed that $f_2\left(Y_2 \mid X, Y_1\right) = f_2\left(Y_2 \mid Y_1\right)$. For a Bayesian analysis an extension to more than two surveys would be needed, together with specification of appropriate prior distributions for the parameters. It does not seem to be straightforward to model the distribution of Y_2 given Y_1.

Acknowledgements

The authors are grateful to the reviewers whose extensive comments have resulted in a paper that has greater focus and increased breadth. They also appreciate research allocation grants from ACCESS's Pittsburgh Supercomputing Center.

Appendix

Small Area Health Insurance Estimates (SAHIE) Program

The following summary paraphrases relevant parts of US Census Bureau (2021). To avoid distortion of the authors' meaning we have retained the first-person text.

The SAHIE program produces model-based estimates of health insurance coverage for demographic groups within counties and states. We publish county estimates by sex, age and income. The income

groups are defined by the income-to-poverty ratio (IPR) – the ratio of family income to the appropriate federal poverty level.

For estimation, SAHIE uses models that combine survey data from the American Community Survey (ACS) with administrative records data and Census data. The models are "area-level" models because we use survey estimates and administrative data at certain levels of aggregation, rather than individual survey and administrative records. Our modeling approach is similar to that of common models developed for small area estimation, but with additional complexities.

The published estimates are based on aggregates of modeled demographic groups. For counties, we model at a base level defined by age, sex and income groups.

We use estimates from the Census Bureau's Population Estimates Program for the population in groups defined for county by age and sex. We treat these populations as known. Within each of these groups, the number with health insurance coverage in any of the income categories is given by that population multiplied by two unknown proportions to be estimated: the proportion in the income category and the proportion insured within that income category. The models have two largely distinct parts – an "income part" and an "insurance part" – that correspond to these proportions. We use survey estimates of the proportions in the income groups and of the proportions insured within those groups. We assume these survey estimates are unbiased and follow known distributions. We also assume functional forms for the variances of the survey estimates that involve parameters that are estimated. We treat supplemental variables that predict one or both of unknown income and insurance proportions in one of two ways:

Some of these variables are used as fixed predictors in a regression model. There is a regression component in both the income and insurance parts of the model. In each case, a transformation of the proportion is predicted by a linear combination of fixed predictors. Some of these predictors are categorical variables that define the demographic groups we model. Others are continuous. The continuous fixed predictors include variables regarding employment, educational attainment, and demographic population.

We also utilize random continuous predictors, which include data from 5-year ACS, Internal Revenue Service, Supplemental Nutrition Assistance Program, and Medicaid/Children's Health Insurance Program. These are not fixed predictors in the model. Instead, we treat them as random, in a way similar to survey estimates, but not as unbiased estimators of the numbers. Instead, we assume that their expectations are linear functions of the number in an income group or the number insured within an income group. We typically assume they are normally distributed with variances that depend on unknown parameters.

We formulate the model in a Bayesian framework and report the posterior means as the point estimates. We use the posterior means and variances together with a normal approximation to calculate symmetric 90-percent confidence intervals, and report their half-widths as the margins of error.

We control the estimates to be consistent with specified national totals.

References

Bauder, M., Luery, D. and Szelepka, S. (2018). *Small Area Estimation of Health Insurance Coverage in 2010-2016*. Technical Report, U.S. Census Bureau.

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 and Survey Methodology*, 17, 1, 67-90, https://doi.org/10.21307/stattrans-2016-006.

Dahl, D., Day, R. and Tsai, J. (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112, 721-732.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.

Evans, R., and Sedransk, J. (1999). Methodology for pooling subpopulation regressions when there is uncertainty about which subpopulations are similar. *Statistica Sinica*, 9, 345-359.

Evans, R., and Sedransk, J. (2001). Combining data from experiments that may be similar. *Biometrika*, 88(3), 643-656.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Ha, N., and Sedransk, J. (2019). Assessing health insurance coverage in Florida using the Behavioral Risk Factor Surveillance System. *Statistics in Medicine*, 38(13), 2332-2352, https://doi.org/10.1002/sim.8108.

Jara, A., Hanson, T., Quintana, F., Müller, P. and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5), 1-30, https://doi.org/10.18637/jss.v040.i05.

Kim, J.-K., Park, S. and Kim, S.-Y. (2015). Small area estimation combining information from several sources. *Survey Methodology*, 41, 1, 21-36. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14150-eng.pdf.

Lohr, S., and Raghunathan, T. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293-312.

Malec, D., and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79(3), 593-601.

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, A*, 174(1), 31-50, https://doi.org/10.1111/j.1467-985X.2010.00648.x.

Muller, P., Quintana, F., Jara, A. and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Nandram, B., Berg, E. and Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environmental and Ecological Statistics*, 21, 507-530, https://doi.org/10.1007/s10651-013-0266-z.

Park, S., Kim, J. and Stukel, D. (2017). A measurement error model approach to survey data integration: Combining information from two surveys. *Metron*, 75, 345-357.

Pierannunzi, C., Xu, F., Wallace, R., Garvin, W., Greenlund, K., Bartoli, W., Ford, D., Eke, P. and Town, G. (2016). A methodological approach to small area estimation for the Behavioral Risk Factor Surveillance System. *Preventing Chronic Disease*, 2016 Jul. 14, 13, http://dx.doi.org/10.5888/pcd13.150480.

Polettini, S. (2017).A generalised semiparametric Bayesian Fay-Herriot model for small area estimation shrinking both means and variances. *Bayesian Analysis*, 2017, 12, 729-752.

Rao, J., and Molina, I. (2015). *Small Area Estimation*. 2nd Edition. New York: John Wiley & Sons, Inc.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

US Census Bureau (2021). SAHIE 2008 – 2015 demographic and income model methodology: Summary for counties and for states. https://www.Census.gov/Small Area Health Insurance Estimates (SAHIE) Program/Technical Documentation/Methodology/Demographic and Income Model Methodology (2008-2015).

Yan, G., and Sedransk, J. (2011). Improved inference for a linear mixed-effects model when the subpopulations are clustered. *Journal of Statistical Planning and Inference*, 141, 3489-3497.

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.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 89-115
Statistics Canada, Catalogue No. 12-001-X

Survey data integration for regression analysis using model calibration

Zhonglei Wang, Hang J. Kim and Jae Kwang Kim[1]

Abstract

We consider regression analysis in the context of data integration. To combine partial information from external sources, we employ the idea of model calibration which introduces a "working" reduced model based on the observed covariates. The working reduced model is not necessarily correctly specified but can be a useful device to incorporate the partial information from the external data. The actual implementation is based on a novel application of the information projection and model calibration weighting. The proposed method is particularly attractive for combining information from several sources with different missing patterns. The proposed method is applied to a real data example combining survey data from Korean National Health and Nutrition Examination Survey and big data from National Health Insurance Sharing Service in Korea.

Key Words: Big data; Empirical likelihood; Information projection; Measurement error models; Missing covariates.

1. Introduction

Data integration is an emerging research area in survey sampling. By incorporating the partial information from external samples, one can improve the efficiency of the resulting estimator and obtain a more reliable analysis. Lohr and Raghunathan (2017), Yang and Kim (2020), and Rao (2021) provide reviews of statistical methods of data integration for finite population inference. Many existing methods (e.g., Hidiroglou, 2001; Merkouris, 2010; Zubizarreta, 2015) are mainly concerned with estimating population means or totals while combining information for analytic inference such as regression analysis is not fully explored in the existing literature.

In this paper, we consider regression analysis in the context of data integration. When we combine data sources to perform a combined regression analysis, we may encounter some problems: covariates may not be fully observed or be subject to measurement errors. Thus, one may consider the problem as a missing-covariate regression problem. Robins, Rotnitzky and Zhao (1994) and Wang,Wang, Zhao and Ou (1997) discussed semiparametric estimation in regression analysis with missing covariate data under the missing-at-random covariate assumption. In our setup, the external data source with missing covariates can be a census or big data.

Under this setup, Chatterjee, Chen, Maas and Carroll (2016) developed a data integration method based on the constrained maximum likelihood, which uses a fully parametric model for the likelihood specification and a constraint developed from a reduced model for data integration. The constrained maximum likelihood method is efficient when the model is correctly specified but is not applicable when it is difficult or impossible to specify a correct density function. Kundu, Tang, and Chatterjee (2019) generalized the method of Chatterjee et al. (2016) to consider multiple regression models based on the

1. Zhonglei Wang, Assistant Professor, Wang Yanan Institute for Studies in Economics (WISE) and School of Economics, Xiamen University, Xiamen, Fujian 361005, PRC. E-mail: wangzl@xmu.edu.cn; Hang J. Kim, Associate Professor, Division of Statistics and Data Science, University of Cincinnati, Cincinnati, OH 45221, U.S.A. E-mail: kim3h4@ucmail.uc.edu; Jae Kwang Kim, Professor, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-mail: jkim@iastate.edu.

theory of generalized method of moments (Hansen, 1982, GMM). Recently, Xu and Shao (2020) develop a data integration method using a generalized method of moments technique, but their method implicitly assumes that the reduced model is correctly specified. Under a nested case-control design, Shin, Pfeiffer, Graubard and Gail (2020a) proposed to use the fully observed sample in the phase 2 to fit a parametric model, and missing covariates in the phase 1 sample are imputed; also see Shin, Pfeiffer, Graubard and Gail (2020b). Zhang, Deng, Wheeler, Qin and Yu (2021) developed a retrospective empirical likelihood framework to account for sampling bias in case-control studies. Sheng, Sun, Huang and Kim (2021) developed a penalized empirical likelihood approach to incorporate such information in the logistic regression setup.

To combine partial information from external sources, we employ the idea of model calibration (Wu and Sitter, 2001) which introduces a "working" reduced model based on observed covariates. The model parameters in the reduced model are estimated from external sources and then combined through a novel application of the empirical likelihood method (Owen, 1991; Qin and Lawless, 1994), which can be viewed as information projection (Csiszár and Shields, 2004). The working reduced model is not necessarily specified correctly, but a good working model can improve the efficiency of the resulting analysis. The proposed method is particularly attractive for combining information from several data sources with different missing patterns. In this case, we only need to specify different working models for different missing patterns.

Besides, our proposed method is based on the first moment conditions like usual regression analyses, so weak assumptions can broaden the applicability of the proposed method to many practical problems. In particular, the proposed method is directly applicable to survey sample data which is the main focus of our paper. We consider a more general regression setup and our proposed empirical likelihood method does not require that the working reduced model to be correctly specified.

We highlight the contribution of our paper as follows. First, we propose a unified framework for incorporating external data sources in the context of regression analysis. The proposed method uses weaker assumptions than the parametric model-based method of Chatterjee et al. (2016) and thus provides more robust estimation results. Second, the proposed method is widely applicable as it can easily handle multiple external data sources as demonstrated in Section 5. It can also be applied to the case where the external data source is subject to selection bias. In the real data application in Section 7, we demonstrated that our proposed method can utilize the external big data with unknown selection probabilities by applying propensity score weighting adjustment. Finally, our proposed method is easy to implement and fully justified theoretically. The computation is simple as it is a direct application of the standard empirical likelihood method and can be implemented using the existing software.

The paper is organized as follows. In Section 2, a basic setup is introduced, and the existing methods are presented. Section 3 presents the proposed approach, and Section 4 provides its asymptotic properties. In Section 5, an application to multiple data integration is presented. Section 6 presents simulation studies, followed by the application of the proposed method to real data in Section 7. Some concluding remarks are made in Section 8.

2. Basic setup

Consider a finite population $\mathcal{U}=\{1,\ldots,N\}$ of size N. Associated with the i^{th} unit, let y_i denote the study variable of interest and $\mathbf{x}_i=(\mathbf{x}_{i1},\mathbf{x}_{i2})$ the corresponding auxiliary vector of length p. We are interested in estimating a population parameter $\boldsymbol{\beta}_0$, which solves $\mathbf{U}_1(\boldsymbol{\beta})=\sum_{i\in\mathcal{U}}\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\mathbf{0}$ where $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x},y)$ is a pre-specified estimating function for $\boldsymbol{\beta}$. One example of the estimating function is $\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\{y_i-m_1(\mathbf{x}_i;\boldsymbol{\beta})\}\mathbf{h}_1(\mathbf{x}_i;\boldsymbol{\beta})$, which is implicitly based on a regression model $E(Y_i\mid\mathbf{x}_i)=m_1(\mathbf{x}_i;\boldsymbol{\beta})$ on the super-population level for some $\mathbf{h}_1(\mathbf{x}_i;\boldsymbol{\beta})$ satisfying certain identification conditions (e.g., Kim and Rao, 2009). From the finite population a probability sample $S_1\subset\mathcal{U}$ is selected, and a Z-estimator $\hat{\boldsymbol{\beta}}$ can be obtained by solving

$$\hat{\mathbf{U}}_1(\boldsymbol{\beta})\equiv\sum_{i\in S_1}d_{i1}\,\mathbf{U}_1(\boldsymbol{\beta};\mathbf{x}_i,y_i)=\mathbf{0}, \tag{2.1}$$

where d_i is the sampling weight for unit $i\in S_1$.

In addition to S_1, suppose that we observe $\mathbf{x}_{i1}$ and y_i throughout the finite population and wish to incorporate this extra information to improve the estimation efficiency of $\hat{\boldsymbol{\beta}}$. Before proposing our method, we introduce two related works, including Chen and Chen (2000) and Chatterjee et al. (2016).

Chen and Chen (2000) first considered this problem in the context of measurement error models. To explain their idea in our setup, we first consider a "working" reduced model,

$$E(Y_i\mid\mathbf{x}_{i1})=m_2(\mathbf{x}_{i1};\boldsymbol{\alpha}) \tag{2.2}$$

for some $\boldsymbol{\alpha}$. Under the working model (2.2), we can obtain an estimator $\hat{\boldsymbol{\alpha}}$ from the current sample S_1 by solving

$$\hat{\mathbf{U}}_2(\boldsymbol{\alpha})\equiv\sum_{i\in S_1}d_i\,\mathbf{U}_2(\boldsymbol{\alpha};\mathbf{x}_{i1},y_i)=\mathbf{0}, \tag{2.3}$$

where $\mathbf{U}_2(\boldsymbol{\alpha};\mathbf{x}_{i1},y_i)=\{y_i-m_2(\mathbf{x}_{i1};\boldsymbol{\alpha})\}\mathbf{h}_2(\mathbf{x}_{i1};\boldsymbol{\alpha})$ for some $\mathbf{h}_2(\mathbf{x}_{i1};\boldsymbol{\alpha})$ satisfying conditions similar to ones imposed to $\mathbf{h}_1(\mathbf{x}_i;\boldsymbol{\beta})$. In addition, one can get $\boldsymbol{\alpha}^*$ that solves $\sum_{i=1}^{N}\mathbf{U}_2(\boldsymbol{\alpha};\mathbf{x}_{i1},y_i)=\mathbf{0}$. Chen and Chen (2000) proposed using

$$\hat{\boldsymbol{\beta}}^*=\hat{\boldsymbol{\beta}}+\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})\{\hat{V}(\hat{\boldsymbol{\alpha}})\}^{-1}(\boldsymbol{\alpha}^*-\hat{\boldsymbol{\alpha}})$$

as an efficient estimator of $\boldsymbol{\beta}$ where $\hat{V}(\cdot)$ and $\widehat{\mathrm{Cov}}(\cdot)$ denote the design-based variance and covariance estimators, respectively. The working model in (2.2) is not necessarily correctly specified, but a good working model can improve the efficiency of the final estimator. While the estimator of Chen and Chen (2000) is theoretically justified, it can be numerically unstable as the estimation errors of the variance and covariance matrix can be large.

Chatterjee et al. (2016) considered a likelihood-based approach using a conditional distribution of Y_i given $\mathbf{X}_i$ with density $f(y_i\mid\mathbf{x}_i;\boldsymbol{\beta})$ and imposed a constraint based on external information. Specifically, they proposed to maximize

$$\prod_{i\in\mathcal{S}_1} f\left(y_i \mid \mathbf{x}_i;\boldsymbol{\beta}\right) dF\left(\mathbf{x}_i\right) \tag{2.4}$$

subject to

$$\iint \mathbf{U}_2\left(\boldsymbol{\alpha}^*; \mathbf{x}_1, y\right) f\left(y \mid \mathbf{x};\boldsymbol{\beta}\right) dy dF(\mathbf{x}) = \mathbf{0}, \tag{2.5}$$

where $F(\mathbf{x})$ is an unspecified distribution function for $\mathbf{x}$, $dF(\mathbf{x})$ is the Radon-Nikodym derivative of the distribution function $F(\mathbf{x})$ with respect to a certain dominating measure, and $\boldsymbol{\alpha}^*$ is the model parameter available from an external source. Following the likelihood based approach of Chatterjee et al. (2016), $\mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y\right)$ corresponds to the estimating function involving a "reduced" distribution function $g\left(y_i \mid \mathbf{x}_{i1}; \boldsymbol{\alpha}_0\right)$ with model parameter $\boldsymbol{\alpha}_0$, where $g\left(y_i \mid \mathbf{x}_{i1}; \alpha_0\right)$ can be incorrectly specified. That is, $\boldsymbol{\alpha}^*$ is the external information for $\boldsymbol{\alpha}_0$. Chatterjee et al. (2016) estimated $F(\mathbf{x})$ nonparametrically by empirical likelihood. By imposing this constraint into the maximum likelihood estimation, the external information $\boldsymbol{\alpha}^*$ can be naturally incorporated.

The constrained maximum likelihood (CML) method is not directly applicable to our conditional mean model in (2.1) as the likelihood function for $\boldsymbol{\beta}$ is not defined in our setup. Besides, the design feature for the probability sample S_1 is not directly applicable in their method. Nonetheless, one can use an objective function such as that in a generalized method of moments to apply the constrained optimization problem, which is asymptotically equivalent to the empirical likelihood method (Imbens, 2002). The empirical likelihood implementation of CML approach is discussed by Han and Lawless (2019).

3. Proposed approach

We now consider an alternative approach for combining information from several sources. To combine information from several sources, we use the Kullback-Leibler (KL) divergence measure to apply the information projection (Csiszár and Shields, 2004) on the model space with constraints. Let $\hat{P}$ be the empirical distribution of the sample with

$$\hat{P}(x, y) = \frac{1}{\sum_{i\in\mathcal{S}_1} d_i} \sum_{i\in\mathcal{S}_1} d_i \mathbb{I}\left\{(x, y) = (x_i, y_i)\right\}. \tag{3.1}$$

Given the empirical distribution $\hat{P}$, we wish to find the minimizer of

$$D\left(\hat{P} \,\|\, P\right) = \int \log\left\{d\hat{P}(\mathbf{x}, y)\right\} d\hat{P}(\mathbf{x}, y) - \int \log\left\{dP(\mathbf{x}, y)\right\} d\hat{P}(\mathbf{x}, y) \tag{3.2}$$

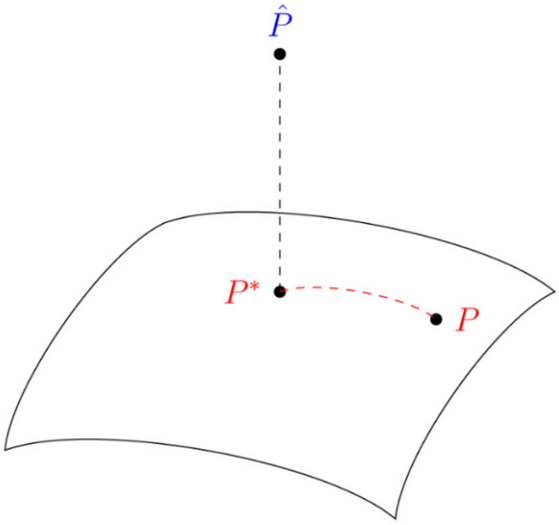
with respect to P in the model space. Notice that the first term is a constant and the minimizer of (3.2) is the pseudo maximum likelihood estimator of $\hat{P}$.

We consider the following constraints in our model at the finite-population level:

$$\sum_{i=1}^{N} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) p\left(\mathbf{x}_i, y_i\right) = 0 \quad \text{and} \quad \sum_{i=1}^{N} \mathbf{U}_2\left(\boldsymbol{\alpha}^*; \mathbf{x}_{i1}, y_i\right) p\left(\mathbf{x}_i, y_i\right) = 0, \tag{3.3}$$

where $p(\mathbf{x}_i, y_i)$ is the point mass assigned to point $(\mathbf{x}_i, y_i)$ in the finite population satisfying $\sum_{i=1}^{N} p(\mathbf{x}_i, y_i) = 1$. See Figure 3.1 for a graphical illustration of the information projection.

Figure 3.1 Information projection for the empirical distribution $\hat{P}$.



Note that P^* minimizes $D(\hat{P} \| P)$ among P satisfying the constraints in (3.3).

Using the weighted empirical distribution in (3.1), the KL divergence measure in (3.2) reduces to $D(\hat{P} \| P) = \text{constant} - \hat{N}^{-1} \sum_{i \in \mathcal{S}_1} d_i \log\{p(\mathbf{x}_i, y_i)\}$ where $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$. Thus, we only have to maximize $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$ subject to $\sum_{i=1}^{N} p_i = 1$ and the constraints in (3.3), where p_i abbreviates $p(\mathbf{x}_i, y_i)$. Note that having $p_i > 0$ for $i \notin \mathcal{S}_1$ will decrease the value of $l(\mathbf{p}) = \sum_{i \in \mathcal{S}_1} d_i \log(p_i)$, the solution $\hat{p}_i$ to this optimization problem should give $\hat{p}_i = 0$ for $i \notin S_1$. Therefore, we can safely set $p_i = 0$ for $i \notin S_1$ and express the problem as finding the maximizer of

$$Q(\mathbf{d}, \mathbf{w}) = \sum_{i \in \mathcal{S}_1} d_i \log(w_i) \tag{3.4}$$

subject to

$$\sum_{i \in \mathcal{S}_1} w_i = 1, \tag{3.5}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2(\boldsymbol{\alpha}^*; \mathbf{x}_{i1}, y_i) = \mathbf{0}, \tag{3.6}$$

$$\sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) = \mathbf{0}.$$

We use w_i instead of p_i to represent the final weights assigned to the sample elements.

Remark 1. *Maximizing the objective function in (3.4) is equivalent to minimizing the following cross entropy:*

$$-\sum_{i\in\mathcal{S}_1}\tilde{d}_i\log(w_i), \tag{3.7}$$

where $\tilde{d}_i = d_i\Big/\left(\sum_{i\in\mathcal{S}_1} d_i\right)$. *The objective function (3.7) is also the pseudo empirical log-likelihood function considered by Chen and Sitter (1999) and Wu and Rao (2006). Instead of (3.4), we may consider other objective functions, including the population empirical likelihood proposed by Chen and Kim (2014) for example.*

Our proposed method is different from Chatterjee et al. (2016) in that we use a more general integral constraint (2.5) which does not involve the conditional density function $f\left(y\middle|\,\mathbf{x};\boldsymbol{\beta}\right)$. Constraint (3.6) still incorporates the extra information in $\boldsymbol{\alpha}^*$. The above optimization can be solved by applying the standard profile empirical likelihood method or using the following two-step estimation method.

1. Find the calibration weights $\hat{\mathbf{w}} = \left\{\hat{w}_i : i\in\mathcal{S}_1\right\}$ maximizing $Q(\mathbf{d},\mathbf{w})$ subject to (3.5)-(3.6).
2. Once the solution $\hat{\mathbf{w}}$ is obtained from the calibration, estimate $\boldsymbol{\beta}$ by solving

$$\sum_{i\in\mathcal{S}_1}\hat{w}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)=\mathbf{0}. \tag{3.8}$$

If the benchmark $\boldsymbol{\alpha}^*$ is not available from the finite population but can be estimated from an independent external sample, we can use the information from both the original internal sample and the external sample to obtain the benchmark estimate. In practical situations, we may not have access to the raw data of the external sample but often be able to have its summary statistics. Suppose that the external sample provides a point estimator $\hat{\boldsymbol{\alpha}}_2$ and its variance estimator $\mathbf{V}_2 = \hat{V}\left(\hat{\boldsymbol{\alpha}}_2\right)$ for the working reduced model in (2.2). Then, an estimator of the benchmark $\boldsymbol{\alpha}^*$ can be obtained by

$$\hat{\boldsymbol{\alpha}}^* = \left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\left(\mathbf{V}_1^{-1}\hat{\boldsymbol{\alpha}}_1+\mathbf{V}_2^{-1}\hat{\boldsymbol{\alpha}}_2\right), \tag{3.9}$$

where $\hat{\boldsymbol{\alpha}}_1$ and $\mathbf{V}_1$ are estimated with the internal sample $\mathcal{S}_1$. Once $\hat{\boldsymbol{\alpha}}^*$ is obtained by (3.9), it replaces $\boldsymbol{\alpha}^*$ in the calibration equation in (3.6).

Similarly to Wu and Sitter (2001), the proposed method does not require a "true" working model as explained below. Let $\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha})=0$ be the estimating equation for obtaining $\boldsymbol{\alpha}^*$ computed from the external sample $\mathcal{S}_2$. Now, the final estimating function for $\boldsymbol{\beta}$ using the model calibration $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) = \sum_{i\in\mathcal{S}_1}\hat{w}_i\,\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)$ can be approximated by

$$\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta}) \doteq \hat{\mathbf{U}}_1(\boldsymbol{\beta}) + \mathbf{K}\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*) - \hat{\mathbf{U}}_2(\boldsymbol{\alpha}^*)\right\} \tag{3.10}$$

for some $\mathbf{K}$ where $\hat{\mathbf{U}}_1(\boldsymbol{\beta})$ and $\hat{\mathbf{U}}_2(\boldsymbol{\alpha})$ are computed by (2.1) and (2.3), respectively, from the internal sample $\mathcal{S}_1$. The approximation in (3.10) can be easily derived using the asymptotic equivalence of the calibration estimator and the regression estimator. Thus, even if $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha}^*)\right\}$ is not equal to zero, the solution to $\hat{\mathbf{U}}_{\text{cal}}(\boldsymbol{\beta})=0$ is consistent as $E\left\{\hat{\mathbf{U}}_{\text{ext}}(\boldsymbol{\alpha})-\hat{\mathbf{U}}_2(\boldsymbol{\alpha})\right\}=0$ by design.

Remark 2. *Although the working model* $E\left(Y_i \middle| \mathbf{x}_{i1}\right) = m_2\left(\mathbf{x}_{i1}; \boldsymbol{\alpha}\right)$ *does not need to be correctly specified, we can systematically find* $\mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right)$ *by casting its construction as a missing covariate problem, relying on the regression calibration technique. For example, suppose that* $\mathbf{x}_i = \left(x_{i1}, x_{i2}\right)$, *we set a predictor* $\hat{x}_{i2} = \beta_0 + \beta_1 x_{i1}$, *and an estimating equation is written by*

$$\mathbf{U}_1\left(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i\right) = \left\{y_i - m_1\left(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta}\right)\right\} \mathbf{h}_1\left(x_{i1}, \hat{x}_{i2}; \boldsymbol{\beta}\right) \tag{3.11}$$

for the control function of the model calibration method where $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)$. *We can either estimate* $\boldsymbol{\beta}$ *from sample* $\mathcal{S}_1$ *or use any fixed parameter value as long as the solution to* $\sum_{i \in \mathcal{S}_1} d_i \mathbf{U}_1\left(\boldsymbol{\beta}; x_{i1}, \hat{x}_{i2}, y_i\right) = \mathbf{0}$ *is unique. A benchmark estimator of* $\boldsymbol{\beta}$ *can be obtained using external samples to apply the proposed model calibration method. If we use the control function in (3.11), then we are essentially treating a regression of* y *on* x_1 *and* $\hat{x}_2$ *as the "working" model for model calibration. This is feasible only when we have direct access to an external sample* $\mathcal{S}_2$ *in addition to the internal sample* $\mathcal{S}_1$.

4. Theoretical properties

In this section, we investigate the asymptotic properties of the the proposed estimator $\hat{\boldsymbol{\beta}}$ to (3.8). Since the population parameters including $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}^*$ are determined by the finite population of size N, we explicitly use a subscript N for those in this section, e.g., $\boldsymbol{\beta}_{0N}$ and $\boldsymbol{\alpha}_N^*$, but we omit this subscript for $\left(d_i, \mathbf{x}_i, y_i\right)$ for simplicity. We consider two scenarios: when $\boldsymbol{\alpha}_N^*$ is available from the finite population and when we only have an external sample to estimate $\boldsymbol{\alpha}_N^*$ by the generalized least square in (3.9).

4.1 $\boldsymbol{\alpha}_N^*$ is available

Let $\tilde{d}_i = \hat{N}^{-1} d_i$ where $\hat{N} = \sum_{i \in \mathcal{S}_1} d_i$ is the Horvitz-Thompson estimator of the population size N. Replacing d_i by $\tilde{d}_i$ in (3.4), we consider the Lagrangian problem that maximizes

$$l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) = \sum_{i \in \mathcal{S}_1} \tilde{d}_i \log\left(w_i\right) + \boldsymbol{\lambda}^{\top} \sum_{i \in \mathcal{S}_1} w_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) + \phi\left(\sum_{i \in \mathcal{S}_1} w_i - 1\right),$$

where $\boldsymbol{\lambda}$ and ϕ are the Lagrange multipliers.

By setting $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial \boldsymbol{\lambda} = \mathbf{0}$, $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial \phi = 0$ and $\partial l\left(\mathbf{w}, \boldsymbol{\lambda}, \phi\right) / \partial w_i = 0$ for $i \in \mathcal{S}_1$, we get $\hat{\phi} = -1$ and $\hat{w}_i = \tilde{d}_i \left\{1 - \boldsymbol{\lambda}^{\top} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)\right\}^{-1}$. Then, the proposed method is equivalent to solving $g\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = \mathbf{0}$ where

$$g\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^{\top} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) \\ \sum_{i \in \mathcal{S}_1} \dfrac{\tilde{d}_i}{1 - \boldsymbol{\lambda}^{\top} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix}. \tag{4.1}$$

Denote the solution to (4.1) as $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\beta}}^{\top}, \boldsymbol{\lambda}^{\top}\right)^{\top}$. To investigate asymptotic properties of $\hat{\boldsymbol{\eta}}$, we propose the following regularity conditions.

C1. There exists a compact set $\mathcal{A}$ such that $Z_S = \sup_{\boldsymbol{\alpha}\in\mathcal{A}} \max_{i\in\mathcal{S}_1} \left\| \mathbf{U}_2\left(\boldsymbol{\alpha}; \mathbf{x}_{i1}, y_i\right)\right\| = o_p\left(n^{1/2}\right)$ and $\boldsymbol{\alpha}_N^* \in \mathcal{A}$ for $N \in \mathbb{N}$, where $\|\cdot\|$ denotes the Euclidean norm and the stochastic order is with respect to the sampling design.

C2. The sampling design satisfies the following convergence results.

a. There exist a compact set Ω such that $\boldsymbol{\beta}_{0N} \in \Omega$ for $N \in \mathbb{N}$ and an interior point of Ω, $\boldsymbol{\beta}_p$, such that $\lim_{N\to\infty} \boldsymbol{\beta}_{0N} = \boldsymbol{\beta}_p$.

b. There exists a continuous function $\mathbf{U}_0(\boldsymbol{\beta})$ over Ω such that $\sup_{\boldsymbol{\beta}\in\Omega} \left\| \sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}_i, y_i) - \mathbf{U}_0(\boldsymbol{\beta})\right\| \to 0$ in probability, where $\boldsymbol{\beta}_p$ is the unique solution to $\mathbf{U}_0(\boldsymbol{\beta}) = \mathbf{0}$.

c. $\sum_{i\in\mathcal{S}_1} \tilde{d}_i \partial \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right)/\partial \boldsymbol{\beta}^\top = \boldsymbol{\mathcal{I}}_{11} + o_p(1)$, where $\boldsymbol{\mathcal{I}}_{11}$ is non-stochastic and invertible.

d. $\sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right) \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)^\top = \boldsymbol{\mathcal{I}}_{12} + o_p(1)$, where $\boldsymbol{\mathcal{I}}_{12}$ is non-stochastic.

e. $\sum_{i\in\mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right)^{\otimes 2} = \boldsymbol{\mathcal{I}}_{22} + o_p(1)$, where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$ for any matrix $\mathbf{A}$ and $\boldsymbol{\mathcal{I}}_{22}$ is non-stochastic and positively definitive.

C3. The sampling design satisfies

$$n^{1/2} \sum_{i\in\mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1\left(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i\right) \\ \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix} \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right)$$

in distribution, where $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right)$ is a normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

C1 is a technical condition to obtain the asymptotic order of $\hat{\boldsymbol{\lambda}}$, and a similar condition is also assumed by Wu and Rao (2006); see their condition C1 for details. C2 assumes several convergence results for the two estimating functions. Specifically, C2a shows the parameter space of the finite population parameter $\boldsymbol{\beta}_{0N}$, and the convergence of $\boldsymbol{\beta}_{0N}$ can be satisfied under regularity conditions. Condition C2b is necessary to show $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_p \to 0$ in probability, then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N} \to 0$ in probability, coupled with C2a. Conditions C2c-C2e guarantee the central limit theorem for $\hat{\boldsymbol{\eta}}$. Note that $\boldsymbol{\mathcal{I}}_{22}$ is symmetric by C2e, but $\boldsymbol{\mathcal{I}}_{11}$ in C2c may be asymmetric for a certain estimating function $\mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}, y\right)$. Condition C3 is satisfied under regularity conditions for general sampling designs; see Fuller (2009, Section 1.3) for details.

Theorem 1. *Suppose that conditions C1-C3 hold. Then,* $n^{1/2}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\right) \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\eta\right)$ *in distribution, where* $\boldsymbol{\Sigma}_\eta = \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\Sigma}_u \left(\boldsymbol{\mathcal{I}}^{-1}\right)^\top$ *and*

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \mathbf{0} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix}.$$

The proof of Theorem 1 is presented in Appendix A. By Theorem 1, we can obtain that $n^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0N}\right) \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\beta\right)$ in distribution, where

$$\boldsymbol{\Sigma}_{\beta} = \boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{11}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}} - \boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{21}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}} - \boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}} + \boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}$$

and $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ correspond to the asymptotic variances of $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)$ and $n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$, respectively. Furthermore, we have the following result regarding the optimality of $\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i\right)$.

Corollary 1. *Suppose that the conditions in Theorem 1 hold. For a fixed estimating function* $\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x},y\right)$, $\hat{\boldsymbol{\beta}}$ *is optimal if* $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_1,y\right) = E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\mathbf{x}_1,y\right\}$ *holds almost surely for the working reduced model, where* $\mathbf{x}=\left(\mathbf{x}_1,\mathbf{x}_2\right)$ *and the expectation is taken with respect to the super-population model.*

The proof of Corollary 1 is relegated to Appendix B. Corollary 1 presents a sufficient condition on the reduced model to guarantee an optimal estimator $\hat{\boldsymbol{\beta}}$ if the working model is correctly specified. That is, even if we do not require that the reduced model is correctly specified for consistency, the efficiency gain is guaranteed only under the correct model specification. By Corollary 1, an optimal estimator of $\boldsymbol{\alpha}_N^*$ can be obtained by solving $E\left\{\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x},y\right)\middle|\mathbf{x}_1,y\right\}=\mathbf{0}$.

Under regularity conditions, it can be shown that $\boldsymbol{\Sigma}_{\beta} = \boldsymbol{\mathcal{I}}_{11}^{-1}\left(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}$ for simple random sampling with or without replacement. Since $\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\Sigma}_{11}\left(\boldsymbol{\mathcal{I}}_{11}^{-1}\right)^{\mathrm{T}}$ is the asymptotic variance of $n^{1/2}\left(\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{0N}\right)$, where $\hat{\boldsymbol{\beta}}_m$ solves $\sum_{i\in\mathcal{S}_1}d_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)=0$, the proposed approach achieves efficient estimation under simple random sampling; see Section S1 of the Supplementary Material for details.

4.2 An external estimator $\hat{\boldsymbol{\alpha}}_2$ is available

When $\boldsymbol{\alpha}^*$ is not available but an external sample is available to get $\hat{\boldsymbol{\alpha}}^*$ in (3.9), we consider

$$\tilde{g}\left(\boldsymbol{\eta}\right) = \begin{pmatrix} \sum_{i\in\mathcal{S}_1}\dfrac{\tilde{d}_i}{1-\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right) \\ \sum_{i\in\mathcal{S}_1}\dfrac{\tilde{d}_i}{1-\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right)}\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{i1},y_i\right) \end{pmatrix}. \tag{4.2}$$

Denote $\tilde{\boldsymbol{\eta}}$ to be the solution of $\tilde{g}\left(\boldsymbol{\eta}\right)=\mathbf{0}$. Then, the following additional assumptions are required to get the asymptotic properties for $\tilde{\boldsymbol{\eta}}$.

C4. $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\partial\mathbf{U}_2\left(\boldsymbol{\alpha};\mathbf{x}_{i1},y_i\right)/\partial\boldsymbol{\alpha}^{\mathrm{T}} = \boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}\right)+o_p\left(1\right)$ uniformly for $\boldsymbol{\alpha}\in\mathcal{A}$, where $\boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}\right)$ is non-stochastic. Besides, there exists an invertible matrix $\boldsymbol{\mathcal{I}}_0$ such that $\lim_{N\to\infty}\boldsymbol{\mathcal{I}}\left(\boldsymbol{\alpha}_N^*\right)=\boldsymbol{\mathcal{I}}_0$.

C5. The sampling design and the external sample satisfy the following convergence results.

a. Both $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ are consistent for $\boldsymbol{\alpha}^*$.

b. $\mathbf{V}_1$ and $\mathbf{V}_2$ are design consistent variance estimators of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$, respectively.

c. $\mathbf{V}_1^{-1}$, $\mathbf{V}_2^{-1}$, and $\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}$ exist in probability.

d. $\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}=\mathbf{W}+o_p\left(1\right)$, where $\mathbf{W}$ is non-stochastic.

e. There exists a scaling function $\gamma(n)$ such that $\gamma(n)\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*\right) \to \mathcal{N}\left(0, \boldsymbol{\Sigma}_2\right)$ in distribution, where $\boldsymbol{\Sigma}_2$ satisfies $\gamma(n)^2 \mathbf{V}_2 = \boldsymbol{\Sigma}_2 + o_p(1)$.

C4 is used to obtain the asymptotic order and the variance of $\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}_N^*$, and a similar condition was used by Yuan and Jennrich (1998). C5a and C5b assume the consistency of $\hat{\boldsymbol{\alpha}}_2$ and $\mathbf{V}_2$ obtained by an external sample. For the consistency of $\hat{\boldsymbol{\alpha}}_1$, a sufficient condition is similar with C2b. The design consistency of the variance estimator $\mathbf{V}_1$ can be obtained under general sampling designs; see Fuller (2009, Chapter 1) for details. C5c guarantees the existence of $\hat{\boldsymbol{\alpha}}^*$ for the proposed method. C5e shows the central limit theorem with respect to the summary statistic $\hat{\boldsymbol{\alpha}}_2$, and it is used to derive a similar result as C3 with $\boldsymbol{\alpha}^*$ replaced by $\hat{\boldsymbol{\alpha}}^*$. Specifically, the convergence rate of $\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*\right)$ is $\gamma(n)^{-1}$, which is determined by the external sample.

The following theorem establishes an asymptotic distribution similar to that in C3.

Theorem 2. *Suppose that conditions C1 and C3-C5 hold. Then,*

$$n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \begin{pmatrix} \mathbf{U}_1\left(\boldsymbol{\beta}_0; \mathbf{x}_i, y_i\right) \\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i\right) \end{pmatrix} \to \mathcal{N}\left(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_u\right)$$

in distribution, where

$$\tilde{\boldsymbol{\Sigma}}_u = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{pmatrix}.$$

Case 1. Specifically, if there exists a non-stochastic matrix $\boldsymbol{\Sigma}_c$ *such that* $n\mathbf{V}_2 = \boldsymbol{\Sigma}_c + o_p(1)$, *then* $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12} = \boldsymbol{\Sigma}_{12}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \boldsymbol{\mathcal{I}}_0^{\mathrm{T}}$, $\tilde{\boldsymbol{\Sigma}}_{21} = \tilde{\boldsymbol{\Sigma}}_{12}^{\mathrm{T}}$ *and* $\tilde{\boldsymbol{\Sigma}}_{22} = \boldsymbol{\mathcal{I}}_0 \mathbf{W}\left\{\boldsymbol{\Sigma}_c + \boldsymbol{\mathcal{I}}_0^{-1} \boldsymbol{\Sigma}_{22} (\boldsymbol{\mathcal{I}}_0^{-1})^{\mathrm{T}}\right\} \mathbf{W}^{\mathrm{T}} \boldsymbol{\mathcal{I}}_o^{\mathrm{T}}$;
Case 2. If $\mathbf{W} = 0$, *then* $\tilde{\boldsymbol{\Sigma}}_{ij} = \mathbf{0}$ *for* $(i, j) \neq (1,1)$ *and* $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}$.

The proof of Theorem 2 is presented in Appendix C. For Case 1, if $\hat{\boldsymbol{\alpha}}_2$ estimated from an external sample is much more efficient than $\hat{\boldsymbol{\alpha}}$ in the sense of $\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) = o_p\left(n^{-1/2}\right)$, then $\mathbf{W}$ is an identity matrix and $\tilde{\boldsymbol{\Sigma}}_{ij} = \boldsymbol{\Sigma}_{ij}$ for $i, j = 1, 2$. Thus, we can ignore the variability of the summary statistic $\hat{\boldsymbol{\alpha}}_2$ from the external sample and get the same asymptotic distribution as in C3. Although the asymptotic distributions are the same, C3 with known $\boldsymbol{\alpha}_N^*$ is not a special case of Theorem 2 since $\hat{\boldsymbol{\alpha}}_2 = \boldsymbol{\alpha}_N^*$ has zero variance, which violates C5c-C5e. On the other hand, if $\left(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_N^*\right) \asymp n^{-1/2}$ in probability, then $\hat{\boldsymbol{\alpha}}_2$ is as efficient as $\hat{\boldsymbol{\alpha}}_1$. Thus, $\mathbf{W}$ is not an identity matrix nor a zero matrix, and the proposed method is more efficient than one replacing $\boldsymbol{\alpha}^*$ by $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}_2$ due to the extra information provided by the external sample. It is trivial that we cannot use $\hat{\boldsymbol{\alpha}}_1$ to replace $\boldsymbol{\alpha}^*$ in (3.6); otherwise, we get $\hat{w}_i = \tilde{d}_i$, and (3.8) is equivalent to the traditional estimation equation $\sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1\left(\boldsymbol{\beta}; \mathbf{x}_i, y_i\right) = \mathbf{0}$ without calibration. If $\hat{\boldsymbol{\alpha}}_2$ is much less efficient than $\hat{\boldsymbol{\alpha}}_1$ in terms of convergence rate, then we should not use such an external sample for the proposed method because $\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}^* + o_p\left(n^{-1/2}\right)$ and $n^{1/2} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*; \mathbf{x}_{i1}, y_i\right) = o_p(1)$; see Appendix C for details. By C5, we can obtain the same consistency results in Lemmas A1-A2 for (4.2) under the same conditions. Thus, by Theorem 2, we obtain the following asymptotic distribution for $\tilde{\boldsymbol{\eta}}$.

Corollary 2. *Suppose that conditions C1-C5 hold. Then, we have* $n^{1/2}\left(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right)\to\mathcal{N}\left(0,\tilde{\boldsymbol{\Sigma}}_\eta\right)$ *in distribution, where* $\tilde{\boldsymbol{\Sigma}}_\eta=\boldsymbol{\mathcal{I}}^{-1}\tilde{\boldsymbol{\Sigma}}_u\left(\boldsymbol{\mathcal{I}}^{-1}\right)^{\mathrm{T}}$, *the form of* $\boldsymbol{\mathcal{I}}$ *is in Theorem 1, and the form of* $\tilde{\boldsymbol{\Sigma}}_\eta$ *is in Theorem 2.*

Remark 3. *It is worthy pointing out that when deriving the asymptotic properties in this section, we do not consider the weighting adjustments such as nonresponse adjustment, trimming, and raking. However, those weighting adjustments are commonly used in survey sampling. Thus, it is a promising research topic to generalize the proposed method incorporating those weighting adjustments.*

5. Multiple data integration

We now consider regression analysis combining partial information from external samples. To explain the idea, Table 5.1 shows an example data structure with three data sources $\left(A,B,C\right)$, where Sample A contains all the observations while samples B and C contain partial observations.

Table 5.1
Data structure for survey integration

Sample	Sampling Weight	z	x_1	x_2	y
A	d_a	X	X	X	X
B	d_b	X	X		X
C	d_c	X		X	X

Under the setup of Table 5.1, suppose that we are interested in estimating the parameters in the regression model $E\left(Y\middle|\, x_1,x_2\right)=m_1\left(\beta_0+\beta_1x_1+\beta_2x_2\right)$, where $m_1(\cdot)$ is known but $\boldsymbol{\beta}=\left(\beta_0,\beta_1,\beta_2\right)$ is unknown. The estimating equation for $\boldsymbol{\beta}$ using sample A can be written as

$$\hat{\mathbf{U}}_a\left(\boldsymbol{\beta}\right)\equiv\sum_{i\in A}d_{a,i}\left\{y_i-m\left(x_{i1},x_{i2};\boldsymbol{\beta}\right)\right\}\mathbf{h}\left(x_{i1},x_{i2};\boldsymbol{\beta}\right)=\mathbf{0} \tag{5.1}$$

for some $\mathbf{h}\left(x_{i1},x_{i2};\boldsymbol{\beta}\right)$ such that $\hat{\mathbf{U}}_a\left(\boldsymbol{\beta}\right)$ is linearly independent almost everywhere.

Now, we wish to incorporate the partial information from sample B. To do this, suppose that we have a "working" model for $E\left(Y\middle|\, x_1,z\right)$:

$$E\left(Y\middle|\, x_1,z\right)=m_2\left(x_1,z;\boldsymbol{\alpha}\right) \tag{5.2}$$

for some $\boldsymbol{\alpha}$. Note that, since $\left(z_i,x_{1i},y_i\right)$ are observed, we can use sample B to estimate $\boldsymbol{\alpha}$ by solving $\sum_{i\in B}d_{b,i}\mathbf{U}_b\left(\boldsymbol{\alpha};x_{i1},z_i,y_i\right)=\mathbf{0}$ for some $\mathbf{U}_b$ satisfying $E\left\{\mathbf{U}_b\left(\boldsymbol{\alpha};x_1,z,Y\right)\middle|\, x_1,z\right\}=\mathbf{0}$ under the working model (5.2).

Similarly, to incorporate the partial information from sample C, suppose that we have a "working" model for $E\left(Y\middle|\, x_2,z\right)$:

$$E\left(Y\middle|\, x_2,z\right)=m_3\left(x_2,z;\boldsymbol{\gamma}\right) \tag{5.3}$$

for some $\boldsymbol{\gamma}$. We can also construct an unbiased estimating equation $\sum_{i\in C} d_{c,i}\mathbf{U}_c\left(\boldsymbol{\gamma}; x_{i2}, z_i, y_i\right)=\mathbf{0}$ for some $\mathbf{U}_c$ satisfying $E\left\{\mathbf{U}_c\left(\boldsymbol{\gamma}; x_2, z, Y\right)\middle| x_2, z\right\}=\mathbf{0}$ under the working model (5.3). Once $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are obtained, we can use this extra information to improve the efficiency of $\hat{\boldsymbol{\beta}}$ in (5.1). To incorporate the extra information, we can formulate it as maximizing $Q\left(\mathbf{d}_a, \mathbf{w}\right)=\sum_{i\in A} d_{a,i}\log\left(w_i\right)$ subject to $\sum_{i\in A} w_i = N$ and

$$\sum_{i\in A} w_i\left[\mathbf{U}_b\left(\hat{\boldsymbol{\alpha}}; x_{i1}, z_i, y_i\right), \mathbf{U}_c\left(\hat{\boldsymbol{\gamma}}; x_{i2}, z_i, y_i\right)\right]=\mathbf{0}, \tag{5.4}$$

where $\mathbf{d}_a$ and $\mathbf{w}$ are sets containing the sampling weights and calibration weights with respect to sample A. Constraint (5.4) incorporates the extra information. Once the solution $\hat{w}_i$ is obtained, we can use $\sum_{i\in A}\hat{w}_i\left\{y_i - m\left(x_{i1}, x_{i2}; \boldsymbol{\beta}\right)\right\}\mathbf{h}\left(x_{i1}, x_{i2}; \boldsymbol{\beta}\right)=\mathbf{0}$ to estimate $\boldsymbol{\beta}$. The asymptotic results can be obtained similarly in Section 4.

Remark 4. *In this paper, we implicitly assume that the populations for the internal sample and the external samples are the same, but it is possible that those populations differ in some scenarios. For example, the external estimator* $\hat{\boldsymbol{\alpha}}$ *may be obtained based on a non-probability sample, whose sampling frame differs from the one for the probability sample due to the coverage bias in many opt-in surveys. There are several data integration methods incorporating information from heterogeneous populations. For example, Taylor, Choi and Han (2022) proposed to use ratios of coefficients to incorporate the external information under regularity conditions even when the populations for the internal and external samples differ. See also Zhai and Han (2022) and Sheng, Sun, Huang and Kim (2022) for penalized approaches when incorporating external information from heterogeneous populations. The aforementioned existing methods do not take the complex sampling properties into consideration, so it is promising to investigate data integration for heterogeneous populations under survey sampling in a future project.*

6. Simulation study

To evaluate the finite sample performance of the proposed estimator, we conducted simulation studies assuming several scenarios. We generated a finite population of size $N=100{,}000$, each record consisting of auxiliary variables $\mathbf{x}_i=\left(x_{i1}, x_{i2}\right)^{\mathrm{T}}$ of length $p=2$ and a response variable y_i. We assume that $\left(\mathbf{x}_i, y_i\right)$ is available for the internal sample $\mathcal{S}_1$ while only $\left(x_{i1}, y_i\right)$ is available for the external sample $\mathcal{S}_2$.

We evaluate the performance of the proposed estimator under a linear regression setup. In this case, we are interested in making statistical inference for $\boldsymbol{\beta}=\left(\beta_0, \beta_1, \beta_2\right)^{\mathrm{T}}$ that solves $\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)\left(1, x_{i1}, x_{i2}\right)^{\mathrm{T}}=\mathbf{0}$.

First, we consider two scenarios to generate covariates for the finite population: (i) $x_{i1}\sim N\left(3,1\right)$ and $x_{i2}\sim N\left(11, 6.5^2\right)$, where x_{i1} and x_{i2} are independent; (ii) $x_{i1}\sim N\left(3,1\right)$ and $x_{i2}=x_{i1}^2+\epsilon_i$ with $\epsilon_i\sim N\left(0,1\right)$. The simulation parameters are chosen such that the marginal mean and variance of x_{i2} are similar in the independent and the dependent settings. Second, the response variable is generated as $Y_i=\mu_i+\varepsilon_i$ with $\mu_i=1+2x_{i1}+x_{i2}$ under two scenarios: (i) homogeneous variance with $\varepsilon_i\sim N\left(0,9\right)$ and (ii) heteroscedasticity with $\varepsilon_i\,|\,\mathbf{x}_i\sim N\left(0,\sigma_i^2\right)$ with $\sigma_i=0.2\,|\,\mu_i\,|$. Third, we consider two sampling designs to

generate a probability sample $\mathcal{S}_1$ of (expected) size $n_1 = 1{,}000$: (i) simple random sampling without replacement (SRS), and (ii) Poisson sampling with inclusion probabilities satisfying $\pi_{1i} \propto \left(y_i - \min\{y_i : i = 1, \ldots, N\} + 10\right)^{1/2}$ and $\sum_{i=1}^{N} \pi_{1i} = n_1$. Last, we consider two sampling designs to generate an external sample $\mathcal{S}_2$ of (expected) size $n_2 = 10{,}000$: (i) SRS and (ii) Poisson sampling with inclusion probabilities satisfying $\pi_{2i} \propto \left\{1 + \exp\left(0.2x_{i1} + 0.1x_{i2} - 0.6\right)\right\}^{-1}$ and $\sum_{i=1}^{N} \pi_{2i} = n_2$. It is worthy pointing out that the sampling design for the internal sample is informative (Pfeffermann, 1993) under Poisson sampling, so ignoring the design feature may result in erroneous inference.

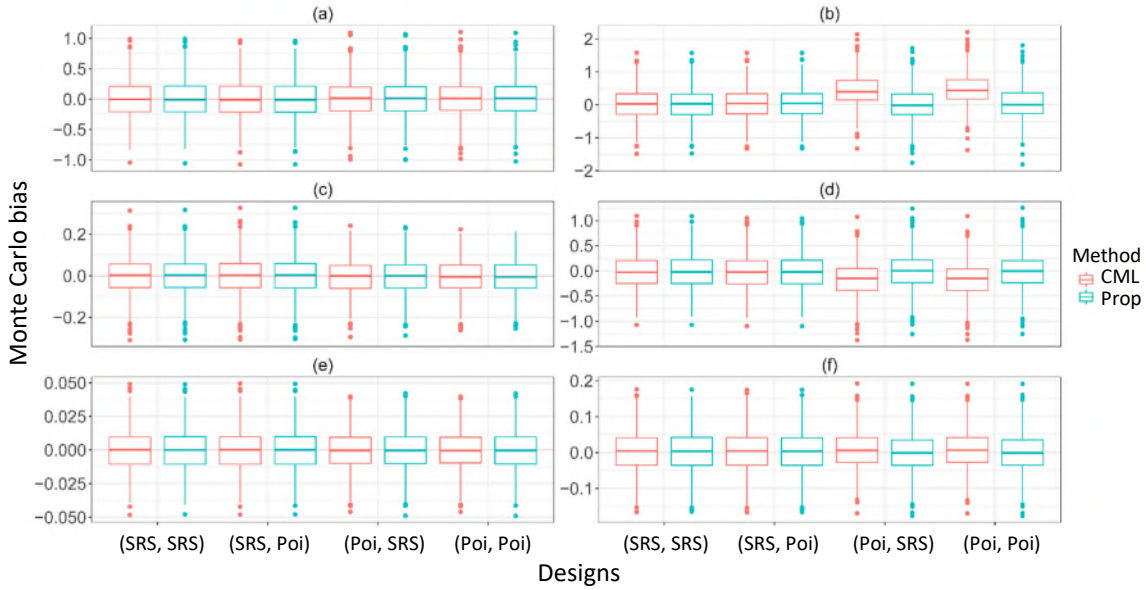
For the proposed estimator, we consider a working reduced model, $\sum_{i \in \mathcal{S}_2} \pi_{i2}^{-1} \left(y_i - \alpha_0 - \alpha_1 x_{i1}\right)\left(1, x_{i1}\right)^{\mathrm{T}} = \mathbf{0}$, whose solution is denoted as $\hat{\boldsymbol{\alpha}}_2$. Based on the external sample $\mathcal{S}_2$, we assume that a point estimator $\hat{\boldsymbol{\alpha}}_2$ and its variance estimator $\mathbf{V}_2 = \hat{V}\left(\hat{\boldsymbol{\alpha}}_2\right)$ are available as discussed in Section 3. Linearization is adopted to obtain a variance estimator $\mathbf{V}_2$; see the proof of Theorem 1 in Appendix A for details.

In the simulation study, the proposed estimator is compared with the constrained maximum likelihood (CML) estimator (Chatterjee et al., 2016). We assume a normal distribution for the likelihood function, i.e., $y_i \mid \mathbf{x}_i \sim N\left\{\left(1, \mathbf{x}_i^{\mathrm{T}}\right)\boldsymbol{\beta}, \sigma_{\text{full}}^2\right\}$. We also suppose that an analyst assumes $y_i \mid \mathbf{x}_{i1} \sim N\left\{\left(1, x_{i1}\right)\boldsymbol{\alpha}, \sigma_{\text{red}}^2\right\}$ for the working reduced model. See Section S2 of the Supplementary Material for the computation details. We consider the CML estimator under the setting where the extra information of $\left(y_i, x_{i1}\right)$ is available for an external sample, not for the entire population.

We conduct $M = 1{,}000$ Monte Carlo simulations, and Figures 6.1 and 6.2 show the Monte Carlo bias of the proposed and CML estimators for the homogeneous and heterogeneous variance setups, respectively. From Figure 6.1, when the variance of the error term is homogeneous and the internal sample is generated by SRS, the proposed estimator performs approximately the same as CML estimator in terms of Monte Carlo bias and variance. However, when the auxiliaries are correlated and the internal sample is generated by Poisson sampling, the CML estimator is questionable, since its model is wrongly specified under the informative Poisson sampling design. For example, the Monte Carlo bias of the CML estimator is not negligible when estimating β_0 and β_1. Because the proposed estimator incorporates the design features, its performance is satisfactory for all setups. As shown in Figure 6.2, even when the internal sample is generated by SRS, the CML estimator is slightly less efficient than the proposed estimator. The reason is that the CML estimator fails to take the heterogeneous variance into consideration, but the proposed estimator does not make any distribution assumption. When the internal sample is generated by an informative Poisson sampling design, the CML performs poorly, since it is not unbiased, and since its variance is larger than the proposed estimator.

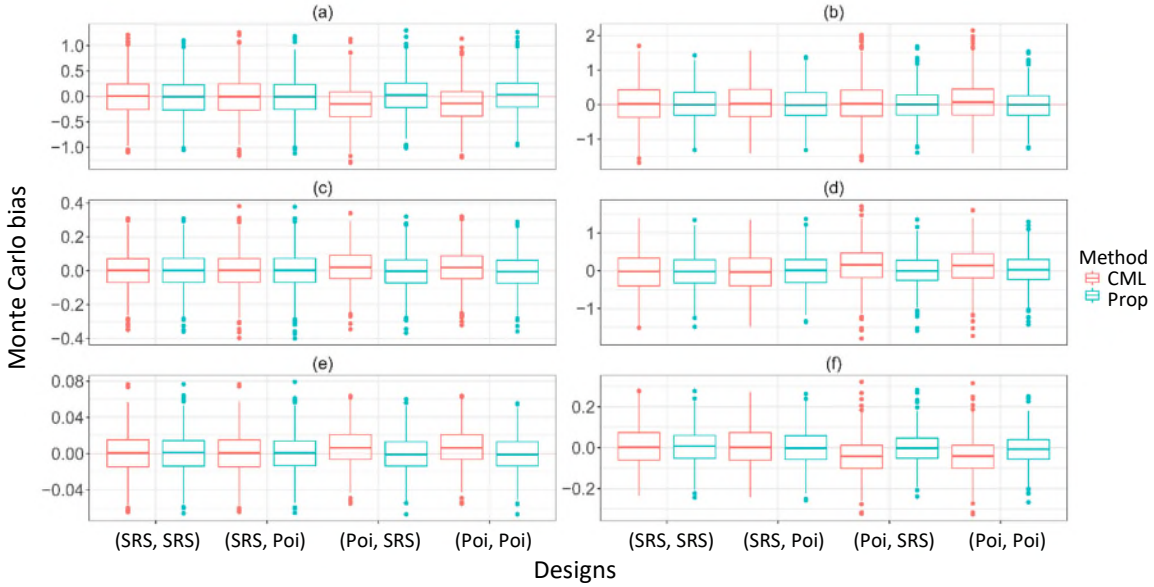
Table 6.1 shows the coverage rate of a 95% confidence interval for the proposed estimator under different settings. Chatterjee et al. (2016) only investigated the theoretical properties of their estimator when the population-level information is available. Thus, no interval estimator can be provided if only an external sample is available. By Table 6.1, we conclude that the coverage rates of the confidence intervals are all close to its nominal truth 0.95 under different settings. One possible reason for this phenomenon is that the proposed estimator is model free, so the proposed model is more robust and can be used under complex sampling designs.

Figure 6.1 Monte Carlo bias of the proposed and CML estimators based on 1,000 Monte Carlo simulations under the homogeneous variance setup.



The first to the third rows stand for the Monte Carlo bias for estimating β_0, β_1 and β_2, respectively. The three plots, including (a), (c) and (e), in the left column show the results when the auxiliary variables are independently generated, and those, including (b), (d) and (f), in the right column are for the case when the auxiliaries are dependent. "CML" and "Prop" stands for the CML estimator and the proposed estimator, respectively. The first design in the parenthesis is used to generate the internal sample $\mathcal{S}_1$, and the second one to generate the external sample $\mathcal{S}_2$. "SRS" and "Poi" represents Sampling random sampling and Poisson sampling.

Figure 6.2 Monte Carlo bias of the proposed and CML estimators based on 1,000 Monte Carlo simulations under the heterogeneous variance setup.



The first to the third rows stand for the Monte Carlo bias for estimating β_0, β_1 and β_2, respectively. The three plots, including (a), (c) and (e), in the left column show the results when the auxiliary variables are independently generated, and those, including (b), (d) and (f), in the right column are for the case when the auxiliaries are dependent. "CML" and "Prop" stands for the CML estimator and the proposed estimator, respectively. The first design in the parenthesis is used to generate the internal sample $\mathcal{S}_1$, and the second one to generate the external sample $\mathcal{S}_2$. "SRS" and "Poi" represents Sampling random sampling and Poisson sampling.

Table 6.1
Coverage rate of a 95% confidence interval by the proposed method based on 1,000 Monte Carlo simulations under different setups

	$\mathcal{S}_1$ Des	$\mathcal{S}_2$ Des	Independent			Dependent		
			β_0	β_1	β_2	β_0	β_1	β_2
Homo	SRS	SRS	0.948	0.952	0.939	0.945	0.948	0.934
		Poi	0.945	0.951	0.938	0.946	0.946	0.934
	Poi	SRS	0.957	0.966	0.949	0.935	0.943	0.940
		Poi	0.962	0.964	0.951	0.936	0.943	0.938
Hete	SRS	SRS	0.944	0.942	0.933	0.933	0.925	0.935
		Poi	0.949	0.942	0.935	0.935	0.934	0.931
	Poi	SRS	0.959	0.955	0.935	0.948	0.950	0.941
		Poi	0.961	0.956	0.944	0.952	0.949	0.946

Note: "Homo" and "Hete" stands for the homogeneous and heterogeneous variance for the error term, respectively. "$\mathcal{S}_1$ Des" and "$\mathcal{S}_2$ Des" show the sampling design used to generate the internal sample $\mathcal{S}_1$ and the external sample $\mathcal{S}_2$. "SRS" and "Poi" stands for Sampling random sampling and Poisson sampling, respectively. "Independent" and "Dependent" correspond to the cases when the auxiliary variables are independent and dependent, respectively.

An additional simulation with a logistic regression setup is relegated to Section S3 of the Supplementary Material, and similar conclusions can be reached.

7. Application study

7.1 Data description and problem formulation

As an application example, we apply the proposed method to analyze a subset of the data from the Korea National Health and Nutrition Examination Survey (KNHANES). The annual survey includes approximately 5,000 individuals each year and collects information regarding health-related behaviors by interviews, basic health conditions by physical and blood tests, and dietary intake by nutrition survey. The sampling design of KNHANES is a stratified sampling using age, sex, and region as stratification variables. The final sampling weights are computed via nonresponse adjustment and post-stratification, then provided to data users with survey variables.

To improve the efficiency of data analysis with KNHANES of size $n_1 = 4{,}929$, we used an external public database provided by the National Health Insurance Sharing Service (NHISS) in Korea. The big data provided by NHISS contain about $n_2 = 1{,}000{,}000$ individuals with health-related information, some of whose variables are a subset of variables in KNHANES.

These data structures, with the small n_1, the large n_2, and the big data having a subset of variables in the internal sample, are suited well to the setting we addressed in Section 2. However, there is another complication in applying the proposed method to the real application. In the NHISS data, its selection probabilities are unknown, so the design consistent estimator $\hat{\boldsymbol{\alpha}}_2$ in (3.9) is unavailable. Section 7.2

addresses this issue by using a propensity weighting approach and Section 7.3 presents the analysis results of the application study.

7.2 Propensity weighting for external data with unknown selection probability

We now consider an extension of the proposed method to the case where the external sample $\mathcal{S}_2$ is a big data with unknown selection probabilities. In this case, the working model for $E\left(Y_i \mid \mathbf{x}_{i1}\right) = m\left(\boldsymbol{\alpha}^{\top}\mathbf{x}_{i1}\right)$ may not hold for the sample S_2. Nonetheless, we may still solve

$$\sum_{i\in\mathcal{S}_2}\left\{y_i - m\left(\boldsymbol{\alpha}^{\top}\mathbf{x}_{i1}\right)\right\}\mathbf{x}_{i1} = \mathbf{0} \tag{7.1}$$

to obtain $\hat{\alpha}_0$ and $\hat{\alpha}_1$. If the sampling mechanism for $\mathcal{S}_2$ is ignorable or non-informative, then the solution of (7.1) is unbiased; otherwise, the resulting estimator is biased.

To remove the selection biases in the big data estimate, Kim and Wang (2019) suggested using propensity score weights in (7.1) to obtain an unbiased estimator of $\boldsymbol{\alpha}$. To construct the propensity score weights, we employ a nonignorable nonresponse model, $P\left(\delta_i = 1 \mid \mathbf{x}_{i1}, y_i\right) = \pi\left(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi}\right)$, where $\delta_i = 1$ if $i \in S_2$ and zero otherwise. Note that we can express $\pi\left(\mathbf{x}_{i1}, y_i\right)^{-1} = 1 + \left(N_0/N_1\right) r\left(\mathbf{x}_{i1}, y_i\right)$, where $r\left(\mathbf{x}_{i1}, y_i\right) = f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 0\right) \big/ f\left(\mathbf{x}_{i1}, y_i \mid \delta_i = 1\right)$ is the density ratio function with $N_1 = \sum_{i=1}^{N}\delta_i$ and $N_0 = N - N_1$. Using the motivation of Wang and Kim (2021), we may assume a log-linear density ratio model, $\log\left\{r\left(x_{i1}, y_i; \boldsymbol{\phi}\right)\right\} = \phi_0 + \phi_1 x_{i1} + \phi_2 y_i$. The maximum entropy estimator of $\boldsymbol{\phi}$ is obtained by solving $\left(1/N_1\right)\sum_{i=1}^{N}\delta_i \exp\left(\phi_0 + \phi_1 x_{i1} + \phi_2 y_i\right)\left(1, x_{i1}, y_i\right) = \left(1, \hat{\bar{x}}_1, \hat{\bar{y}}\right)$, where $\left(\hat{\bar{x}}_1, \hat{\bar{y}}\right) = \left(1/\hat{N}_0\right)\left\{\sum_{i\in\mathcal{S}_1} d_i\left(x_{i1}, y_i\right) - \sum_{i=1}^{N}\delta_i\left(x_{i1}, y_i\right)\right\}$, $\hat{N}_0 = \sum_{i\in\mathcal{S}_1} d_i - N_1$, and $\mathcal{S}_1$ is the internal sample. Once $\hat{\boldsymbol{\phi}}$ is obtained, we can construct $\hat{\pi}\left(x_{i1}, y_i\right)$ and solve

$$\sum_{i\in\mathcal{S}_2}\frac{1}{\hat{\pi}\left(x_{i1}, y_i\right)}\left\{y_i - m\left(\alpha_0 + \alpha_1 x_{i1}\right)\right\}\left(1, x_{i1}\right) = \left(0, 0\right) \tag{7.2}$$

to obtain $\hat{\boldsymbol{\alpha}}_2 = \left(\hat{\alpha}_0, \hat{\alpha}_1\right)$.

In addition, we can use the internal sample $\mathcal{S}_1$ to fit the same working model to obtain $\hat{\boldsymbol{\alpha}}_1$. After that, we obtain $\hat{\boldsymbol{\alpha}}^{*}$ using (3.9) and apply the proposed calibration weighting method to combine information from the big data. In practice, $\mathbf{V}_2$ in (3.9) is difficult to compute, but it is negligibly small if the sample size for $\mathcal{S}_2$ is huge. In this case, we may simply use $\hat{\boldsymbol{\alpha}}^{*} = \hat{\boldsymbol{\alpha}}_2$ in the calibration problem.

7.3 Application study results: Korea National Health and Nutrition Examination Survey

In this application study, we use $n_1 = 4{,}929$ records of KNHANES data that have no missing values in four variables: Total cholesterol, Hemoglobin, Triglyceride, and high-density lipoprotein (HDL)

cholesterol. For demonstration purpose, we assume that an analyst is interested in conducing the following linear regression analysis,

$$E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_i\right) = \beta_0 + \beta_1 \text{Hemoglobin}_i + \beta_2 \text{Triglyceride}_i + \beta_3 \text{HDL}_i \quad \text{for} \quad i \in \mathcal{S}_1;$$

check Section S4 of the Supplementary Material for details about the linearity assumption. In our data, the biggest absolute value of the pairwise correlation among covariates is -0.40 observed between Triglyceride and HDL cholesterol, which is similar to a scenario in Section 6 where the covariates were highly correlated. The big external data consist of $n_2 = 1{,}000{,}000$ records of NHISS data with fully observed items in Total cholesterol, Hemoglobin, and Triglyceride. The assumed working reduced model is

$$E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_{i1}\right) = \alpha_0 + \alpha_1 \text{Hemoglobin}_i + \alpha_2 \text{Triglyceride}_i \quad \text{for} \quad i \in \mathcal{S}_1 \cup \mathcal{S}_2.$$

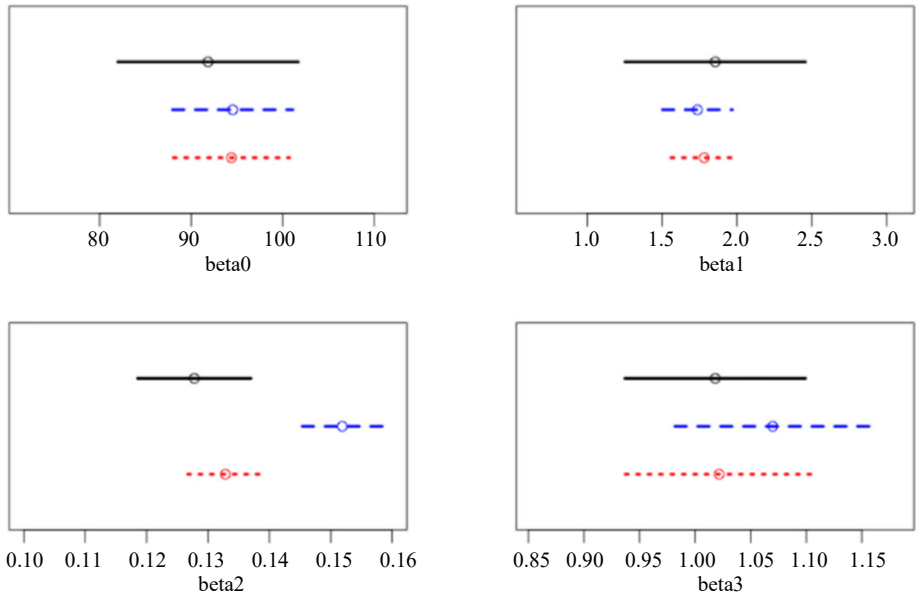
In this application study, we implement our proposed methods with the external sample, where $\hat{\boldsymbol{\alpha}}_2$ is used instead of $\boldsymbol{\alpha}^*$ that is unavailable as we do not have information regarding the entire population. With the external sample whose selection probabilities are unknown, we prepare two versions of proposed methods: (i) considering $\mathcal{S}_2$ as SRS, i.e., without propensity weighting, and (ii) with the propensity weighting adjustment introduced in Section 7.2. For the propensity weighting, we fit the log-linear density ratio model to the external data, $\log\left\{r\left(\mathbf{x}_{i1}, y_i; \boldsymbol{\phi}\right)\right\} = \phi_0 + \phi_1 \text{Hemoglobin}_i + \phi_2 \text{Triglyceride}_i + \phi_3 \text{Total Cholesterol}_i$, calculate $\hat{\pi}\left(\mathbf{x}_{i1}, y_i\right)$ given $\hat{\boldsymbol{\phi}}$, then solve

$$\sum_{i \in \mathcal{S}_2} \frac{1}{\hat{\pi}\left(x_{i1}, y_i\right)} \left\{y_i - m\left(\alpha_0 + \alpha_1 x_{i1}\right)\right\} \left(1, x_{i1}\right) = \left(0, 0\right)$$

to obtain $\hat{\boldsymbol{\alpha}}_2$. The above logistic regression model is commonly assumed in the literature; see Elliott and Valliant (2017), Chen, Li and Wu (2020), Wang and Kim (2021) and the references within for details. Since the CML estimator fails to incorporate the design features, it is not considered in the application section. The performances of proposed methods are compared with the reference method that uses the internal sample $\mathcal{S}_1$ only to get weighted least square estimates considering the sampling weights.

Figure 7.1 shows the point estimates and the 95% confidence intervals of $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$ for each method. The proposed methods show smaller variances for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ than using the internal sample only. This result coincides with our findings in the simulation studies of the previous section. For β_2, the estimator of the proposed method without propensity weighting shows a systematic difference from the other two estimators. When the propensity weighting adjustment is coupled with the proposed method, its confidence interval of β_2 is contained by that of using the internal sample only. This result implies that the systematic bias due to the disregard of the sampling probabilities is addressed by the propensity weighting adjustment. No efficiency gain in estimating β_3 was expected as the external data contain information of x_{i1} (Hemoglobin) and x_{i2} (Triglyceride), not x_{i3} (HDL).

Figure 7.1 Comparison of the regression analysis for $E\left(\text{Total Cholesterol}_i \mid \mathbf{x}_i\right) = \beta_0 + \beta_1 \text{Hemoglobin}_i + \beta_2 \text{Triglyceride}_i + \beta_3 \text{HDL}_i$ using the internal data from Korea National Health and Nutrition Examination Survey supported by the big external data from the National Health Insurance Sharing Service database.



For each panel, circles are point estimates and lines are their 95% confidence intervals for using the internal sample $\mathcal{S}_1$ only with the weighted least square (top solid line), the proposed method without adjustment (middle dashed line), and the proposed method with propensity score weighting adjustment (bottom dotted line).

8. Conclusion

Incorporating external data sources into the regression analysis of the internal sample is an important practical problem. We have addressed this problem using a novel application of the information projection (Csiszár and Shields, 2004) and the model calibration weighting (Wu and Sitter, 2001). The proposed method is directly applicable to survey sampling and can be easily extended to multiple data integration. The proposed method is easy to implement and does not require direct access to external data. As long as the estimated regression coefficients and their standard errors for the working reduced model are available, we can incorporate the extra information into our analysis.

There are several possible directions on future research extensions. First, a Bayesian approach can be developed under the same setup. One may use the Bayesian empirical likelihood method of Zhao, Ghosh, Rao and Wu (2020) in this setup. The proposed method can potentially be used to combine the randomized clinical trial data with big real-world data (Yang, Zheng and Wang, 2020); such extensions will be presented elsewhere. It will be also interesting to connect the proposed approach to two-phase (double) sampling design whose efficient design and estimation has been recently studied actively (Rivera-Rodriguez, Spiegelman and Haneuse, 2019; Rivera-Rodriguez, Haneuse, Wang and Spiegelman, 2020; Wang, Williams, Chen and Chen, 2020). The data structure of the two-phase sampling with the large-n, small-p first stage sample and the small-n, large-p second stage sample is well suited to the set-up assumed by the suggested model calibration approach.

Acknowledgements

We appreciate the constructive comments of the reviewers and the associate editor. The research of Z. Wang was partially supported by the National Natural Science Foundation of China grants (Award no: 11901487, 72033002) and the Fundamental Scientific Center of National Natural Science Foundation of China grant (Award no: 71988101). The research of J.K. Kim was partially supported by the National Science Foundation grant (Award no: CSSI-1931380), a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University, and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

Supplement

The supplementary material can be found in the document https://arxiv.org/abs/2107.06448, and it contains special case under simple random sampling (S1), implementation of Chatterjee et al. (2016) (S2), an additional simulation study (S3), and validation for the linearity assumption for the KNHANES dataset (S4).

Appendix

A. Proof of Theorem 1

Lemma A1. *Suppose that conditions C1, C2e and C3 hold. Then,* $\left\| \hat{\boldsymbol{\lambda}} \right\| = O_p\left(n^{-1/2}\right)$.

Proof of Lemma A1. Denote $\hat{\boldsymbol{\lambda}} = \rho\boldsymbol{\theta}$, *where* $\rho = \left\| \hat{\boldsymbol{\lambda}} \right\|$ *and* $\boldsymbol{\theta} = \rho^{-1}\hat{\boldsymbol{\lambda}}$ *is a vector of unit length. Then, we have*

$$
\begin{aligned}
\mathbf{0} &= \left| \sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i}{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right| \\
&= \left| \boldsymbol{\theta}^{\mathrm{T}} \sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i}{1-\rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right| \\
&= \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) + \rho \sum_{i\in\mathcal{S}_1} \frac{\tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta}}{1-\rho\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)} \right| \\
&\geq \frac{\rho}{1+\rho Z_S} \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta} \right| - \left| \sum_{i\in\mathcal{S}_1} \tilde{d}_i \boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) \right|, \qquad \text{(A.1)}
\end{aligned}
$$

where the first equality holds since $g\left(\hat{\boldsymbol{\eta}}\right) = \mathbf{0}$, and the last inequality holds by the triangular inequality.

By C2e and the Rayleigh-Ritz Theorem (Horn and Johnson, 2012, Section 4.2), there exists a constant $\sigma_0 > 0$ such that

$$\sum_{i\in\mathcal{S}_1}\tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\left\{\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{\mathrm{T}}\boldsymbol{\theta} > \sigma_0 + o_p(1). \tag{A.2}$$

By C3 and the Slutsky's theorem, we have

$$\sum_{i\in\mathcal{S}_1}\tilde{d}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right) = O_p\left(n^{-1/2}\right). \tag{A.3}$$

Thus, by C1 and (A.1)-(A.3), we have proved Lemma A1.

Lemma A2. *Suppose that conditions C1, C2a-C2e and C3 hold. Then,* $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0N}=o_p(1)$.

Proof of Lemma A2. By Lemma A1 and C1, we conclude that

$$\begin{aligned}\max_{i\in\mathcal{S}_1}\left|\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right| &\le \max_{i\in\mathcal{S}_1}\left\|\hat{\boldsymbol{\lambda}}\right\|\left\|\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\| \\ &= \left\|\hat{\boldsymbol{\lambda}}\right\|\max_{i\in\mathcal{S}_1}\left\|\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\| = o_p(1).\end{aligned} \tag{A.4}$$

First, we show that

$$\sum_{i\in\mathcal{S}_1}\frac{\tilde{d}_i}{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)-\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\to\mathbf{0} \tag{A.5}$$

in probability uniformly for $\boldsymbol{\beta}\in\Omega$. By (A.4), we have

$$\begin{aligned}&\left\|\sum_{i\in\mathcal{S}_1}\tilde{d}_i\left\{\frac{1}{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)}-1\right\}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\right\| \\ &\quad=\left\|\sum_{i\in\mathcal{S}_1}\tilde{d}_i\left\{\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+o_p\left(\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right)\right\}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\right\| \\ &\quad\le\left(1+o_p(1)\right)\max_{i\in\mathcal{S}_1}\left|\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right|\left\|\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\right\|.\end{aligned} \tag{A.6}$$

By C2a-C2b, there exists a constant $C_{u1}>0$ such that $\sup_{\boldsymbol{\beta}\in\Omega}\left\|\mathbf{U}_0\left(\boldsymbol{\beta}\right)\right\|<C_{u1}$. Since $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)$ converge uniformly to $\mathbf{U}_0\left(\boldsymbol{\beta}\right)$ in probability, we conclude that

$$\left\|\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\right\|<C_{u1}+o_p(1) \tag{A.7}$$

uniformly over Ω. By (A.4) and (A.6)-(A.7), we have validated (A.5).

By C2b and (A.5), we conclude that $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\left\{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{-1}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)$ converges uniformly to $U_0\left(\boldsymbol{\beta}\right)$ in probability. Denote $Q_0\left(\boldsymbol{\beta}\right)=-U_0\left(\boldsymbol{\beta}\right)^2$ and $Q_s\left(\boldsymbol{\beta}\right)=-\left[\sum_{i\in\mathcal{S}_1}\tilde{d}_i\left\{1-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\right\}^{-1}\mathbf{U}_1\left(\boldsymbol{\beta};\mathbf{x}_i,y_i\right)\right]^2$. Then, $\boldsymbol{\beta}_p$ uniquely maximizes $Q_0\left(\boldsymbol{\beta}\right)$ by (C2b), and $\hat{\boldsymbol{\beta}}$ maximizes $Q_s\left(\boldsymbol{\beta}\right)$. In addition, $Q_s\left(\boldsymbol{\beta}\right)$ converge uniformly to $Q_0\left(\boldsymbol{\beta}\right)$ in probability over the compact set Ω. Thus, by C2a and Theorem 2.1 of Engle and McFadden (1994, Chapter 36), we have finished the proof for Lemma A2.

Proof of Theorem 1. By Lemmas A1-A2, we have shown that

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}_0 + o_p(1), \tag{A.8}$$

where $\hat{\boldsymbol{\eta}}^{\mathrm{T}} = \left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\right)$, $\boldsymbol{\eta}_0^{\mathrm{T}} = \left(\boldsymbol{\beta}_{0N}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right)$, and $\mathbf{0}$ is a vector of zero with the same length of $\hat{\boldsymbol{\lambda}}$.

By (A.8) and the Taylor expansion, we have

$$\begin{aligned}
\mathbf{0} \;=\; & g(\hat{\boldsymbol{\eta}}) = g(\boldsymbol{\eta}_0) + \frac{\partial g}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + o_p\left(\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \|\right) \\
= & \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \\ \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i) \end{pmatrix} \\
& + \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \frac{\partial \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i)\right\}^{\mathrm{T}} \\ \mathbf{0} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i)\right\}^{\otimes 2} \end{pmatrix} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\
& + o_p\left(\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \|\right).
\end{aligned} \tag{A.9}$$

By (C3), we have

$$n^{1/2} \begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \\ \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i) \end{pmatrix} \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u) \tag{A.10}$$

in distribution. By (C2e)-(C2c), we conclude that

$$\begin{pmatrix} \sum_{i \in \mathcal{S}_1} \tilde{d}_i \frac{\partial \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \mathbf{U}_1(\boldsymbol{\beta}_{0N}; \mathbf{x}_i, y_i) \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i)\right\}^{\mathrm{T}} \\ \mathbf{0} & \sum_{i \in \mathcal{S}_1} \tilde{d}_i \left\{\mathbf{U}_2(\boldsymbol{\alpha}_N^*; \mathbf{x}_{1i}, y_i)\right\}^{\otimes 2} \end{pmatrix} \to \boldsymbol{\mathcal{I}} \tag{A.11}$$

in probability, where

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \mathbf{0} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix}.$$

By (A.9)-(A.11), we conclude that

$$n^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \tag{A.12}$$

in distribution, where $\boldsymbol{\Sigma}_\eta = \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\Sigma}_u (\boldsymbol{\mathcal{I}}^{-1})^{\mathrm{T}}$.

B. Proof of Corollary 1

Since $\mathbf{U}_1(\boldsymbol{\beta}; \mathbf{x}, y)$ is given, it is enough to consider

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}} + \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}^{\mathrm{T}},$$

the asymptotic variance of $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)$, where $\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) = n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i)$ and $\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*) = n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{i1},y_i)$.

Consider

$$\begin{aligned}
&\mathrm{Var}\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\right\} \\
&\quad = E\left[\mathrm{Var}\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\,\middle|\,\mathcal{A}_N\right\}\right] + \mathrm{Var}\left[E\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\,\middle|\,\mathcal{A}_N\right\}\right] \\
&\quad \succeq E\left[\mathrm{Var}\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N})\right\}\right],
\end{aligned}$$

where $\mathcal{A}_N = \left\{(\mathbf{x}_{i1},y): i\in\mathcal{S}_1\right\}$, $A\succeq B$ is equivalent to that $A-B$ is non-negatively definitive for two matrices A and B with the same dimension, and the last inequality holds since $\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)$ is non-stochastic conditional on $\mathcal{A}_N$. Thus, $\mathrm{Var}\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\right\}$ achieves minimum if $\mathrm{Var}\left[E\left\{\tilde{\mathbf{U}}_1(\boldsymbol{\beta}_{0N}) - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\tilde{\mathbf{U}}_2(\boldsymbol{\alpha}_N^*)\,\middle|\,\mathcal{A}_N\right\}\right] = 0$, which is induced by the condition $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_1,y) = E\left\{\mathbf{U}_1(\boldsymbol{\beta}_{0N};\mathbf{x},y)\,\middle|\,\mathbf{x}_1,y\right\}$.

C. Proof of Theorem 2

Before proving Theorem 2, we need the following result.

Lemma A3. *Suppose that conditions C1, C3-C5 hold. Then, we have*

$$\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_N^* = O_p(n^{-1/2}).$$

Proof of Lemma A3. Since $\hat{\boldsymbol{\alpha}}_2$ is obtained by an independent external survey, we conclude that the variance of $\hat{\boldsymbol{\alpha}}^*$ can be estimated by $\left(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1}\right)^{-1}$. Thus, the order of the variance of $\hat{\boldsymbol{\alpha}}^*$ is determined by the less efficient estimator between $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$. If we showed

$$\hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_N^* = O_p(n^{-1/2}), \tag{C.1}$$

we could have $\mathbf{V}_1 = O_p(n^{-1})$ by (C5b). Since $\hat{\boldsymbol{\alpha}}^*$ is at least as efficient as $\hat{\boldsymbol{\alpha}}_1$, we have completed the proof of Lemma A3.

Thus, it remains to show (C.1). By C4 and C5, we have

$$\begin{aligned}
\mathbf{0} &= \sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\hat{\boldsymbol{\alpha}}_1;\mathbf{x}_{1i},y_i) \\
&= \sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i) + \left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\tilde{\boldsymbol{\alpha}};\mathbf{x}_{1i},y_i)\right\}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_N^*),
\end{aligned} \tag{C.2}$$

where $\tilde{\boldsymbol{\alpha}}$ lies on the segment joining $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_N^*$. By C3-C5 and (C.2), we conclude that

$$\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_N^* = -\boldsymbol{\mathcal{I}}_0^{-1}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i) + o_p(n^{-1/2}). \tag{C.3}$$

Thus, by C3 and (C.3), we have shown (C.1).

Proof of Theorem 2. Consider

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)\\
&\quad=\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+\left\{\frac{\partial}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\tilde{\boldsymbol{\alpha}};x_{1i},y_i\right)\right\}\left(\hat{\boldsymbol{\alpha}}^*-\boldsymbol{\alpha}_N^*\right)\\
&\quad=\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\\
&\qquad+\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_1^{-1}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_N^*\right)+o_p\left(n^{-1/2}\right),\\
&\quad=\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\\
&\qquad-\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_1^{-1}\boldsymbol{\mathcal{I}}_0^{-1}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+o_p\left(n^{-1/2}\right),\\
&\quad=\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)+\boldsymbol{\mathcal{I}}_0\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\boldsymbol{\mathcal{I}}_0^{-1}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\\
&\qquad+o_p\left(n^{-1/2}\right)\\
&\quad=\boldsymbol{\mathcal{I}}_0\mathbf{W}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)+\boldsymbol{\mathcal{I}}_0\mathbf{W}\boldsymbol{\mathcal{I}}_0^{-1}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)+o_p\left(\kappa(n)\right),
\end{aligned}\tag{C.4}$$

where $\tilde{\boldsymbol{\alpha}}$ lies on the segment joining $\hat{\boldsymbol{\alpha}}^*$ and $\boldsymbol{\alpha}_N^*$, the second equality holds by C4 and Lemma A3, the third equality holds by (C.3), the last equality holds by C5d, $\kappa(n)=\gamma(n)$ if $\gamma(n)\,n^{1/2}\to\infty$ and $\kappa(n)=n^{-1/2}$ otherwise, and $\gamma(n)$ is the convergence order of $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)$ in (C5e).

If there exists a non-stochastic matrix $\boldsymbol{\Sigma}_c$ such that $n\mathbf{V}_2=\boldsymbol{\Sigma}_c+o_p(1)$, then $\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)=O_p\left(n^{1/2}\right)$ and $\mathbf{W}$ is not a zero matrix. Then, by (C.4), we have

$$\begin{aligned}
&n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}\\
&\qquad=\begin{pmatrix}\mathbf{0}\\ n^{1/2}\boldsymbol{\mathcal{I}}_0\mathbf{W}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)\end{pmatrix}+n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \boldsymbol{\mathcal{I}}_0\mathbf{W}\boldsymbol{\mathcal{I}}_0^{-1}\mathbf{U}_2\left(\boldsymbol{\alpha}_N^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}+o_p(1).
\end{aligned}\tag{C.5}$$

Since the external sample is independent with the internal sample and $\boldsymbol{\Sigma}_c$ is the asymptotic variance of $n^{1/2}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right)$, by (C3), (C5e) and (C.5), we conclude that

$$n^{1/2}\sum_{i\in\mathcal{S}_1}\tilde{d}_i\begin{pmatrix}\mathbf{U}_1\left(\boldsymbol{\beta}_{0N};\mathbf{x}_i,y_i\right)\\ \mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)\end{pmatrix}\to\mathcal{N}\left(0,\tilde{\boldsymbol{\Sigma}}_u\right),$$

where $\tilde{\boldsymbol{\Sigma}}_{11}=\boldsymbol{\Sigma}_{11}$, $\tilde{\boldsymbol{\Sigma}}_{12}=\boldsymbol{\Sigma}_{12}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_0^{\mathrm{T}}$, $\tilde{\boldsymbol{\Sigma}}_{21}=\tilde{\boldsymbol{\Sigma}}_{12}^{\mathrm{T}}$, and $\tilde{\boldsymbol{\Sigma}}_{22}=\boldsymbol{\mathcal{I}}_0\mathbf{W}\left\{\boldsymbol{\Sigma}_c+\boldsymbol{\mathcal{I}}_0^{-1}\boldsymbol{\Sigma}_{22}\left(\boldsymbol{\mathcal{I}}_0^{-1}\right)^{\mathrm{T}}\right\}\mathbf{W}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_o^{\mathrm{T}}$. Thus, we have proved the first case of Theorem 2.

If $\mathbf{W}=0$, then $\gamma(n)\,n^{1/2}\to\infty$ and the rate of $\kappa(n)$ is slower than $n^{-1/2}$ in (C.4). Thus, the remainder term of (C.4) is no longer $o_p\left(n^{-1/2}\right)$ for $\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right)$. Instead, for this case, we investigate the asymptotic order of $\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}$ first. By C3, C5b and (C.3), we have

$$\mathbf{V}_1\asymp n^{-1}\quad\text{and}\quad\mathbf{V}_2\asymp\gamma(n)^{-2}\tag{C.6}$$

in probability. Thus, (C.6) leads to

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1} \asymp n^{-1}\gamma(n)^2 \tag{C.7}$$

in probability by the fact that $\gamma(n)\, n^{1/2} \to \infty$. Thus, by (C5e) and (C.7), we have

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right) \asymp n^{-1}\gamma(n) \tag{C.8}$$

in probability. By $\gamma(n)n^{1/2} \to \infty$, (C.7) and (C.8), we have shown

$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1} = o_p(1),$$
$$\left(\mathbf{V}_1^{-1}+\mathbf{V}_2^{-1}\right)^{-1}\mathbf{V}_2^{-1}\left(\hat{\boldsymbol{\alpha}}_2-\boldsymbol{\alpha}_N^*\right) = o_p\left(n^{-1/2}\right)$$

in probability. Thus, by the fourth equality of (C.4), we can show that

$$\sum_{i\in\mathcal{S}_1}\tilde{d}_i\mathbf{U}_2\left(\hat{\boldsymbol{\alpha}}^*;\mathbf{x}_{1i},y_i\right) = o_p\left(n^{-1/2}\right),$$

and we have proved the third case of Theorem 2.

References

Chatterjee, N., Chen, Y.H., Maas, P. and Carroll, R.J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111, 107-117.

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9(2), 385-406.

Chen, S., and Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24(1), 335-355.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Chen, Y.H., and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B*, 62, 449-460.

Csiszár, I., and Shields, P.C. (2004). Information theory and statistics: A tutorial. *Foundations and Trands in Communications and Information Theory*, 1, 417-528.

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

Engle, R.F., and McFadden, D.L. (1994). *Handbook of Econometrics: Volume IV*. Amsterdam.

Fuller, W.A. (2009). *Sampling Statistic*. Hoboken, New York: John Wiley & Sons, Inc.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

Han, P., and Lawless, J.F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29(3), 1321-1342.

Hidiroglou, M. (2001). Double sampling. *Survey Methodology*, 27, 2, 143-154. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6091-eng.pdf.

Horn, R.A., and Johnson, C.R. (2012). *Matrix Analysis*. Cambridge university press, New York.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20, 493-506.

Kim, J.K., and Rao, J.N.K. (2019). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*, 87, S177-S191.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3), 567-585.

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

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19, 1725-1747.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

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

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

Rivera-Rodriguez, C., Haneuse, S., Wang, M. and Spiegelman, D. (2020). Augmented pseudolikelihood estimation for two-phase studies. *Statistical Methods in Medical Research*, 29, 344-358.

Rivera-Rodriguez, C., Spiegelman, D. and Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38, 4611-4624.

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

Sheng, Y., Sun, Y., Huang, C.-Y. and Kim, M.-O. (2021). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*.

Sheng, Y., Sun, Y., Huang, C.-Y. and Kim, M.-O. (2022). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*, 78(2), 679-690.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. and Gail, M.H. (2020a). Weight calibration to improve efficiency for estimating pure risks from the additive hazards model with the nested case-control design. *Biometrics*, 1-13.

Shin, Y.E., Pfeiffer, R.M., Graubard, B.I. and Gail, M.H. (2020b). Weight calibration to improve the efficiency of pure risk estimates from case-control samples nested in a cohort. *Biometrics*, 76(4), 1087-1097.

Taylor, J.M.G., Choi, K. and Han, P. (2022). Data integration: Exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 1-16.

Wang, C.Y., Wang, S., Zhao, L.-P. and Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92, 512-525.

Wang, H., and Kim, J.K. (2021). Propensity score estimation using density ratio model under item nonresponse. *arXiv preprint arXiv:2104.13469*.

Wang, L., Williams, M.L., Chen, Y. and Chen, J. (2020). Novel two-phase sampling designs for studying binary outcomes. *Biometrics*, 76, 210-223.

Wu, C., and Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 34(3), 359-375.

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

Xu, M., and Shao, J. (2020). Meta-analysis of independent datasets using constrained generalised method of moments. *Statistical Theory and Related Fields*, 4, 109-116.

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

Yang, S., Zheng, D. and Wang, X. (2020). Elastic integrated analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579v2*.

Yuan, K.-H., and Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2), 245-260.

Zhai, Y., and Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics*, 0(0), 1-12.

Zhang, H., Deng, L., Wheeler, W., Qin, J. and Yu, K. (2021). Integrative analysis of multiple case-control studies. *Biometrics*, 1-12.

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

Zubizarreta, J.R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110, 910-922.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 117-138
Statistics Canada, Catalogue No. 12-001-X

One-sided testing of population domain means in surveys

Xiaoming Xu and Mary C. Meyer[1]

Abstract

Recent work in survey domain estimation allows for estimation of population domain means under *a priori* assumptions expressed in terms of linear inequality constraints. For example, it might be known that the population means are non-decreasing along ordered domains. Imposing the constraints has been shown to provide estimators with smaller variance and tighter confidence intervals. In this paper we consider a formal test of the null hypothesis that all the constraints are binding, versus the alternative that at least one constraint is non-binding. The test of constant versus increasing domain means is a special case. The power of the test is substantially better than the test with the same null hypothesis and an unconstrained alternative. The new test is used with data from the National Survey of College Graduates, to show that salaries are positively related to the subject's father's educational level, across fields of study and over several years of cohorts.

Key Words: Survey domain; Order constraints; Monotone; Block monotone.

1. Introduction

Methods for estimation of population domain means under *a priori* assumptions in the form of linear inequality constraints have been recently established. Suppose interest is in estimating $\bar{\mathbf{y}}_U \in \mathbb{R}^D$, a vector of population domain means, where D is the number of domains. Wu, Meyer and Opsomer (2016) derived an isotonic survey estimator of $\bar{\mathbf{y}}_U$, where it is assumed that $\bar{y}_{U_1} \leq \ldots \leq \bar{y}_{U_D}$. They showed that the constrained estimator is equivalent to a "pooled" estimator, where weighted averages of adjacent sample domain means are used to form an isotonic vector of domain mean estimates. Advantages to the ordered mean estimates are that they "make sense" in terms of satisfying the assumptions, and the confidence intervals for the estimates are typically reduced in length. Oliva-Aviles, Meyer and Opsomer (2019) proposed an information criterion to check the validity of the monotone assumption; that is, determining whether the domain means are ordered or unordered.

Oliva-Aviles, Meyer and Opsomer (2020) proposed a framework for estimation and inference with more general shape and order constraints in survey contexts. Examples include block orderings, and orderings of domain means arranged in grids. For example, average cholesterol level may be assumed to be increasing in age category and body mass index (BMI) level, but decreasing in exercise category. In another context, suppose average salary is to be estimated by job rank, job type, and location, with average salary assumed to be increasing with rank, and block orderings imposed on job type and location. More recently, Xu, Meyer and Opsomer (2021) formulated a mixture covariance matrix for constrained estimation that was shown to improve coverage of confidence intervals while retaining the smaller lengths.

1. Xiaoming Xu, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, U.S.A. E-mail: xiaoming.xu197@duke.edu; Mary C. Meyer, Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A. E-mail: mary.meyer@colostate.edu.

The desired linear inequality constraints may be formulated using an $M \times D$ constraint matrix $\mathbf{A}$, where the assumption is $\mathbf{A}\bar{\mathbf{y}}_U \geq \mathbf{0}$. For the isotonic domain means, $M = D - 1$, and the nonzero elements of the constraint matrix are $\{\mathbf{A}\}_{m,m} = -1$ and $\{\mathbf{A}\}_{m,m+1} = 1$. For block orderings, where domains are grouped by ordered blocks, each domain in block one, for example, is assumed to have a population mean not larger than each domain in block two, and in block two, each population domain mean does not exceed any of those in block three, etc. Here the number of constraints is $M = \sum_{b=1}^{B-1} \sum_{b'=b+1}^{B} D_b D_{b'}$, where B is the number of blocks and D_b is the number of domains in the b^{th} block, $b = 1, \ldots, B$. For example, suppose interest is in mean salaries at an institution, where the domains are four "fields", and it is assumed that fields 3 and 4 have higher salaries than fields 1 and 2. In this case $B = 2$, $D_1 = D_2 = 2$, and the constraint matrix is

$$\mathbf{A} = \begin{pmatrix} -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

For a third example, consider domains arranged in a grid; for a context suppose the population units are lakes in a state, and y_i is the level of a certain pollutant in lake i. We are interested in average levels by county and by distance from an industrial plant. If there are 60 counties and 5 categories of distance, there are 300 domains. If we know that the level of pollutant is non-increasing in the distance variable, then there are $60 \times 4 = 240$ constraints, formulated as antitonic within each county.

We propose a test where the null hypothesis is that $\mathbf{A}\bar{\mathbf{y}}_U = \mathbf{0}$, versus the alternative $\mathbf{A}\bar{\mathbf{y}}_U \geq \mathbf{0}$, and $\mathbf{A}\bar{\mathbf{y}}_U$ has at least one positive element. The simplest example is the null hypothesis of constant domain means, versus the alternative of increasing domain means. (Note that these hypotheses are different from the alternatives in Oliva-Aviles, Meyer and Opsomer (2019), who were deciding between monotone and non-monotone domain means.) For the industrial plant example above, we can test the null hypothesis that, within each county, the domain means are constant in distance. Using the constraints for a one-sided alternative results in improved power over the equivalent two-sided test.

This test has been widely studied outside of the survey context; see Bartholomew (1959); Bartholomew (1961); Chacko (1963); McDermott and Mudholkar (1993); Robertson, Wright and Dykstra (1988); Meyer (2003); Silvapulle and Sen (2005); Sen and Meyer (2017) and others. The null distribution of the likelihood-ratio test statistic for the one-sided test has been derived based on the normal-errors model. In brief, when the error terms are independently and identically distributed with known model variance, the null distribution of the likelihood ratio statistic is shown to be a mixture of chi-square distributions, while for the unknown model variance, the test statistic has the null distribution of a mixture of beta distributions. If the error terms are not independently and identically distributed, the results, based on principles of generalized least squares, still hold provided the covariance structure for the error terms is available. Similar results for the one-sided likelihood ratio test were obtained by Perlman (1969) where

the completely unknown covariance matrix was considered. Meyer and Wang (2012) formally proved that the one-sided test will provide higher power than the test using the unconstrained alternative.

In this paper we extend the one-sided test to the survey context. In the Section 2, the test is derived, and in Section 3 some large sample theory is given. Simulations in Section 4 show that the test performs well compared to the test with the unconstrained alternative, with better power and a test size closer to the target. In Section 5 the methods are applied to the National Survey of College Graduates (NSCG), to test whether salaries are higher for people whose father's education level is higher, controlling for field of study, highest degree attained, and year of degree. The test is available in the `R` package `csurvey`.

2. Formulation of the test statistic

To establish the notation, let $U=\{1,2,\ldots,N\}$ be the finite population. A sample $s\subset U$ of size n is to be drawn based on a probability sampling design p, where $p(s)$ is the probability of drawing the sample s. The first order inclusion probability $\pi_i=\Pr(i\in s)=\sum_{i\in s}p(s)$ and the second order inclusion probability $\pi_{ij}=\Pr(i,j\in s)=\sum_{i,j\in s}p(s)$, determined by the sampling design, are both assumed to be positive. The assumed positive π_i and π_{ij} ensure that the design-based estimator of the population parameter and the associated design-based variance estimator can be obtained, respectively. In terms of the domains of interest, let $\{U_d:d=1,\ldots,D\}$ be a partition of the population U and N_d be the population size of domain d, where D is the number of domains. We denote by s_d the intersection of s and U_d, and let n_d be the sample size for s_d. Sample size n_d arises from a random sampling procedure and thus is not fixed in general.

Let y be the variable of interest and denote by y_i the value for the i^{th} unit in the population. The population domain means are $\bar{\mathbf{y}}_U=\left(\bar{y}_{U_1},\ldots,\bar{y}_{U_D}\right)^{\mathsf{T}}$, and $\bar{y}_{U_d}$ is given by:

$$\bar{y}_{U_d}=\frac{\sum_{i\in U_d}y_i}{N_d}\quad d=1,\ldots,D.$$

Two common design-based estimators of the population means are the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) or the Hájek estimator (Hájek, 1971); because the Hájek estimator $\tilde{y}_{s_d}$ does not require information about the population domain size N_d and has other advantages in practice, we will focus on the Hájek estimator. The results for the Horvitz-Thompson estimator, however, can be derived analogously. The Hájek estimator for domain means is $\tilde{\mathbf{y}}_s=\left(\tilde{y}_{s_1},\ldots,\tilde{y}_{s_D}\right)$, where

$$\tilde{y}_{s_d}=\frac{\sum_{i\in s_d}y_i/\pi_i}{\hat{N}_d}$$

and $\hat{N}_d=\sum_{i\in s_d}1/\pi_i$.

We are concerned with testing:

$$H_0 : \overline{\mathbf{y}}_U \in V \quad \text{versus} \quad H_1 : \overline{\mathbf{y}}_U \in C \setminus V \tag{2.1}$$

where $V = \{\mathbf{y} : \mathbf{A}\mathbf{y} = \mathbf{0}\}$ is the null space of $\mathbf{A}$ and the alternative set is the convex cone $C = \{\mathbf{y} : \mathbf{A}\mathbf{y} \geq \mathbf{0}\}$ excluding the set V. A set C is a convex cone if for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in C, $\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2$ is in C for any non-negative α_1 and α_2.

We start with a brief review of the properties of the unconstrained estimator $\tilde{\mathbf{y}}_s$. By the Taylor expansion, we can linearize the $\tilde{\mathbf{y}}_s$ as follows:

$$\tilde{\mathbf{y}}_s = \overline{\mathbf{y}}_U + \hat{\mathbf{y}}^{\text{center}} + O_p\left(n^{-1}\right)$$

where

$$\hat{\mathbf{y}}^{\text{center}} = \left(\frac{1}{N_1} \sum_{i \in s_1} \frac{\left(y_i - \overline{y}_{U_1}\right)}{\pi_i}, \ldots, \frac{1}{N_D} \sum_{i \in s_D} \frac{\left(y_i - \overline{y}_{U_D}\right)}{\pi_i} \right)^{\mathrm{T}}.$$

The properties of $\tilde{\mathbf{y}}_s - \overline{\mathbf{y}}_U$ can be approximated by $\hat{\mathbf{y}}^{\text{center}}$ and we have that $\mathrm{E}\left(\hat{\mathbf{y}}^{\text{center}}\right) = 0$ and the variance of $\hat{\mathbf{y}}^{\text{center}}$ is $\boldsymbol{\Sigma}$, where the dd'^{th} element of $\boldsymbol{\Sigma}$ is:

$$\{\boldsymbol{\Sigma}\}_{dd'} = \frac{1}{N_d N_{d'}} \sum_{i \in U_d} \sum_{j \in U_{d'}} \Delta_{ij} \frac{\left(y_i - \overline{y}_{U_d}\right)\left(y_j - \overline{y}_{U_{d'}}\right)}{\pi_i \pi_j}, \quad d, d' = 1, 2, \ldots, D$$

where $\Delta_{ij} = \mathrm{cov}\left(I_i, I_j\right) = \pi_{ij} - \pi_i \pi_j$ and I_i is the indicator variable of whether unit i is selected by sampling design. By the design normal assumption (A5) in the appendix, we have $\boldsymbol{\Sigma}^{-1/2}\hat{\mathbf{y}}^{\text{center}} \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \mathbf{I}\right)$, hence:

$$\boldsymbol{\Sigma}^{-1/2}\left(\tilde{\mathbf{y}}_s - \overline{\mathbf{y}}_U\right) = \boldsymbol{\Sigma}^{-1/2}\hat{\mathbf{y}}^{\text{center}} + o_p(1) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \mathbf{I}\right).$$

We denote by $\hat{\boldsymbol{\Sigma}}$ a consistent estimator of $\boldsymbol{\Sigma}$, in the sense that $n\left(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right) = o_p(1)$. For testing (2.1), we propose the following weighted least squares test statistic:

$$\hat{T} = \frac{\min_{\boldsymbol{\theta}_0 \in V}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right) - \min_{\boldsymbol{\theta}_1 \in C}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)}{\min_{\boldsymbol{\theta}_0 \in V}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)}.$$

Assuming the second order inclusion probability π_{ij} to be known, the dd'^{th} element of the design based consistent estimator $\hat{\boldsymbol{\Sigma}}$ has the following expression:

$$\{\hat{\boldsymbol{\Sigma}}\}_{dd'} = \frac{1}{\hat{N}_d \hat{N}_{d'}} \sum_{i \in s_d} \sum_{j \in s_{d'}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\left(y_i - \tilde{y}_{s_d}\right)\left(y_j - \tilde{y}_{s_{d'}}\right)}{\pi_i \pi_j}, \quad d, d' = 1, 2, \ldots, D. \tag{2.2}$$

See Särndal, Swensson and Wretman (1992) Chapter 5 on page 185 for more details. Particularly, under a fixed size design, the Sen-Yates-Grundy variance estimator, derived as an alternative form of (2.2), can

also be used. In addition, under many complex survey designs, the second order inclusion probability π_{ij} might be zero or unknown so that the design based covariance estimator $\hat{\boldsymbol{\Sigma}}$ cannot be obtained. In such cases, the use of consistent replication-based variance estimators (such as Jackknife estimator, bootstrap estimator) can be considered, since the calculation of replication variance estimator does not involve the second order inclusion probabilities. As long as the replication-based variance estimators are good approximation for $\boldsymbol{\Sigma}$, the asymptotic properties of $\hat{T}$, which will be developed shortly, would hold.

We will reject H_0 if $\hat{T}$ is large. This is similar in structure to the classical test (as was presented in, for example, Silvapulle and Sen (2005) Chapter 3). If $\tilde{\mathbf{y}}_s$ were normal with $\text{cov}(\tilde{\mathbf{y}}_s) = \hat{\boldsymbol{\Sigma}}$, then $\hat{T}$ would be distributed as a mixture of beta random variables, under the null hypothesis. In the survey context, we approximate the distribution of $\hat{T}$.

3. Asymptotic distribution of the test statistic

The assumptions needed to derive an approximate distribution of $\hat{T}$ are listed in Appendix, and are similar to those in Xu et al. (2021).

To derive the asymptotic null distribution of $\hat{T}$, we first show the following result.

Lemma 1. *The test statistic* $\hat{T}$ *can be written as:*

$$\begin{aligned}\hat{T} &= \frac{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0) - \min_{\boldsymbol{\theta}_1 \in C}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)} \\ &= \frac{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0) - \min_{\boldsymbol{\theta}_1 \in C}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)} + o_p(1).\end{aligned}$$

Proof. Let $\hat{\mathbf{A}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}^{1/2}$, $\hat{\mathbf{Z}}_s = \hat{\boldsymbol{\Sigma}}^{-1/2}\tilde{\mathbf{y}}_s$, $\hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\theta}_1$, $\hat{\mathbf{V}} = \{\hat{\boldsymbol{\theta}}_0 : \hat{\mathbf{A}}\hat{\boldsymbol{\theta}}_0 = 0\}$ and $\hat{C} = \{\hat{\boldsymbol{\theta}}_1 : \hat{\mathbf{A}}\hat{\boldsymbol{\theta}}_1 \geq 0\}$. Then by a transformation, we have:

$$\begin{aligned}\hat{T} &= \frac{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0) - \min_{\boldsymbol{\theta}_1 \in C}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)} \\ &= \frac{\min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0) - \min_{\hat{\boldsymbol{\theta}}_1 \in \hat{C}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1)}{\min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)} \\ &= 1 - \frac{\min_{\hat{\boldsymbol{\theta}}_1 \in \hat{C}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1)}{\min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)}.\end{aligned}$$

Let $\hat{V}^{\perp}$ be the linear space of vectors in $\mathbb{R}^D$ that are orthogonal to vectors in $\hat{V}$. Note that $\min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)^{\mathrm{T}}(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0)$ is the squared length of the projection of $\hat{\mathbf{Z}}_s$ onto $\hat{V}^{\perp}$ and the projection of

$\hat{\mathbf{Z}}_s$ onto $\hat{V}$ has the explicit expression $\hat{\boldsymbol{\theta}}_0^* = \left(\mathbf{I} - \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\right)\hat{\mathbf{Z}}_s$, where $\left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^-$ is the generalized inverse of $\hat{\mathbf{A}}\hat{\mathbf{A}}^\top$. Hence, by the consistency of $\hat{\boldsymbol{\Sigma}}$, we have the following:

$$\begin{aligned}
\min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}} \frac{1}{n}\left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0\right)^\top \left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0\right) &= \frac{1}{n}\left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0^*\right)^\top \left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0^*\right) \\
&= \frac{1}{n}\left(\hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\hat{\mathbf{Z}}_s\right)^\top \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\hat{\mathbf{Z}}_s \\
&= \frac{1}{n}\hat{\mathbf{Z}}_s^\top \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\hat{\mathbf{Z}}_s \\
&= \frac{1}{n}\tilde{\mathbf{y}}_s^\top \mathbf{A}^\top \left(\mathbf{A}\hat{\boldsymbol{\Sigma}}\mathbf{A}^\top\right)^- \mathbf{A}\tilde{\mathbf{y}}_s \\
&= \tilde{\mathbf{y}}_s^\top \mathbf{A}^\top \left(\mathbf{A}n\boldsymbol{\Sigma}\mathbf{A}^\top\right)^- \mathbf{A}\tilde{\mathbf{y}}_s + o_p(1) \\
&= \min_{\boldsymbol{\theta}_0 \in V} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right) + o_p(1). \qquad (3.1)
\end{aligned}$$

By (3.1) and the result that $\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) - \min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) = o_p(1)$ by Lemma 4 in the Appendix, we get

$$\begin{aligned}
\hat{T} &= 1 - \frac{\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)}{\min_{\boldsymbol{\theta}_0 \in V} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)} \\
&= 1 - \frac{\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)}{\min_{\boldsymbol{\theta}_0 \in V} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)} + o_p(1)
\end{aligned}$$

the proof is complete.

The denominator in above expression must be bounded away from zero in probability, which is indeed the case because it can be shown that the $\min_{\boldsymbol{\theta}_0 \in V}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top (n\boldsymbol{\Sigma})^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)$ has, asymptotically, $\chi^2(M)$ distribution under the null and design normal assumption.

Next, let $\tilde{\mathbf{Z}}_s = \boldsymbol{\Sigma}^{-1/2}\tilde{\mathbf{y}}_s, \mathbf{Z}_U = \boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{y}}_U, \tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_1 = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}_1$ and define $\tilde{V} = \left\{\tilde{\boldsymbol{\theta}}: \tilde{\mathbf{A}}\tilde{\boldsymbol{\theta}} = 0\right\}, \tilde{C} = \left\{\tilde{\boldsymbol{\theta}}: \tilde{\mathbf{A}}\tilde{\boldsymbol{\theta}} \geq 0\right\}$, where $\tilde{\mathbf{A}} = \mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Then, we have the following main result of the paper.

Theorem 1. *Define*

$$T = \frac{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \mathbf{Z} - \tilde{\boldsymbol{\theta}}_0 \right\|^2 \min_{\tilde{\boldsymbol{\theta}}_1 \in \tilde{C}} \left\| \mathbf{Z} - \tilde{\boldsymbol{\theta}}_1 \right\|^2}{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \mathbf{Z} - \tilde{\boldsymbol{\theta}}_0 \right\|^2}$$

where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$. *Then under the null,* $\hat{T}$ *converges in distribution to* T. *That is,*

$$\hat{T} \xrightarrow{\mathcal{D}} T.$$

Proof. According to the transformation above, we can express $\hat{T}$ as:

$$\hat{T} = \frac{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \tilde{\mathbf{Z}}_s - \tilde{\boldsymbol{\theta}}_0 \right\|^2 - \min_{\tilde{\boldsymbol{\theta}}_1 \in \tilde{C}} \left\| \tilde{\mathbf{Z}}_s - \tilde{\boldsymbol{\theta}}_1 \right\|^2}{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \tilde{\mathbf{Z}}_s - \tilde{\boldsymbol{\theta}}_0 \right\|^2} + o_p(1)$$

$$= \frac{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \tilde{\mathbf{Z}}_s - \mathbf{Z}_U + \mathbf{Z}_U - \tilde{\boldsymbol{\theta}}_0 \right\|^2 - \min_{\tilde{\boldsymbol{\theta}}_1 \in \tilde{C}} \left\| \tilde{\mathbf{Z}}_s - \mathbf{Z}_U + \mathbf{Z}_U - \tilde{\boldsymbol{\theta}}_1 \right\|^2}{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \tilde{\mathbf{Z}}_s - \mathbf{Z}_U + \mathbf{Z}_U - \tilde{\boldsymbol{\theta}}_0 \right\|^2} + o_p(1)$$

$$= \frac{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \mathbf{Z}^{\text{center}} - \tilde{\boldsymbol{\theta}}_0 \right\|^2 - \min_{\tilde{\boldsymbol{\theta}}_1 \in \tilde{C}} \left\| \mathbf{Z}^{\text{center}} - \tilde{\boldsymbol{\theta}}_1 \right\|^2}{\min_{\tilde{\boldsymbol{\theta}}_0 \in \tilde{V}} \left\| \mathbf{Z}^{\text{center}} - \tilde{\boldsymbol{\theta}}_0 \right\|^2} + o_p(1)$$

where $\mathbf{Z}^{\text{center}} = \tilde{\mathbf{Z}}_s - \mathbf{Z}_U$, and recall that under H_0, $\mathbf{Z}_U \in \tilde{V}$, so that, in the above expression, minimizing over $\tilde{\boldsymbol{\theta}}_0$ is equivalent to minimizing over $-\mathbf{Z}_U + \tilde{\boldsymbol{\theta}}_0$, and similarly for minimizing over $\tilde{\boldsymbol{\theta}}_1$.

Then, we have $\hat{T} \xrightarrow{\mathcal{D}} T$. This follows from the Lipschitz continuity of the projection of $\mathbf{Z}$ onto a convex cone; that is, if $\hat{\boldsymbol{\theta}}$ is the projection of $\mathbf{Z}$ onto the cone C, then $\hat{\boldsymbol{\theta}}$ is a continuous function of $\mathbf{Z}$; see Proposition 1 and its proof in Meyer and Woodroofe (2000).

The random variable T defined in Theorem 1 has been shown to be distributed as a mixture of beta random variables under H_0. See Robertson et al. (1988) in Chapter 2 and Meyer (2003) for more details. Also, the mixing distribution can be found (to within a desired precision) via simulation. Specifically, if $M_0 \leq M$ is the rank of the constraint matrix $\mathbf{A}$,

$$\Pr(T \leq c) = \sum_{m=0}^{M_0} \Pr\left\{ \text{Be}\left(\frac{M_0 - m}{2}, \frac{m}{2} \right) \leq c \right\} p_m,$$

where the mixing probabilities $p_0, \ldots, p_{M_0}$ are approximated through simulations, and $\text{Be}(\alpha, \beta)$ represents a Beta random variable with parameters α and β, respectively. By convention, $\text{Be}(0, \beta) = 0$ and $\text{Be}(\alpha, 0) = 1$.

If m is the dimension of the space spanned by the rows of $\mathbf{A}$ that represent binding constraints, then each p_m represents the probability that m constraints are binding, $m = 0, \ldots, M_0$. Each row of $\hat{\mathbf{A}}$ represents a constraint, and we say that the j^{th} constraint is binding if the j^{th} element of $\hat{\mathbf{A}}\hat{\boldsymbol{\theta}}$ is zero. The quantity $D - m$, where m is the number of binding constraints, can be thought of as the observed degrees of freedom of the fit. For more information about this mixing distribution, see Silvapulle and Sen (2005), Chapter 3. The mixing probabilities are approximated as follows:

(1) Generate $\mathbf{Z}$ from a standard multivariate normal distribution $N(\mathbf{0}, \mathbf{I})$.

(2) Project the generated $\mathbf{Z}$ onto the convex cone $\hat{C} = \{\boldsymbol{\theta} : \hat{\mathbf{A}}\boldsymbol{\theta} \geq 0\}$ to obtain the J set, where $\hat{\mathbf{A}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}^{1/2}$. Specifically, let $\hat{\boldsymbol{\theta}}$ be the projection of $\mathbf{Z}$ onto the $\hat{C}$, then $J = \{j : \hat{\mathbf{A}}_j \hat{\boldsymbol{\theta}} = 0\}$, where $\hat{\mathbf{A}}_j$ is the j^{th} row of $\hat{\mathbf{A}}$. That is, J indexes the set of "binding constraints". The `R`

package `coneproj` (Liao and Meyer (2014)) finds $\hat{\boldsymbol{\theta}}$ given the generated $\mathbf{Z}$ and $\hat{\mathbf{A}}$, and also returns the set of binding constraints J.

(3) Repeat the previous steps R times (say $R = 1{,}000$).

(4) Estimate p_m by the proportion of times that the set J has m elements, $m = 0, 1, \ldots, M_0$. When the matrix $\mathbf{A}$ has more constraints than dimensions, then, the cone projection routine in `coneproj` can always find a minimal unique J set. (See Meyer (2013) for details.)

3.1 The properties of asymptotic power of the test

In this section, we prove consistency and monotonicity of the power function of this test. First, we show that if the alternative hypothesis is true, then the probability of rejecting the null hypothesis increases to one as N and n increase without bound.

Theorem 2. *Let α be the test size and c_α be the corresponding critical value of the test. Then, the power of the test converges to 1 under the alternative. That is:*

$$P\left(\hat{T} > c_\alpha \,\middle|\, \overline{\mathbf{y}}_U \in \mathcal{C} \setminus V\right) \to 1, \quad as \quad N \to \infty.$$

Proof. Since $\hat{T} = 1 - \frac{\min_{\boldsymbol{\theta}_1 \in C} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)}$, it suffices to show that:

$$\frac{\min_{\boldsymbol{\theta}_1 \in C} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)}{\min_{\boldsymbol{\theta}_0 \in V} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)} = o_p(1)$$

under the the alternative. For the numerator, we have

$$\begin{aligned} \min_{\boldsymbol{\theta}_1 \in C} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) &\le \left(\tilde{\mathbf{y}}_s - \overline{\mathbf{y}}_U\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\tilde{\mathbf{y}}_s - \overline{\mathbf{y}}_U\right) \\ &= O_p\left(\frac{1}{\sqrt{n}}\right) O_p(n)\, O_p\left(\frac{1}{\sqrt{n}}\right) = O_p(1) \end{aligned}$$

where we use the fact that $\tilde{\mathbf{y}}_s - \overline{\mathbf{y}}_U = O_p\left(n^{-1/2}\right)$ and $\hat{\boldsymbol{\Sigma}} = O_p\left(n^{-1}\right)$ element-wise. For the denominator, we have:

$$\begin{aligned} \min_{\boldsymbol{\theta}_0 \in V} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0\right) &= \min_{\hat{\boldsymbol{\theta}}_0 \in \hat{V}} \left\| \hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_0 \right\|^2 \\ &= \left[\hat{\mathbf{Z}}_s - \left(\mathbf{I} - \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\right) \hat{\mathbf{Z}}_s\right]^\top \left[\hat{\mathbf{Z}}_s - \left(\mathbf{I} - \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\right) \hat{\mathbf{Z}}_s\right] \\ &= \hat{\mathbf{Z}}_s^\top \hat{\mathbf{A}}^\top \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^- \hat{\mathbf{A}}\hat{\mathbf{Z}}_s \\ &= \tilde{\mathbf{y}}_s^\top \mathbf{A}^\top \left(\mathbf{A}\hat{\boldsymbol{\Sigma}}\mathbf{A}^\top\right)^- \mathbf{A}\tilde{\mathbf{y}}_s. \end{aligned}$$

Hence, we have:

$$\frac{\min_{\boldsymbol{\theta}_1 \in C} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)} = \frac{\min_{\boldsymbol{\theta}_1 \in C} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)^\top (n\hat{\boldsymbol{\Sigma}})^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1)}{\min_{\boldsymbol{\theta}_0 \in V} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)^\top (n\hat{\boldsymbol{\Sigma}})^{-1} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_0)}$$

$$= O_p(n^{-1}) \frac{1}{\tilde{\mathbf{y}}_s^\top \mathbf{A}^\top (\mathbf{A} n \hat{\boldsymbol{\Sigma}} \mathbf{A}^\top)^- \mathbf{A} \tilde{\mathbf{y}}_s}$$

$$= O_p(n^{-1}) \frac{1}{\bar{\mathbf{y}}_U^\top \mathbf{A}^\top (\mathbf{A} n \boldsymbol{\Sigma} \mathbf{A}^\top)^- \mathbf{A} \bar{\mathbf{y}}_U + o_p(1)}$$

$$= O_p(n^{-1}) O_p(1) = o_p(1)$$

because $\tilde{\mathbf{y}}_s$ and $\hat{\boldsymbol{\Sigma}}$ are consistent for $\bar{\mathbf{y}}_U$ and $\boldsymbol{\Sigma}$ respectively. Therefore, under the alternative, $\hat{T}$ goes to 1 asymptotically.

4. Simulation studies

The simulations involve one or two dimensional grids, with several constraints and population domain means. We present the results in table form from three scenarios: for each, we record the proportions of times the null is rejected in various cases, with different sample sizes, significance levels and the variances for generating the study variables. In each case, we generate a population of size N, then we draw 10,000 samples from the population according to a sampling design. For each sample, we compute the test statistic value and the estimated covariance matrix. We compare the test statistic with the critical values under different significance levels, where the critical values are obtained from the asymptotic null distribution of the test statistics. Further, we compare the power of this one-sided test with that of ANOVA F test using the unconstrained alternative. That is,

$$H_0 : \mathbf{A}\bar{\mathbf{y}}_U = \mathbf{0} \quad \text{versus} \quad H_2 : \mathbf{A}\bar{\mathbf{y}}_U \neq \mathbf{0}.$$

Here, we use `svyglm` function in `survey` package to fit the ANOVA model and compute the P-values of the ANOVA F test by applying the `anova` function in `survey` package.

4.1 Monotonicity in one variable

As in Xu et al. (2021) and Oliva-Aviles et al. (2020), the limiting domain means for generating the study variables are given by the functions as follows:

$$\mu_d^{(0)} \equiv 1, \quad \mu_d^{(1)} = \frac{\exp(12d/D - 6)}{3.5\left(1 + \exp(12d/D - 6)\right)}, \quad \mu_d^{(2)} = \frac{\exp(12d/D - 6)}{2.5\left(1 + \exp(12d/D - 6)\right)}$$

for $d = 1, 2, \ldots, D,$ where $D = 12$ is the number of domains. The study variables $y_1, \ldots, y_N$ are generated by adding independent and identically distributed $N\left(0, \sigma_i^2\right)\ (i = 1, 2)$ errors to the μ_d values from above three functions, respectively, with $\sigma_1 = 1$ and $\sigma_2 = 1.5.$ We compare the test size and power for the test of constant versus increasing domain means, with the standard ANOVA test of constant versus non-constant domain means. Notice that under $\boldsymbol{\mu}^{(0)} = \left(\mu_1^{(0)}, \ldots, \mu_D^{(0)}\right)^{\mathrm{T}}$, the null hypothesis is true, while under $\boldsymbol{\mu}^{(0)} = \left(\mu_1^{(1)}, \ldots, \mu_D^{(1)}\right)^{\mathrm{T}}$ and $\boldsymbol{\mu}^{(2)} = \left(\mu_1^{(2)}, \ldots, \mu_D^{(2)}\right)^{\mathrm{T}}$, the population domain means have increasing order and thus the alternative is true, with $\boldsymbol{\mu}^{(2)}$ having larger effect size.

We draw the samples from a stratified random sampling design without replacement, with $H = 4$ strata that cut across the D domains. The strata are determined using an auxiliary variable $z,$ which is correlated with study variable $y.$ The values of z are created by adding i.i.d. standard normal errors to $\left(d/D\right).$ By ranking the values of $z,$ we can create 4 blocks of N/H elements. Then, the stratum membership of the population element is determined by the corresponding ranked $z.$ Finally, the population sizes are set to be $N = 9{,}600,$ $N = 19{,}200,$ $N = 57{,}600$ and $N = 76{,}800$ with population domain size $N_d = N/D$ for $d = 1, \ldots, D.$ The total sample sizes $n = 200,$ $n = 400,$ $n = 1{,}200$ and $n =$ 1,600 are assigned to the 4 strata with sample size (25, 50, 50, 75), (50, 100, 100, 150), (150, 300, 300, 450), (200, 400, 400, 600) in each stratum, respectively.

The results in Table 4.1 show that the test size for the proposed one-sided test is closer to the target, while the two-sided test size is somewhat inflated even for the larger sample sizes. For the simulations where the alternative hypothesis is true, the one-sided test has substantially higher power.

Table 4.1
Monotonicity in one variable: the proportions of times null is rejected under various settings and power comparison between the constrained one-sided test (top half) and the unconstrained test (bottom half)

	σ	n	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
			$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$
One-sided test	$\sigma = 1$	n = 200	0.0996	0.4689	0.6686	0.0533	0.3218	0.5055	0.0134	0.1194	0.2230
		n = 400	0.0840	0.6352	0.8529	0.0403	0.4780	0.7268	0.0085	0.2028	0.4054
		n = 1,200	0.1039	0.9657	0.9986	0.0537	0.9027	0.9941	0.0121	0.6444	0.9133
		n = 1,600	0.0981	0.9867	0.9999	0.0489	0.9550	0.9988	0.0110	0.7533	0.9654
	$\sigma = 1.5$	n = 200	0.0994	0.3128	0.4370	0.0528	0.2008	0.2938	0.0133	0.0625	0.1056
		n = 400	0.0839	0.4101	0.5946	0.0402	0.2740	0.4338	0.0084	0.0873	0.1770
		n = 1,200	0.1037	0.7838	0.9461	0.0532	0.6327	0.8679	0.0120	0.3142	0.5773
		n = 1,600	0.0980	0.8544	0.9751	0.0488	0.7253	0.9334	0.0109	0.3900	0.6928
ANOVA F test	$\sigma = 1$	n = 200	0.1412	0.2677	0.4017	0.0746	0.1627	0.2685	0.0147	0.0457	0.0973
		n = 400	0.1280	0.3618	0.6034	0.0658	0.2385	0.4627	0.0147	0.0835	0.2259
		n = 1,200	0.1123	0.8139	0.9854	0.0590	0.7121	0.9694	0.0117	0.4736	0.8943
		n = 1,600	0.1111	0.9253	0.9986	0.0576	0.8633	0.9964	0.0126	0.6868	0.9814
	$\sigma = 1.5$	n = 200	0.1412	0.1909	0.2502	0.0746	0.1087	0.1495	0.0147	0.0261	0.0408
		n = 400	0.1280	0.2195	0.3278	0.0658	0.1296	0.2094	0.0147	0.0313	0.0661
		n = 1,200	0.1123	0.4670	0.7538	0.0590	0.3320	0.6361	0.0117	0.1397	0.3902
		n = 1,600	0.1111	0.5947	0.8795	0.0576	0.4602	0.8014	0.0126	0.2367	0.5932

4.2 Block monotonic in one variable

In "block monotonic" ordering case, we assume the population means are ordered among blocks, but there is no ordering imposed within the blocks. Specifically, we organize the limiting domain means in four blocks of three domains as following:

$$\boldsymbol{\mu}^{(0)} = (0.05\ \ 0.05\ \ 0.05 \mid 0.05\ \ 0.05\ \ 0.05 \mid 0.05\ \ 0.05\ \ 0.05 \mid 0.05\ \ 0.05\ \ 0.05)$$

$$\boldsymbol{\mu}^{(1)} = (-0.06\ \ 0\ \ 0.06 \mid 0.12\ \ 0.06\ \ 0.18 \mid 0.18\ \ 0.24\ \ 0.30 \mid 0.30\ \ 0.36\ \ 0.30)$$

$$\boldsymbol{\mu}^{(2)} = (-0.08\ \ 0\ \ 0.08 \mid 0.16\ \ 0.08\ \ 0.24 \mid 0.24\ \ 0.32\ \ 0.40 \mid 0.40\ \ 0.48\ \ 0.40)$$

where the blocks are separated by the vertical lines. Hence, under the alternative, we expect the population mean for each of the domains in block b would be at least as large as those in block $b-1$, for $b = 2, 3, 4$. The effect size of $\bar{\mathbf{y}}_U^{(2)}$ generated from $\boldsymbol{\mu}^{(2)}$ would be larger than that of $\bar{\mathbf{y}}_U^{(1)}$ from $\boldsymbol{\mu}^{(1)}$. We use the same stratified simple random sampling design as in the previous example.

The results in Table 4.2 show again that one-sided test has substantially higher power for simulations where the alternative is true, and for simulations under the null hypothesis, the test size is approximately correct for the one-sided test and the two-sided ANOVA test has inflated test size.

Table 4.2
Block monotonicity in one variable: the proportions of times null is rejected under various settings and power comparison between the constrained one-sided test (top half) and the unconstrained test (bottom half)

	σ	n	$\alpha_1 = 0.1$			$\alpha_2 = 0.05$			$\alpha_3 = 0.01$		
			$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(0)}$	$\mu^{(1)}$	$\mu^{(2)}$
One-sided test	$\sigma = 1$	n = 200	0.1013	0.5114	0.6795	0.0568	0.3590	0.5216	0.0119	0.1397	0.2391
		n = 400	0.1036	0.7368	0.8856	0.0534	0.5838	0.7878	0.0109	0.2840	0.4722
		n = 1,200	0.0964	0.9718	0.9978	0.0487	0.9224	0.9880	0.0089	0.6671	0.8801
		n = 1,600	0.0976	0.9877	0.9998	0.0492	0.9635	0.9958	0.0098	0.7668	0.9339
	$\sigma = 1.5$	n = 200	0.1014	0.3421	0.4535	0.0567	0.2191	0.3124	0.0117	0.0731	0.1144
		n = 400	0.1031	0.4992	0.6616	0.0534	0.3544	0.5028	0.0109	0.1335	0.2235
		n = 1,200	0.0965	0.8187	0.9422	0.0485	0.6794	0.8672	0.0091	0.3474	0.5661
		n = 1,600	0.0974	0.8830	0.9743	0.0497	0.7652	0.9232	0.0099	0.4367	0.6746
ANOVA F test	$\sigma = 1$	n = 200	0.1412	0.2941	0.4368	0.0746	0.1847	0.2951	0.0147	0.0551	0.1155
		n = 400	0.1280	0.4220	0.6556	0.0658	0.2912	0.5231	0.0147	0.1123	0.2712
		n = 1,200	0.1123	0.8940	0.9921	0.0590	0.8177	0.9840	0.0117	0.6099	0.9363
		n = 1,600	0.1111	0.9678	0.9995	0.0576	0.9293	0.9986	0.0126	0.8094	0.9911
	$\sigma = 1.5$	n = 200	0.1412	0.2052	0.2611	0.0746	0.1173	0.1583	0.0147	0.0281	0.0431
		n = 400	0.1280	0.2445	0.3543	0.0658	0.1457	0.2333	0.0147	0.0389	0.0787
		n = 1,200	0.1123	0.5399	0.8012	0.0590	0.4099	0.6932	0.0117	0.1926	0.4549
		n = 1,600	0.1111	0.6799	0.9091	0.0576	0.5539	0.8468	0.0126	0.3153	0.6589

4.3 Monotonicity in two variables

Here we take into consideration a grid of domains, which represent two variables. The null hypothesis is that the population domain means are constant in one of the variables, and the alternative is that the

population means are increasing in that variable, while the domain means unconstrained in the other variable. In other words, we test for monotonicity in one variable while "controlling for" the effects of the other. In particular, we set the limiting domain means as follows:

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \end{pmatrix},$$

while

$$\boldsymbol{\mu}^{(1)} = \begin{pmatrix} 0 & 0.04 & 0.16 & 0.24 & 0.28 \\ 0.04 & 0.08 & 0.20 & 0.32 & 0.40 \\ 0.04 & 0.12 & 0.12 & 0.20 & 0.28 \\ 0.04 & 0.04 & 0.12 & 0.24 & 0.28 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\mu}^{(2)} = \begin{pmatrix} 0 & 0.05 & 0.20 & 0.30 & 0.35 \\ 0.05 & 0.10 & 0.25 & 0.40 & 0.50 \\ 0.05 & 0.15 & 0.15 & 0.25 & 0.35 \\ 0.05 & 0.05 & 0.15 & 0.30 & 0.35 \end{pmatrix}.$$

The sampling mechanism and the way we generate the study variable y are the same as that in one dimensional case. However, because there are more domains in this case, we set the sample size to be $n = 400$, $n = 800$, $n = 1{,}200$ and $n = 2{,}000$, respectively, corresponding to the population size $N = 8{,}000$, $N = 16{,}000$, $N = 24{,}000$ and $N = 40{,}000$, where the sample sizes are divided among the strata as $(50, 100, 100, 150)$, $(100, 200, 200, 300)$, $(150, 300, 300, 450)$ and $(250, 500, 500, 750)$, respectively. The simulation results in Table 4.3 demonstrate similar properties as those in the previous scenarios: the tests have higher power as sample size gets larger and the effect size of the population domain means is larger.

Table 4.3
Monotonicity in two variables: the proportions of times null is rejected under various settings and power comparison between the constrained one-sided test (top half) and the unconstrained test (bottom half)

	σ	n	$\alpha_1 = 0.1$			$\alpha_2 = 0.05$			$\alpha_3 = 0.01$		
			$\boldsymbol{\mu}^{(0)}$	$\boldsymbol{\mu}^{(1)}$	$\boldsymbol{\mu}^{(2)}$	$\boldsymbol{\mu}^{(0)}$	$\boldsymbol{\mu}^{(1)}$	$\boldsymbol{\mu}^{(2)}$	$\boldsymbol{\mu}^{(0)}$	$\boldsymbol{\mu}^{(1)}$	$\boldsymbol{\mu}^{(2)}$
One-sided test	$\sigma = 1$	n = 400	0.1770	0.7738	0.8755	0.1000	0.6415	0.7757	0.0255	0.3460	0.4907
		n = 800	0.1203	0.8732	0.9576	0.0590	0.7677	0.8975	0.0129	0.4706	0.6598
		n = 1,200	0.1097	0.9571	0.9921	0.0562	0.8972	0.9762	0.0102	0.6556	0.8523
		n = 2,000	0.1093	0.9929	0.9994	0.0558	0.9794	0.9975	0.0103	0.8661	0.9700
	$\sigma = 1.5$	n = 400	0.1778	0.5837	0.6840	0.1006	0.4301	0.5382	0.0255	0.1844	0.2586
		n = 800	0.1210	0.6512	0.7783	0.0594	0.4967	0.6399	0.0133	0.2257	0.3421
		n = 1,200	0.1098	0.7701	0.8908	0.0565	0.6247	0.7881	0.0100	0.3235	0.4909
		n = 2,000	0.1089	0.9019	0.9725	0.0560	0.8040	0.9292	0.0103	0.5150	0.7236
ANOVA F test	$\sigma = 1$	n = 400	0.1584	0.4337	0.5642	0.0828	0.3005	0.4255	0.0184	0.1075	0.1886
		n = 800	0.1338	0.5817	0.7748	0.0703	0.4407	0.6600	0.0154	0.2165	0.4058
		n = 1,200	0.1273	0.7028	0.8922	0.0662	0.5773	0.8149	0.0140	0.3224	0.6055
		n = 2,000	0.1289	0.9174	0.9912	0.0697	0.8577	0.9789	0.0149	0.6664	0.9198
	$\sigma = 1.5$	n = 400	0.1584	0.2899	0.3578	0.0828	0.1759	0.2285	0.0184	0.0510	0.0732
		n = 800	0.1338	0.3283	0.4443	0.0703	0.2138	0.3133	0.0154	0.0717	0.1274
		n = 1,200	0.1273	0.3803	0.5358	0.0662	0.2606	0.4009	0.0140	0.1014	0.1883
		n = 2,000	0.1289	0.5759	0.7811	0.0697	0.4434	0.6683	0.0149	0.2148	0.4215

5. Application to NSCG 2019 data

To demonstrate the utility of the proposed one-sided test procedure in real survey data, we consider the 2019 National Survey of College Graduates, which is conducted by the U.S. Census Bureau. NSCG is a repeated cross-sectional biennial complex survey that provides data on the characteristics of the nation's college graduates, with a focus on those in the science and engineering workforce. In all survey cycles, NSCG used a stratified sampling design to select its sample from the eligible sampling frame, which is the American Community Survey (ACS). Specifically, sample cases were selected from the returning sample members in 2013 NSCG (originally selected from the 2011 ACS), 2015 NSCG (originally selected from the 2013 ACS), 2017 NSCG (originally selected from the 2015 ACS) and the 2017 ACS. Within the sampling strata, probability proportional to size (PPS) or systematic random sampling techniques was used to select the NSCG sample. Due to its various complexities, NSCG implemented replication based approach to variance estimation. The variance-covariance matrix is computed by using the 2019 NSCG replicate weights, which are based on Successive Difference and Jackknife replication methods. The number of replicate weights is 320, which is a decent number to provide a stable variance estimate. Both the replicate weights and replicate adjustment factors were calculated by NSCG and are available upon request. The public use files and relevant documentation are available to the public on the NCSES website (https://www.nsf.gov/statistics/srvygrads/).

The annual salary is the study variable (denoted by SALARY in the dataset), restricted to observations with an annual salary between $30,000 and $900,000. As the annual salary variable distribution is skewed, a log transformation is implemented. Four variables are considered:

- Field (denoted by NDGMEMG in the dataset): This nominal variable defines the field of study for the highest degree. There are six levels: (1) Computer and mathematical sciences; (2) Biological, agricultural and environmental life sciences; (3) Physical and related sciences; (4) Social and related sciences; (5) Engineering; (6) Other.
- Father's education level (denoted by EDDAD in the dataset): This ordinal variable denotes the highest level of education completed by the respondents' father (or male guardian). The six levels are: (1) Less than high school completed; (2) High school diploma or equivalent; (3) Some college, vocational, or trade school (including 2-year degrees); (4) Bachelors degree (e.g., BS, BA, AB); (5) Masters degree (e.g., MS, MA, MBA); (6) Professional degree (e.g., JD, LLB, MD, DDS, etc.) and Doctorate (e.g., PhD, DSc, EdD, etc.).
- Academic year of award for the highest degree (denoted by HDACYR).
- Highest degree type (denoted by DGRDG): The four levels are: (1) Bachelor's; (2) Master's; (3) Doctorate; (4) Professional.

Suppose interest is in the question: for wage-earners whose highest degree is a bachelor's, does the father's education level influence the salary, when controlling for field of study and time since degree? To answer this, we perform separate tests for cohorts in years that the degree was attained, as in Table 5.1.

Within each cohort, there are 36 domains, with six levels each of field and father's education level. The sample sizes for the five cohorts in Table 5.1 are 2,021; 4,032; 5,259; 2,969 and 1,813, respectively. So, the domain sample sizes are generally not small within each cohort. We test the null hypothesis that the salary is constant over father's education level, within each field, against the alternative that the salary is increasing in father's education level. We compare the p-values for this test with constrained alternative to the ANOVA test with unconstrained alternative. The `svyglm` function in `survey` package is used for the unconstrained alternative, and the F test by applying the `anova` function in `survey` package gives the p-value. The results of the tests for five recent cohorts are in Table 5.1.

Table 5.1
***p*-values for the null hypothesis that salary is constant in father's education level, controlling for field of study**

year	2006-2007	2008-2010	2011-2013	2014-2015	2016-2017
one-sided test	0.01951	0.00248	0.00029	0.00622	0.00052
ANOVA F test	0.15198	0.10045	0.01357	0.22231	0.06551

For each cohort, the p-value for the one-sided test is below 0.05, indicating that salaries increase significantly with father's education level, consistently across years. In contrast, the p-value for the two-sided test is consistently larger, and does not indicate a significant trend for some of the cohorts, and for other years the test results could be considered "borderline". Using the *a priori* knowledge that if father education level affects salary, it must be a positive effect, helps increase the power to see the trend.

6. Discussion

In this paper, we developed a testing procedure for testing the linear inequality restrictions of the population domain means within the survey context. Under the design normal assumption of the survey domain means, the proposed test statistic $\hat{T}$ has the asymptotic mixture beta densities, where the mixing probabilities (or the weights) can be easily computed via simulations. The covariance estimator $\hat{\boldsymbol{\Sigma}}$ and the unconstrained estimator $\tilde{\mathbf{y}}_s$ are obtained from the `survey` package in `R` and the constrained least square projection obtained by using the `coneA` function in `coneproj` package. We showed that the power of the test tends to one as the sample size increases, when the alternative hypothesis is true. Simulations show that the test behaves well, with both increased power and improved test size.

The proposed test procedure can be applied to all kinds of complex sampling designs, including stratified sampling, multistage sampling and so on. In practice, though the total sample size n is large, n_d, the number of randomly selected sample in domain d, may be small, or even zero. In such a case, the degrees of freedom (DF) on the estimate of the covariance matrix is small. The degrees of freedom associated with variance estimators was suggested to be (the number of sampled Primary Sampling Units (PSU) with sampled observations in domain d) minus (the number of strata with sampled observations in

domain d), see Graubard and Korn (1996) for more details. Thus, neither the design based variance estimator nor the replication based variance estimator can provide accurate covariance estimate, which may undermine the effectiveness of the proposed test. To address the issue of n_d being small or zero, one might need to apply appropriate imputation methods to create proxy responses for domain d (Haziza and Vallée (2020) considered the use of imputed data in variance estimation). The proposed procedure is expected to work properly as long as the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ accounts for the complex design and the sample size for each domain is not too small. Taking the stratified design as an example, even if the sample size is zero for domain d within certain strata, the test procedure is still applicable provided a decent number of samples for domain d were selected from other strata and the covariance estimate $\hat{\boldsymbol{\Sigma}}$ properly took into account the specific stratified sampling design being considered. In addition, the simulations gave a partial guide for minimum sample sizes needed for the proposed test under stratified simple random sampling design. For more complex sampling design, the effective sample size, defined as the original sample size divided by their design effect, can be considered. Also, it is important to check the weights for units with very low selection probabilities, because extremely small π_i's or π_{ij}'s will result in rather unstable covariance estimate and thus make the proposed test invalid.

Another related issue is that the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}$ may not be positive definite or even positive semidefinite in finite samples. This problem is not uncommon in survey practice, see Théberge (2022), Haslett (2019), Haslett (2016) for more information. This practical issue will have an impact on the inverses of covariance matrix estimates and thus affect the stability of the proposed test procedure. Hence, we suggest survey practitioners check if covariance estimate is positive definite before applying the proposed test in real application.

The implementation of the test in the `csurvey` package borrows from the `survey` package. For example, suppose we have a grid of domains in two variables `x1` and `x2` and study variable `y`. The survey design is specified with the `svydesign` command in the `survey` package, and the design object `ds` is used in the implementation of the test. The p-value for the test of constant versus increasing domain means along the `x1` variable, without constraining the domain means in the `x2` variable, is obtained as follows.

```
ansc=csvy(y~incr(x1)*x2, data=data_set_name, design=ds, nD=M, test=TRUE)
ansc$pval
```

The `csurvey` package also provides the cone information criteron (CIC) for the fitted model, with and without constraints. The CIC was proposed by Oliva-Aviles, Meyer and Opsomer (2019), for checking monotonicity assumptions in the estimation of order-restricted survey domain means, but is valid for any type of constraints. The command `ansc$CIC`, using the above `csurvey` object, returns the CIC for the data fitted with the constraints. The command `ansc$CIC.un` returns the CIC for the data fitted with no constraints. If the CIC is smaller for the constrained fit, this is evidence that the constraints hold. On the other hand, if the unconstrained CIC is larger, this indicates that the assumptions may be incorrect.

For more information and examples, see the `csurvey` manual.

Acknowledgements

This work was partially supported by NSF MMS 1533804.

Appendix

A. Assumptions

(A1) The number of domains D is a known fixed integer and $\liminf_{N\to\infty} \frac{N_d}{N} > 0$, $\limsup_{N\to\infty} \frac{N_d}{N} < 1$ for $d = 1, 2, \ldots, D$.

(A2) The boundedness property of the finite population fourth moment holds. That is, we have:

$$\limsup_{N\to\infty} N^{-1} \sum_{i\in U} y_i^4 < \infty.$$

(A3) The sample size n is non-random and for a sequence of finite populations U_N with corresponding sequence of samples of size n (for simplicity in notation, we omit the subscript N from n_N), we have $n/N \to \lambda$, as $N \to \infty$, where $0 < \lambda < 1$. There exists a constant vector $\boldsymbol{\mu} \in \mathbb{R}^D$ called the "limiting domain means" so that $\bar{y}_{U_d} \to \mu_d$, for $d = 1, \ldots, D$. In addition, there exists a $\pi \in (0,1)$ such that, $\min_d \frac{n_d}{N_d} \geq \pi$, as $N \to \infty$, for $d = 1, \ldots, D$.

(A4) For all N, $\min_{i\in U} \pi_i \geq \lambda_1 > 0$ and $\min_{i,j\in U} \pi_{ij} \geq \lambda_2 > 0$, and

$$\limsup_{N\to\infty} n \max_{i,j\in U, i\neq j} \left| \Delta_{ij} \right| < \infty$$

where $\Delta_{ij} = \mathrm{cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$ and I_i is the sample membership indicator for subject i.

(A5) For any vector $\mathbf{x} \in \mathbb{R}^D$ with finite fourth population moment, we have:

$$\mathrm{var}(\hat{\mathbf{x}}_s)^{-1/2} (\hat{\mathbf{x}}_s - \bar{\mathbf{x}}_U) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}_{D\times D})$$

where $\hat{\mathbf{x}}_s$ is the HT estimator of $\bar{\mathbf{x}}_U = \left(N_1^{-1} \sum_{i\in U_1} x_i, \ldots, N_D^{-1} \sum_{i\in U_D} x_i \right)^{\mathsf{T}}$, $\mathbf{I}_{D\times D}$ is the identity matrix of dimension D, the design covariance matrix $\mathrm{var}(\hat{\mathbf{x}}_s)$ is positive definite.

The assumption (A1) states that the number of domains remains constant as the population size N changes and ensures that there is no asymptotically vanishing domains. Assumption (A2) is a condition needed for showing the variance consistency of the Horvitz-Thompson estimator and this condition generally can be satisfied for most survey data.

In (A3), the assumption of $n/N \to \lambda$ asymptotically ensures that the sample and the population size are of the same order. In addition, by assuming $\min_d \frac{n_d}{N_d} \geq \pi$, as $N \to \infty$, we guarantee that there is no vanishing sampling fraction for each domain d asymptotically, which is a mild condition in the

design-based context. Further, the non-random sample size assumption can be relaxed to accommodate a random sample size by imposing particular conditions on the expected sample size $\mathrm{E}_p(n)$.

Assumption (A4) illustrates that the design is both a probability sampling design and a measurable design. The assumption on the Δ_{ij} states that the covariance between sample membership indicators is sufficiently small, which goes to zero at rate of n^{-1}. These conditions hold in many classical sampling designs, including simple random sampling with and without replacement, and many other unequal probability sampling designs.

The asymptotic normal assumption in (A5) is usually assumed explicitly and it is satisfied for many specific sampling designs, including simple random sampling with or without replacement. Also, it holds for Poisson sampling and unequal probability sampling with replacement. The design asymptotic normal assumption, taken together with the variance consistency of the Horvitz-Thompson estimator, can be used to derive the asymptotic distribution of the constrained domain mean estimator. More importantly, it is this normal assumption that makes it possible for us to take advantage of the available techniques in the one-sided test literatures and obtain the null distribution of the test statistics approximately. Otherwise, we have to resort to the bootstrap method to get the empirical distribution of the test statistics when the properties of the design estimator are completely unknown.

It is useful to note that all the results developed in this paper remains design-based. Only the design variability is accounted for by the asymptotic variance in the main results. While the design normal assumption can be viewed as "model-like" assumption, it does not imply a random structure for the population and the inference does not involve any type of model variability. The distributional properties derived in the main text follow from the design and sample size assumptions (A3)-(A5).

B. Supplemental materials for Section 3

In this section, we will show the following result

$$\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\top} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) - \min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) = o_p(1)$$

to complete the proof for Lemma 1. Based on the result from (2.1) in Xu et al. (2021), for the term $\min_{\boldsymbol{\theta}_1 \in C}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\top} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) = \min_{\hat{\boldsymbol{\theta}}_1 \in \hat{C}}\left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1\right)^{\top}\left(\hat{\mathbf{Z}}_s - \hat{\boldsymbol{\theta}}_1\right)$, the projection of $\hat{\mathbf{Z}}_s$ onto the cone $\hat{C}$ can be expressed as:

$$\hat{\boldsymbol{\theta}}_1^{*} = \sum_{J}\left(\mathbf{I} - \hat{\mathbf{A}}_J^{\top}\left(\hat{\mathbf{A}}_J \hat{\mathbf{A}}_J^{\top}\right)^{-} \hat{\mathbf{A}}_J\right)\hat{\mathbf{Z}}_s \mathcal{I}_J(s) \tag{B.1}$$

where the sum is over $J \subseteq \{1, \ldots, M\}$ such that the rows of $\hat{\mathbf{A}}_J$ form a linearly independent set and for each sample s, there is only one subset J for which $\mathcal{I}_J(s) = 1$. Using the above explicit form of $\hat{\boldsymbol{\theta}}_1^{*}$, we prove the following results.

Lemma 2. *Let* $\boldsymbol{\mu}$ *be the limiting domain means. Let* J *be the set that is associated with* $\hat{\boldsymbol{\theta}}_1^*$ *in (B.1) and* $J_{\boldsymbol{\mu}}^0$ *be the corresponding set for the solution* $\boldsymbol{\theta}_{\boldsymbol{\mu}}^*$ *that minimizes* $\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_1\right)^{\top}\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_1\right)$ *subject to* $\boldsymbol{\theta}_1 \in C_{\boldsymbol{\mu}}=\left\{\boldsymbol{\theta}:\mathbf{A}_{\boldsymbol{\mu}}\boldsymbol{\theta}\geq 0\right\}$, *where* $\mathbf{Z}_{\boldsymbol{\mu}}=\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1/2}\boldsymbol{\mu}$, $C_{\boldsymbol{\mu}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ *are limiting versions of* $\hat{\mathbf{Z}}_s$, $\hat{C}$, $\hat{\boldsymbol{\Sigma}}$ *and* $\mathbf{A}_{\boldsymbol{\mu}}=\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2}$. *Define* $J_{\boldsymbol{\mu}}^1=\left\{j:\mathbf{A}_j\boldsymbol{\mu}=0\right\}$ *and let* $J_{\boldsymbol{\mu}}=J_{\boldsymbol{\mu}}^0\cup J_{\boldsymbol{\mu}}^1$, *Then, we have:*

$$\Pr\left(J\not\subseteq J_{\boldsymbol{\mu}}\right)=o(1)\quad \textit{and}\quad \Pr\left(J_{\boldsymbol{\mu}}^0\not\subseteq J\right)=o(1).$$

Proof. Firstly, consider the event $J\not\subseteq J_{\boldsymbol{\mu}}$. Define

$$\begin{aligned}\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right) &= \left(\hat{\mathbf{Z}}_s-\hat{\boldsymbol{\theta}}_1^*\right)^{\top}\left(\hat{\mathbf{Z}}_s-\hat{\boldsymbol{\theta}}_1^*\right)\\ &= \left[\hat{\mathbf{Z}}_s-\left(\mathbf{I}-\hat{\mathbf{A}}_J^{\top}\left(\hat{\mathbf{A}}_J\hat{\mathbf{A}}_J^{\top}\right)^{-}\hat{\mathbf{A}}_J\right)\hat{\mathbf{Z}}_s\right]^{\top}\left[\hat{\mathbf{Z}}_s-\left(\mathbf{I}-\hat{\mathbf{A}}_J^{\top}\left(\hat{\mathbf{A}}_J\hat{\mathbf{A}}_J^{\top}\right)^{-}\hat{\mathbf{A}}_J\right)\hat{\mathbf{Z}}_s\right]\\ &= \hat{\mathbf{Z}}_s^{\top}\hat{\mathbf{A}}_J^{\top}\left(\hat{\mathbf{A}}_J\hat{\mathbf{A}}_J^{\top}\right)^{-}\hat{\mathbf{A}}_J\hat{\mathbf{Z}}_s\\ &= \tilde{\mathbf{y}}_s^{\top}\mathbf{A}_J^{\top}\left(\mathbf{A}_J\hat{\boldsymbol{\Sigma}}\mathbf{A}_J^{\top}\right)^{-}\mathbf{A}_J\tilde{\mathbf{y}}_s\end{aligned}$$

similarly, we define:

$$\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)=\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)^{\top}\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)=\boldsymbol{\mu}^{\top}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}^0}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\right)^{-}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}\boldsymbol{\mu}.$$

Note that the projection of $\mathbf{Z}_{\boldsymbol{\mu}}$ onto the linear space spanned by rows of $\mathbf{A}_{\boldsymbol{\mu}}$ in position $J_{\boldsymbol{\mu}}^0$ is the same as the projection onto the linear space spanned by rows of $\mathbf{A}_{\boldsymbol{\mu}}$ in position $J_{\boldsymbol{\mu}}$, so we have:

$$\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)=\boldsymbol{\mu}^{\top}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}^0}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\right)^{-}\mathbf{A}_{J_{\boldsymbol{\mu}}^0}\boldsymbol{\mu}=\boldsymbol{\mu}^{\top}\mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\right)^{-}\mathbf{A}_{J_{\boldsymbol{\mu}}}\boldsymbol{\mu}.$$

Further, denote:

$$\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right)=\left(\hat{\mathbf{Z}}_s-\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right)^{\top}\left(\hat{\mathbf{Z}}_s-\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right)=\tilde{\mathbf{y}}_s^{\top}\mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}}\hat{\boldsymbol{\Sigma}}\mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\right)^{-}\mathbf{A}_{J_{\boldsymbol{\mu}}}\tilde{\mathbf{y}}_s$$

$$\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)=\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)^{\top}\left(\mathbf{Z}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)=\boldsymbol{\mu}^{\top}\mathbf{A}_J^{\top}\left(\mathbf{A}_J\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{A}_J^{\top}\right)^{-}\mathbf{A}_J\boldsymbol{\mu}$$

where $\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}=\left(\mathbf{I}-\hat{\mathbf{A}}_{J_{\boldsymbol{\mu}}}^{\top}\left(\hat{\mathbf{A}}_{J_{\boldsymbol{\mu}}}\hat{\mathbf{A}}_{J_{\boldsymbol{\mu}}}^{\top}\right)^{-}\hat{\mathbf{A}}_{J_{\boldsymbol{\mu}}}\right)\hat{\mathbf{Z}}_s$ and $\boldsymbol{\theta}_{\boldsymbol{\mu},J}=\left(\mathbf{I}-\mathbf{A}_{\boldsymbol{\mu},J}^{\top}\left(\mathbf{A}_{\boldsymbol{\mu},J}\mathbf{A}_{\boldsymbol{\mu},J}^{\top}\right)^{-}\mathbf{A}_{\boldsymbol{\mu},J}\right)\mathbf{Z}_{\boldsymbol{\mu}}$. Then, we must have

$$\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)<\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)\quad\text{and}\quad\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right)<\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right)$$

and due to the consistency of $\tilde{\mathbf{y}}_s$ and $\hat{\boldsymbol{\Sigma}}$, respectively, we also have:

$$\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right)-\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)\right)=o_p(1)\quad\text{and}\quad\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right)-\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right)=o_p(1).$$

Finally, by Markov's inequality, we get:

$$\begin{aligned}\Pr\left(J \not\subseteq J_{\boldsymbol{\mu}}\right) \leq\ & \Pr\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right) - \widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right) + \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right) > \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right) \\ \leq\ & \frac{E\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right) - \widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right) + \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right)}{\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)} \\ =\ & \frac{E\left(\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right)\right) - E\left(\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)\right)\right)}{\frac{1}{n}\left(\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right)} \\ \to\ & 0\end{aligned}$$

since $E\left(\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_{1,J_{\boldsymbol{\mu}}}\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right)\right)\right) = o(1)$ and $E\left(\frac{1}{n}\left(\widetilde{\mathrm{SSE}}\left(\hat{\boldsymbol{\theta}}_1^*\right) - \mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu},J}\right)\right)\right) = o(1)$. Using the similar argument, we can also show that:

$$\Pr\left(J_{\boldsymbol{\mu}}^0 \not\subseteq J\right) = o(1)$$

this completes the proof.

By the same argument as in Lemma 2, we also have the following result.

Lemma 3. *Let $J_{\boldsymbol{\Sigma}}$ (unknown) be the corresponding set of the solution $\tilde{\boldsymbol{\theta}}_1^*$ that minimizes $\left(\tilde{\mathbf{Z}}_s - \boldsymbol{\theta}_1\right)^\top \left(\tilde{\mathbf{Z}}_s - \boldsymbol{\theta}_1\right)$ subject to $\boldsymbol{\theta}_1 \in \tilde{C}$. Then, we have:*

$$\Pr\left(J_{\boldsymbol{\Sigma}} \not\subseteq J_{\boldsymbol{\mu}}\right) = o(1) \quad \textit{and} \quad \Pr\left(J_{\boldsymbol{\mu}}^0 \not\subseteq J_{\boldsymbol{\Sigma}}\right) = o(1),$$

where $J_{\boldsymbol{\mu}}$ and $J_{\boldsymbol{\mu}}^0$ are defined in Lemma 2.

Lemma 4. *We have:*

$$\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) - \min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) = o_p(1)$$

with respect to the sampling mechanism.

Proof. Let J be the observed set for a given sample s. We can write the difference as follows:

$$\begin{aligned}&\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) - \min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) \\ =\ & \frac{1}{n}\hat{\mathbf{Z}}_s^\top \hat{\mathbf{A}}_J^\top \left(\hat{\mathbf{A}}_J \hat{\mathbf{A}}_J^\top\right)^- \hat{\mathbf{A}}_J \hat{\mathbf{Z}}_s - \frac{1}{n}\tilde{\mathbf{Z}}_s^\top \tilde{\mathbf{A}}_{J_{\boldsymbol{\Sigma}}}^\top \left(\tilde{\mathbf{A}}_{J_{\boldsymbol{\Sigma}}} \tilde{\mathbf{A}}_{J_{\boldsymbol{\Sigma}}}^\top\right)^- \tilde{\mathbf{A}}_{J_{\boldsymbol{\Sigma}}} \tilde{\mathbf{Z}}_s \\ =\ & \tilde{\mathbf{y}}_s^\top \mathbf{A}_J^\top \left(\mathbf{A}_J n\hat{\boldsymbol{\Sigma}} \mathbf{A}_J^\top\right)^- \mathbf{A}_J \tilde{\mathbf{y}}_s - \tilde{\mathbf{y}}_s^\top \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^\top \left(\mathbf{A}_{J_{\boldsymbol{\Sigma}}} n\boldsymbol{\Sigma} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^\top\right)^- \mathbf{A}_{J_{\boldsymbol{\Sigma}}} \tilde{\mathbf{y}}_s \\ =\ & \tilde{\mathbf{y}}_s^\top \mathbf{A}_J^\top \left(\mathbf{A}_J n\hat{\boldsymbol{\Sigma}} \mathbf{A}_J^\top\right)^- \mathbf{A}_J \tilde{\mathbf{y}}_s \left(I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J \subseteq J_{\boldsymbol{\mu}}\right)} + I_{\left(J \not\subseteq J_{\boldsymbol{\mu}} \text{ or } J_{\boldsymbol{\mu}}^0 \not\subseteq J\right)}\right) \\ -\ & \tilde{\mathbf{y}}_s^\top \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^\top \left(\mathbf{A}_{J_{\boldsymbol{\Sigma}}} n\boldsymbol{\Sigma} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^\top\right)^- \mathbf{A}_{J_{\boldsymbol{\Sigma}}} \tilde{\mathbf{y}}_s \left(I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J_{\boldsymbol{\Sigma}} \subseteq J_{\boldsymbol{\mu}}\right)} + I_{\left(J_{\boldsymbol{\Sigma}} \not\subseteq J_{\boldsymbol{\mu}} \text{ or } J_{\boldsymbol{\mu}}^0 \not\subseteq J_{\boldsymbol{\Sigma}}\right)}\right)\end{aligned}$$

by Lemma 2 and Lemma 3, we have that $I_{\left(J \not\subseteq J_{\boldsymbol{\mu}} \text{ or } J_{\boldsymbol{\mu}}^0 \not\subseteq J\right)} = o_p(1)$ and $I_{\left(J_{\boldsymbol{\Sigma}} \not\subseteq J_{\boldsymbol{\mu}} \text{ or } J_{\boldsymbol{\mu}}^0 \not\subseteq J_{\boldsymbol{\Sigma}}\right)} = o_p(1)$. Then, we have:

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\top} \hat{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) - \min_{\boldsymbol{\theta}_1 \in C} \frac{1}{n}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right)^{\top} \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{y}}_s - \boldsymbol{\theta}_1\right) \\
&= \tilde{\mathbf{y}}_s^{\top} \mathbf{A}_J^{\top}\left(\mathbf{A}_J n\hat{\boldsymbol{\Sigma}} \mathbf{A}_J^{\top}\right)^{-} \mathbf{A}_J \tilde{\mathbf{y}}_s I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J \subseteq J_{\boldsymbol{\mu}}\right)} - \tilde{\mathbf{y}}_s^{\top} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\Sigma}}} n\boldsymbol{\Sigma} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\Sigma}}} \tilde{\mathbf{y}}_s I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J_{\boldsymbol{\Sigma}} \subseteq J_{\boldsymbol{\mu}}\right)} + o_p(1) \\
&= \boldsymbol{\mu}^{\top} \mathbf{A}_J^{\top}\left(\mathbf{A}_J n\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_J^{\top}\right)^{-} \mathbf{A}_J \boldsymbol{\mu} I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J \subseteq J_{\boldsymbol{\mu}}\right)} - \boldsymbol{\mu}^{\top} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\Sigma}}} n\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_{J_{\boldsymbol{\Sigma}}}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\Sigma}}} \boldsymbol{\mu} I_{\left(J_{\boldsymbol{\mu}}^0 \subseteq J_{\boldsymbol{\Sigma}} \subseteq J_{\boldsymbol{\mu}}\right)} + o_p(1) \\
&= \boldsymbol{\mu}^{\top} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}^0} n\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\mu}}^0} \boldsymbol{\mu} - \boldsymbol{\mu}^{\top} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}^0} n\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\mu}}^0} \boldsymbol{\mu} + o_p(1) \\
&= o_p(1)
\end{aligned}
$$

where we use the fact that for any set J with $J_{\boldsymbol{\mu}}^0 \subseteq J \subseteq J_{\boldsymbol{\mu}}$, we have that

$$
\begin{aligned}
\mathrm{SSE}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}}^*\right) &= \boldsymbol{\mu}^{\top} \mathbf{A}_J^{\top}\left(\mathbf{A}_J \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_J^{\top}\right)^{-} \mathbf{A}_J \boldsymbol{\mu} \\
&= \boldsymbol{\mu}^{\top} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}^0} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_{J_{\boldsymbol{\mu}}^0}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\mu}}^0} \boldsymbol{\mu} \\
&= \boldsymbol{\mu}^{\top} \mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\left(\mathbf{A}_{J_{\boldsymbol{\mu}}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \mathbf{A}_{J_{\boldsymbol{\mu}}}^{\top}\right)^{-} \mathbf{A}_{J_{\boldsymbol{\mu}}} \boldsymbol{\mu}.
\end{aligned}
$$

References

Bartholomew, D. (1959). A test of homogeneity for ordered alternatives. *Biometrika*, 46, 36-48.

Bartholomew, D. (1961). A test of homogeneity of means under order restrictions. *Journal of the Royal Statistical Society, Series B*, 23(1), 239-272.

Chacko, V.J. (1963). Testing homogeneity against ordered alternatives. *The Annals of Mathematical Statistics*, 34(3), 945-956.

Graubard, B.I., and Korn, E.L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology*, 144, 102-106.

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

Haslett, S. (2016). Positive semidefiniteness of estimated covariance matrices in linear models for sample survey data. *Special Matrices*, 4, 218-224.

Haslett, S. (2019). Best linear unbiased estimation for varying probability with and without replacement sampling. *Special Matrices*, 7, 78-91.

Haziza, D., and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science*, 3, 583-623.

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.

Liao, X., and Meyer, M.C. (2014). coneproj: an R package for the primal or dual cone projections with routines for constrained regression. *Journal of Statistical Software*, 61, 1-22.

McDermott, M.P., and Mudholkar, G.S. (1993). A simple approach to testing homogeneity of order-constrained means. *Journal of the American Statistical Association*, 88, 1371-1379.

Meyer, M.C. (2003). A test for linear versus convex regression function using shape-restricted regression. *Biometrika*, 90(1), 223-232.

Meyer, M.C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*, 42, 1126-1139.

Meyer, M.C., and Wang, J. (2012). Improved power of one-sided tests. *Statistics and Probability Letters*, 82, 1619-1622.

Meyer, M.C., and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, 28(4), 1083-1104.

Oliva-Aviles, C., Meyer, M.C. and Opsomer, J.D. (2019). Checking validity of monotone domain mean estimators. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 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 Methodology*, 46, 2, 145-180. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020002/article/00002-eng.pdf.

Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, 40(2), 549-567.

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.

Sen, B., and Meyer, M.C. (2017). Testing against a linear regression model using ideas from shape-restricted estimation. *Journal of the Royal Statistical Society, Series B*, 79, 423-448.

Silvapulle, M.J., and Sen, P. (2005). *Constrained Statistical Inference*. Hoboken, New Jersey: Wiley.

Théberge, A. (2022). A generalization of inverse probability weighting. *Survey Methodology*, 48, 1, 177-190. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022001/article/00009-eng.pdf.

Wu, J., Meyer, M.C. and Opsomer, J.D. (2016). Survey estimation of domain means that respect natural orderings. *The Canadian Journal of Statistics/La revue canadienne de statistique*, 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, June 2023
Vol. 49, No. 1, pp. 139-162
Statistics Canada, Catalogue No. 12-001-X

An extension of the weight share method when using a continuous sampling frame

Guillaume Chauvet, Olivier Bouriaud and Philippe Brion[1]

Abstract

The definition of statistical units is a recurring issue in the domain of sample surveys. Indeed, not all the populations surveyed have a readily available sampling frame. For some populations, the sampled units are distinct from the observation units and producing estimates on the population of interest raises complex questions, which can be addressed by using the weight share method (Deville and Lavallée, 2006). However, the two populations considered in this approach are discrete. In some fields of study, the sampled population is continuous: this is for example the case of forest inventories for which, frequently, the trees surveyed are those located on plots of which the centers are points randomly drawn in a given area. The production of statistical estimates from the sample of trees surveyed poses methodological difficulties, as do the associated variance calculations. The purpose of this paper is to generalize the weight share method to the continuous (sampled population) – discrete (surveyed population) case, from the extension proposed by Cordy (1993) of the Horvitz-Thompson estimator for drawing points carried out in a continuous universe.

Key Words: Continuous sampling design; Environmental statistics; Forest inventory; Synthetic variable; Variance estimation.

1. Introduction

The definition of statistical units is a recurring issue in the domain of sample surveys. Indeed, not all the populations surveyed have a readily available sampling frame. For these populations, the sampled units (for which a sampling frame is available from which to select units according to a given sampling design), are distinct from the observation units, which constitute the population of interest on which we are willing to infer.

This issue has been raised for a long time for studying populations that are difficult to reach, e.g., homeless people (see for example Ardilly and Le Blanc, 2001; De Vitiis, Falorsi and Inglese, 2014; Laporte, Vandentorren, Détrez, Douay, Le Strat, Le Méner, Chauvin and The Samenta Research Group, 2018), or nomad/non-localized populations (see for example Lohlé-Tart, Clairin, François and Gendreau, 1988; Clairin and Brion, 1996; Himelein, Eckman and Murray, 2014). It has also become recently more and more accurate for business statistics, with the use of a unit "enterprise" not necessarily equivalent to the unit available in the business registers (Lorenc, Smith and Bavdaž, 2018). In this case, producing estimates on the population of interest raises complex questions, linked to the fact that the weights of the observation units need to be based on the design weights of the units selected in the sampling frame.

To deal with this issue, Deville and Lavallée (2006) proposed the so-called weight share method. It is based on a principle of duality between the sampled population and the observed population, where a variable of interest defined on the observed population may be written as a synthetic variable defined on

1. Guillaume Chauvet, Ensai (Irmar), Campus de Ker Lann, Bruz, France. E-mail: guillaume.chauvet@ensai.fr; Olivier Bouriaud, Université Stefan cel Mare de Suceava, 13 rue de l'Université, 720229 Suceava, Roumania and IGN, Laboratoire d'Inventaire Forestier, 14 rue Girardet, 54000 Nancy, France. E-mail: obouriaud.lif@gmail.com; Philippe Brion, Irmar, France. E-Mail: philippe.brion55@gmail.com.

the sampling frame (see also Lavallée, 2009). Because it creates a link between the observation units and the sampling units, this method enables the properties of the sampling design to be used to define unbiased estimators of totals for the observed populations, and to derive variance formulas. In particular, the sampling weights of the sampling units are used to assign estimation weights to the observation units. This paper deals with the extension of this method to the case when the sampled population is a continuous frame. For the sampled population and the observed population, we will use the notations U^A and U^B in case of discrete populations, and $\mathcal{U}^A$ and $\mathcal{U}^B$ in case of continuous populations.

We are particularly interested in applications encountered in forest inventories, in which it is common practice to use a sample of points selected in a continuum and then fixed-shape supports defined from these points to perform the survey on a discrete population of trees. The approach which consists of transporting a variable from the discrete population to the continuous population is not new, and has been considered by Stevens and Urquhart (2000), Gregoire and Valentine (2007) and Mandallaz (2007), for example. While these previous works were quite similar in their overall approach of the indirect sampling, the link between the units from the population sampled and the units of the target population are only implicit.

The work by Stevens and Urquhart (2000) is very similar to ours. They studied the situation when a finite population of interest is linked to a continuous territory, and they considered a way to transfer a variable of interest onto the continuous sampling frame. This is similar to the synthetic variable that we present in equation (3.15). They derived a so-called "aggregation-unbiasedness" requirement, in order to obtain unbiased total estimators. They also proposed a Horvitz-Thompson variance estimator, making use of the theory by Cordy (1993). Despite the importance of this paper, it did not have a significant impact in the literature. It is in particular telling that the article is not quoted in textbooks like Gregoire and Valentine (2007) and Mandallaz (2007). Therefore, we feel the need for a simple presentation of the approach, where we clearly state what are the estimation weights, the resulting estimators of totals for the finite population, and the associated Horvitz-Thompson variance estimators. The weight share method is a very useful and simple tool for this, as illustrated in the applications considered in Section 3.

In natural populations such as forest trees, the units are distributed spatially over a territory. Estimating totals of any given attribute of this population requires undergoing spatial sampling. To this end, Gregoire and Valentine (2007, Chapter 10) introduce the so-called Monte Carlo integration approach, and call the synthetic variable the "attribute density". Several examples are given for devices used in the practice of forest inventory (e.g., point relascope sampling, line intersect sampling). However, the link with Cordy's set-up is not pointed out, and variance estimation is restricted to the case when the points are selected by independent uniform sampling. Mandallaz (2007, Section 4.2) also develops a related approach, where the link between the observed, finite population and the sampled, continuous population is performed by means of the so-called "local density". The approach is first presented for plot sampling, and then extended to more complex situations like cluster (trakt) sampling, which is popular for forest inventories (Lawrence, McRoberts, Tomppo, Gschwantner and Gabler, 2010). However, the method is quite ad-hoc,

since the local density needs to be computed differently in each situation. On the other hand, the weight share method enables one to produce general formulas for both point estimators and variance estimators. In particular, we consider in Section 3.4 the situation of spatial cluster sampling for forest inventories, for which the weight share method provides a general solution for estimation and variance estimation under an arbitrary continuous sampling design.

In what follows, we first recall in Section 2 the basic principles of the weight share method in the case of two discrete populations U^A and U^B. In Section 3, we extend the method to cover the case when a continuous population $\mathcal{U}^A$ is sampled, and we want to infer on a discrete population U^B. The results of two simulation studies are presented in Section 4. We conclude in Section 5.

2. Sampling in a discrete population

In this section, we first define in Section 2.1 our notations when sampling in a discrete population U^A. We then recall in Section 2.2 how the weight share method may be used to produce estimates in another discrete population U^B linked to U^A. A simple example is presented in Section 2.3 for illustration.

2.1 Notations

We are interested in a discrete population U^A, for which the units in the population are enumerable and a sampling frame may therefore be available. For example, this may be a population of households or individuals in social surveys, or a register in business surveys. The size of the population U^A is denoted as N^A. Suppose that we are interested in a variable of interest y^A taking the value y_i^A for unit $i \in U^A$, and that we wish to estimate the population total

$$\tau_y^A = \sum_{i \in U^A} y_i^A. \tag{2.1}$$

A random sample S^A is selected in U^A by means of a sampling design $p^A(\cdot)$, and we let s^A denote a possible realization of S^A. The Horvitz-Thompson (HT) estimator is

$$\hat{\tau}_y^A = \sum_{i \in S^A} d_i^A y_i^A, \tag{2.2}$$

where $d_i^A = 1/\pi_i^A$ is the design weight of unit i, and π_i^A the probability for unit i of inclusion in the sample. This estimator is design-unbiased for τ_y^A, provided that all the π_i^A's are > 0.

The variance of $\hat{\tau}_y^A$ is

$$V_p\left(\hat{\tau}_y^A\right) = \sum_{i,j \in U^A} \frac{y_i^A}{\pi_i^A} \frac{y_j^A}{\pi_j^A} \left(\pi_{ij}^A - \pi_i^A \pi_j^A\right), \tag{2.3}$$

where π_{ij}^A is the probability that the units i and j are jointly selected in S^A. If all the π_{ij}^A's are positive, this variance is unbiasedly estimated by

$$\hat{V}^A\left(\hat{\tau}_y^A\right) = \sum_{i,j\in S^A} \frac{y_i^A}{\pi_i^A}\frac{y_j^A}{\pi_j^A}\left(\frac{\pi_{ij}^A - \pi_i^A\pi_j^A}{\pi_{ij}^A}\right). \tag{2.4}$$

2.2 Weight share method

Suppose that we are interested in another population U^B, with a variable of interest y^B taking the value y_k^B for unit $k \in U^B$. We wish to estimate the population total

$$\tau_y^B = \sum_{k\in U^B} y_k^B. \tag{2.5}$$

We suppose that no sampling frame is available for U^B, but that this population is linked to the population U^A. The link between the units in U^A and U^B is represented by the indicator variables

$$L^{AB}(i,k) = \begin{cases} 1 \text{ if units } i\in U^A \text{ and } k\in U^B \text{ are linked,} \\ 0 \text{ otherwise.} \end{cases} \tag{2.6}$$

The set of *ancestors* for some unit $k \in U^B$ is $\text{Anc}_k = \left\{i \in U^A;\ L^{AB}(i,k) = 1\right\}$. The set of *descendants* for some unit $i \in U^A$ is $\text{Des}_i = \left\{k \in U^B;\ L^{AB}(i,k) = 1\right\}$. For any unit $k \in U^B$,

$$N_{+k}^{AB} = \sum_{i\in U^A} L^{AB}(i,k) \tag{2.7}$$

is the total number of ancestors. It is required that any unit $k \in U^B$ be linked to at least one unit in U^A; that is, we suppose that $N_{+k}^{AB} > 0$ for any unit $k \in U^B$.

A sample S^B is obtained in U^B by surveying all the descendants of the units i selected in S^A. More formally, we have

$$S^B = \bigcup_{i\in S^A} \text{Des}_i. \tag{2.8}$$

To obtain an estimator of τ_y^B, the weight share method (Deville and Lavallée, 2006) makes use of a principle of duality between populations U^A and U^B, based on the link function given in (2.6). The total τ_y^B may be written as

$$\tau_y^B = \sum_{i\in U^A} y_i^A \quad \text{with} \quad y_i^A = \sum_{k\in U^B} \frac{L^{AB}(i,k)\, y_k^B}{N_{+k}^{AB}}, \tag{2.9}$$

see Deville and Lavallée (2006, Result 2). Equation (2.9) represents the fact that the variable y_k^B may be distributed over the units in U^A to obtain a synthetic variable y_i^A. This is done by sharing each value y_k^B equally among the ancestors in Anc_k.

From equation (2.9), the total τ_y^B can be unbiasedly estimated by performing HT-estimation on the sample S^A, with the synthetic variable y_i^A. This HT-estimator may be rewritten as

$$\hat{\tau}_y^B = \sum_{i \in S^A} d_i^A y_i^A = \sum_{k \in S^B} w_k^B y_k^B, \tag{2.10}$$

with

$$w_k^B = \frac{1}{N_{+k}^{AB}} \sum_{i \in S^A} L^{AB}(i,k)\, d_i^A,$$

see Deville and Lavallée (2006, Result 3). Each unit $k \in S^B$ is given the sum of the weights of the sampled units $i \in S^A$ which are linked to k, divided by the number of links N_{+k}^{AB}. The weights d_i^A of the units $i \in S^A$ are therefore shared among the units $k \in S^B$, hence the name of the method. It is important to note that the weights w_k^B can only be computed if the number of ancestors N_{+k}^{AB} is known for any unit $k \in S^B$. Therefore, this information needs to be collected during the survey. In some situations, it may be difficult or even impossible to state whether or not a unit in U^A is related to another unit in U^B. This is referred to as link nonresponse by Xu and Lavallée (2009), who propose treatment methods to handle this problem.

From equation (2.10), the weight share method enables one to attribute to each unit $k \in S^B$ a weight w_k^B, which is usable for any variable of interest y_k^B and such that the estimator $\hat{\tau}_y^B$ is unbiased. This is a very strong property. In contrast, the HT-estimator for the sample S^B cannot be computed. The inclusion probability of unit k in the sample S^B is

$$\Pr\left(k \in S^B\right) = \sum_{\substack{s^A \subset U^A \\ s^A \cap \mathrm{Anc}_k \neq \varnothing}} p^A(s^A). \tag{2.11}$$

Computing these inclusion probabilities would require a full specification of both the sampling design $p^A(\cdot)$ and of the links between both populations, which is usually impossible.

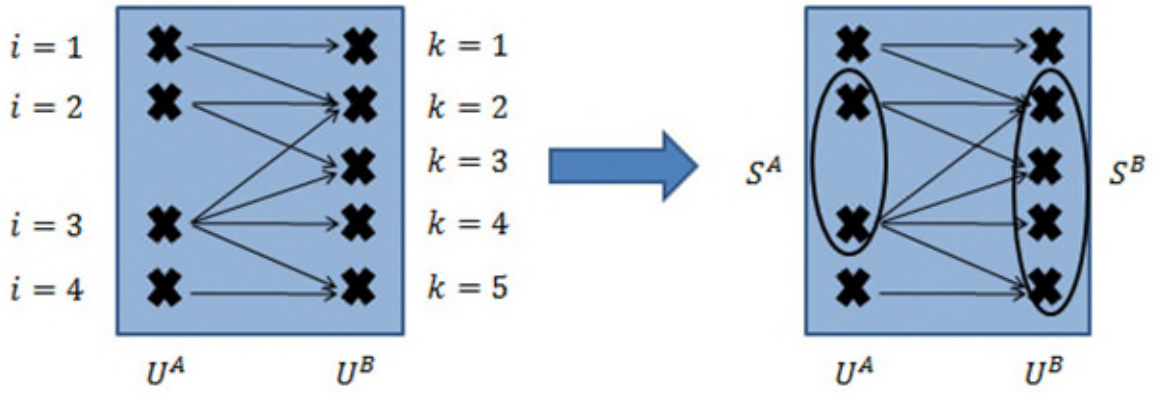
Since $\hat{\tau}_y^B$ may be written as a HT-estimator on the sample S^A, the variance of $\hat{\tau}_y^B$ is given by equation (2.3), with y_i^A the synthetic variable given in (2.9), and a variance estimator is given by (2.4). Note that the variable y_i^A can be exactly computed for any unit $i \in S^A$, since it is assumed that all the units in U^B linked to the units in S^A are surveyed, see equation (2.8).

2.3 A simple example

For illustration, we present a toy example in Figure 2.1. The population U^A contains $N^A = 4$ units, and the population U^B contains $N^B = 5$ units. The links between units are represented by the arrows. For example, unit $i = 3$ in U^A has four descendants, namely units $k = 2, 3, 4$ and 5. Therefore, we have $L^{AB}(3,1) = 0$ and $L^{AB}(3,2) = L^{AB}(3,3) = L^{AB}(3,4) = L^{AB}(3,5) = 1$. Unit $k = 4$ has a single ancestor $i = 3$, whereas unit $k = 5$ has two ancestors $i = 3$ and $i = 4$. Suppose that the sampling design leads to the selection of the subset $s^A = \{2, 3\}$. Then all the descendants of the units in s^A are selected, resulting in the observation of the subset

$$s^B = \{2, 3, 4, 5\}. \tag{2.12}$$

Figure 2.1 A simple example of links between two discrete populations U^A and U^B.



Now, suppose that the sampled values are

$$y_2^B = 1, \quad y_3^B = 3, \quad y_4^B = 3, \quad y_5^B = 5. \tag{2.13}$$

We first compute the synthetic variable y_i^A for the units $i \in s^A$, making use of equation (2.9). We obtain

$$\begin{aligned} y_2^A &= \frac{y_2^B}{3} + \frac{y_3^B}{2} \simeq 1.83, \\ y_3^A &= \frac{y_2^B}{3} + \frac{y_3^B}{2} + \frac{y_4^B}{1} + \frac{y_5^B}{2} \simeq 7.33. \end{aligned}$$

Now, we compute the weights w_k^B for the units $k \in s^B$ by means of the weight share method, making use of equation (2.10). Suppose that the sample S^A is selected in U^A by simple random sampling without replacement, leading to $\pi_i^A = 0.5$ and $d_i^A = 2$ for any $i \in U^A$. We obtain

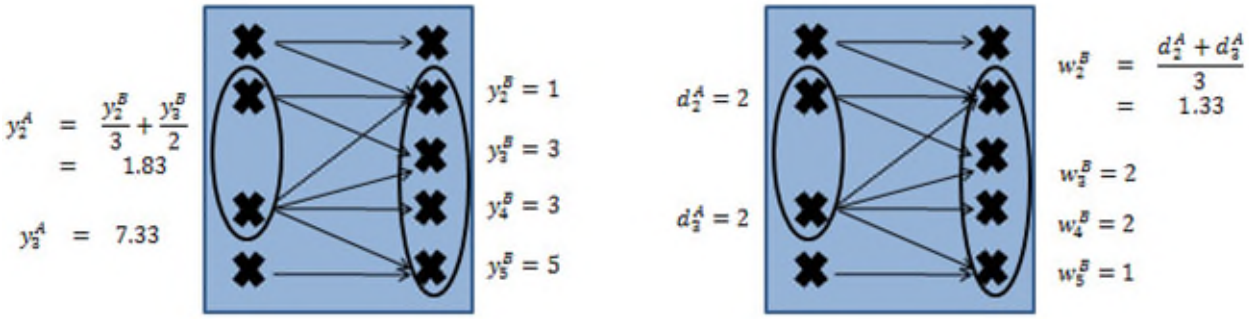
$$\begin{aligned} w_2^B &= \frac{d_2^A + d_3^A}{3} \simeq 1.33, \\ w_3^B &= \frac{d_2^A + d_3^A}{2} = 2, \\ w_4^B &= \frac{d_3^A}{1} = 2, \\ w_5^B &= \frac{d_3^A}{2} = 1. \end{aligned}$$

The estimate for the total τ_y^B is therefore

$$\hat{\tau}_y^B(s^B) = \sum_{k \in s^B} w_k^B y_k^B \simeq 18.33.$$

The results are summarized in Figure 2.2.

Figure 2.2 Computation of the synthetic variable y_i^A and of the weights d_k^B obtained by the weight share method on a simple example.



3. Sampling in a continuous population

In this section, we first define in Section 3.1 our notations when sampling in a continuous universe $\mathcal{U}^A$, following Cordy (1993). We explain in Section 3.2 how the weight share method may be extended to produce estimates in some discrete population U^B linked to $\mathcal{U}^A$. We consider applications to sampling designs used in forest inventories, for which some analog approaches have been proposed in the literature. The case of direct plot sampling is first considered in Section 3.3. The application to cluster sampling is considered in Section 3.4.

3.1 Notations

We first review the set-up for sampling and estimation in a continuous universe $\mathcal{U}^A$ introduced by Cordy (1993). This framework was strongly motivated by environmental applications, and led to an extension of the HT-estimator existing for a discrete population, see Section 2.1.

Suppose that we are interested in a continuous population or universe, i.e., as defined by Gregoire and Valentine (2007, page 93) "that does not naturally divide into smaller discrete units". This may be a landscape or a lake, for example. We suppose that the universe $\mathcal{U}^A$ is included in $\mathbf{R}^q$ with $q \geq 1$. We are interested in some Lebesgue integrable function $y^A : \mathcal{U}^A \to \mathbf{R}$, and we wish to estimate the total (integral)

$$\tau_y^A = \int_{\mathcal{U}^A} y^A(x)\, dx \tag{3.1}$$

of this function over $\mathcal{U}^A$.

A random sample $S^A = \left\{S_1^A, \ldots, S_{n^A}^A\right\}$ of n^A locations is selected from $\mathcal{U}^A$, and we let $s^A = \left\{s_1^A, \ldots, s_{n^A}^A\right\}$ denote a possible realization of S^A. We assume the existence of the joint probability density function (PDF)

$$f\left(s_1^A, \ldots, s_{n^A}^A\right) \tag{3.2}$$

of the sample locations, along with the existence of the marginal PDF and of the joint PDF

$$f_i\left(s_i^A\right) \quad \text{and} \quad f_{ij}\left(s_i^A, s_j^A\right) \tag{3.3}$$

of s_i^A and s_j^A for $i \neq j$. For example, if the sample S^A is obtained by n^A independent selections of some point, performed uniformly in $\mathcal{U}^A$, we have

$$f_i\left(s_i^A\right) = \frac{1\left(s_i^A \in \mathcal{U}^A\right)}{M^A} \tag{3.4}$$

and $f\left(s_1^A, \ldots, s_{n^A}^A\right) = \prod_{i=1}^{n^A} f_i\left(s_i^A\right)$, with $M^A = \int_{\mathcal{U}^A} dx$ the global measure of the universe, and with $1(\cdot)$ an indicator function.

Suppose that the PDF is absolutely continuous with respect to the Lebesgue measure. As noted by Cordy (1993) and Stevens (1997), there are sampling designs used in practice such that this assumption does not hold true, such as systematic sampling, for example. For any point $x \in \mathcal{U}^A$, the inclusion density function is defined by

$$\pi^A(x) = \sum_{i=1}^{n^A} f_i(x). \tag{3.5}$$

This may be seen as a local measure of the number of sampled points by unit of measure. We have in particular $\int_{\mathcal{U}^A} \pi^A(x)\, dx = n^A$, which is a usual property for sampling designs of fixed size n^A. Similarly, the joint inclusion density function is defined by

$$\pi^A\left(x, x'\right) = \sum_{i=1}^{n^A} \sum_{\substack{j=1 \\ j \neq i}}^{n^A} f_{ij}\left(x, x'\right) \tag{3.6}$$

for $x, x' \in \mathcal{U}^A$.

The HT-estimator of τ_y^A is

$$\hat{\tau}_y^A = \sum_{s \in S^A} d^A(s)\, y^A(s), \tag{3.7}$$

with $d^A(x) = 1/\pi^A(x)$ the design weight of some point x. This estimator is unbiased for τ_y, provided that $\pi^A(x) > 0$ almost everywhere, see Cordy (1993, Theorem 1).

If the function $y^A(\cdot)$ is bounded and $\int_{\mathcal{U}^A} \left\{1/\pi^A(x)\right\} dx < \infty$, then the variance of $\hat{\tau}_y^A$ is given by the Horvitz-Thompson formula

$$\begin{aligned} V\left(\hat{\tau}_y^A\right) = & \int_{x \in \mathcal{U}^A} \frac{\left\{y^A(x)\right\}^2}{\pi^A(x)}\, dx \\ & + \int_{x \in \mathcal{U}^A} \int_{x' \in \mathcal{U}^A} \left\{\pi^A\left(x, x'\right) - \pi^A(x)\, \pi^A(x')\right\} \frac{y^A(x)}{\pi^A(x)} \frac{y^A(x')}{\pi^A(x')}\, dx dx', \end{aligned} \tag{3.8}$$

or equivalently by the Sen-Yates-Grundy formula

$$V\left(\hat{\tau}_y^A\right) = \frac{1}{2}\int_{x\in\mathcal{U}^A}\int_{x'\in\mathcal{U}^A}\left\{\pi^A(x)\,\pi^A(x') - \pi^A(x,x')\right\}\left\{\frac{y^A(x)}{\pi^A(x)} - \frac{y^A(x')}{\pi^A(x')}\right\}^2 dx\,dx', \tag{3.9}$$

see Cordy (1993, Section 2). The corresponding variance estimators are respectively

$$\begin{aligned}\hat{V}_{\mathrm{HT}}\left(\hat{\tau}_y^A\right) = &\sum_{s\in S^A}\left\{\frac{y^A(s)}{\pi^A(s)}\right\}^2 \\ &+ \sum_{s\in S^A}\sum_{\substack{s'\in S^A \\ s'\neq s}}\left\{\frac{\pi^A(s,s') - \pi^A(s)\,\pi^A(s')}{\pi^A(s,s')}\right\}\frac{y^A(s)}{\pi^A(s)}\,\frac{y^A(s')}{\pi^A(s')}\end{aligned} \tag{3.10}$$

and

$$\hat{V}_{\mathrm{YG}}\left(\hat{\tau}_y^A\right) = \frac{1}{2}\sum_{s\in S^A}\sum_{\substack{s'\in S^A \\ s'\neq s}}\left\{\frac{\pi^A(s)\,\pi^A(s') - \pi^A(s,s')}{\pi^A(s,s')}\right\}\left\{\frac{y^A(s)}{\pi^A(s)} - \frac{y^A(s')}{\pi^A(s')}\right\}^2. \tag{3.11}$$

If in addition both $\pi^A(x) > 0$ and $\pi^A(x,x') > 0$ almost everywhere on $\mathcal{U}^A$, then these two variance estimators are unbiased, see Cordy (1993, Theorem 2). Note that the condition on $\pi^A(x,x')$ may not be true, for example when using systematic sampling designs.

3.2 Weight share method: the continuous-discrete case

Suppose that we are still interested in the population U^B and in the estimation of the total τ_y^B given in equation (2.5). The units in U^B are not directly sampled, but a continuous universe $\mathcal{U}^A$ linked to U^B is sampled instead. The links between the units inside the populations $\mathcal{U}^A$ and U^B are represented by the indicator function

$$L^{AB}(x,k) = \begin{cases}1 & \text{if } x\in\mathcal{U}^A \text{ and } k\in U^B \text{ are linked,} \\ 0 & \text{otherwise.}\end{cases} \tag{3.12}$$

We keep the same terminology as in Section 2.2, and we note Anc_k for the *ancestor subset* of some $k\in U^B$, and $\mathrm{Des}(x)$ for the *descendant subset* of some $x\in\mathcal{U}^A$. For any $k\in U^B$,

$$M_{+k}^{AB} = \int_{x\in\mathcal{U}^A} L^{AB}(x,k)\,dx \tag{3.13}$$

is the measure of the ancestor subset of unit k. As in the discrete case, we suppose that $M_{+k}^{AB} > 0$ for any $k\in U^B$.

A sample S^B is obtained in U^B by surveying all the descendants of the points selected in S^A. Formally, we have therefore

$$S^B = \bigcup_{s\in S^A}\mathrm{Des}(s). \tag{3.14}$$

To obtain an estimator of τ_y^B, we establish a duality principle between populations $\mathcal{U}^A$ and U^B. This is summarized in Proposition 1. The duality principle is similar to that obtained with a discrete population: each value y_k^B is equally shared among the points in the ancestor subset of k. The synthetic function $y^A(x)$ may therefore be interpreted as a local measure of density of the variable y^B per unit area. This approach has already been considered in the domain of forest inventory by Mandallaz (2007, Section 4.2) and Gregoire and Valentine (2007, Chapter 10), for example, as discussed in the introduction.

Proposition 1. *The total τ_y^B may be written as*

$$\tau_y^B = \int_{x\in\mathcal{U}^A} y^A(x)\,dx \tag{3.15}$$

with

$$y^A(x) = \sum_{k\in U^B} \frac{L^{AB}(x,k)\,y_k^B}{M_{+k}^{AB}}.$$

Proof. *We have*

$$\begin{aligned}\tau_y^B &= \sum_{k\in U^B} y_k^B = \sum_{k\in U^B} y_k^B \times \frac{1}{M_{+k}^{AB}} \int_{x\in\mathcal{U}^A} L^{AB}(x,k)\,dx \\ &= \int_{x\in\mathcal{U}^A} \sum_{k\in U^B} \frac{L^{AB}(x,k)\,y_k^B}{M_{+k}^{AB}}\,dx = \int_{x\in\mathcal{U}^A} y^A(x)\,dx.\end{aligned}$$

Proposition 1 makes it possible to rewrite τ_y^B as an integral over the universe $\mathcal{U}^A$, and therefore to make use of the extended HT-estimator given in (3.7). In turn, this estimator may be written as a weighted sum over the sample S^B. This is summarized in Proposition 2.

Proposition 2. *The total τ_y^B may be unbiasedly estimated by*

$$\hat{\tau}_y^B = \sum_{k\in S^B} w_k^B y_k^B = \sum_{s\in S^A} d^A(s)\,y^A(s) \tag{3.16}$$

with

$$w_k^B = \frac{1}{M_{+k}^{AB}} \sum_{s\in S^A} L^{AB}(s,k)\,d^A(s).$$

Proof. *We can rewrite*

$$\begin{aligned}\hat{\tau}_y^B &= \sum_{s\in S^A}\sum_{k\in S^B} \frac{L^{AB}(s,k)\,y_k^B d^A(s)}{M_{+k}^{AB}} \\ &= \sum_{s\in S^A}\sum_{k\in U^B} \frac{L^{AB}(s,k)\,y_k^B d^A(s)}{M_{+k}^{AB}}\end{aligned}$$

where the last equality follows from the fact that if $s \in S^A$, *all the units* k *in its descendant subset* $\mathrm{Des}(x)$ *are selected in* S^B. *It follows that*

$$\hat{\tau}_y^B = \sum_{s \in S^A} d^A(s) \sum_{k \in U^B} \frac{L^{AB}(s,k)\, y_k^B}{M_{+k}^{AB}} = \sum_{s \in S^A} d^A(s)\, y^A(s),$$

which is simply the HT-estimator of the integral $\int_{x \in \mathcal{U}^A} y^A(x)\, dx$.

The weight share method thus brings a solution to the estimation of τ_y^B, by using a weighted estimator computed on the sample S^B where the weights w_k^B are given in equation (3.16). Each unit $k \in S^B$ is given the sum of the weights of the sampled points $s \in S^A$ which are linked to k, divided by M_{+k}^{AB}, the measure of the ancestor subset of unit k. The principle is therefore the same as with the usual weight share method applied to discrete populations. It is important to note that, for any unit $k \in S^B$, we need to know the measure M_{+k}^{AB} of its ancestor subset.

Since $\hat{\tau}_y^B$ may be written as a HT-estimator on the sample S^A, the variance is obtained from equation (3.8) or equation (3.9), by using the synthetic variable $y^A(x)$ given in equation (3.15). A variance estimator can be obtained by applying equation (3.10) or equation (3.11). Therefore, variance estimation is straightforward for the weight share estimator.

3.3 Application to plot sampling for forest inventories

We first consider an application of the weight share method to the case where a forest inventory is performed by direct plot sampling. We are interested in a population U^B of trees located on a territory $\mathcal{U}^A$. A sample of points S^A is first selected in $\mathcal{U}^A$ by using a continuous sampling design. For each point $s \in S^A$, the circle $C_r(s)$ centered on s with some predetermined radius r is drawn. All the trees $k \in U^B$ such that their center x_k is inside these circles are selected in the sample S^B and surveyed.

The link function is therefore

$$L^{AB}(x,k) = \begin{cases} 1 \text{ if } x_k \in C_r(x), \\ 0 \text{ otherwise,} \end{cases} = \begin{cases} 1 \text{ if } x \in C_r(x_k), \\ 0 \text{ otherwise.} \end{cases} \tag{3.17}$$

For any tree $k \in U^B$, the quantity M_{+k}^{AB} is

$$M_{+k}^{AB} = \int_{x \in \mathcal{U}^A} L^{AB}(x,k)\, dx = \int_{x \in \mathcal{U}^A} 1\{x \in C_r(x_k)\}\, dx, \tag{3.18}$$

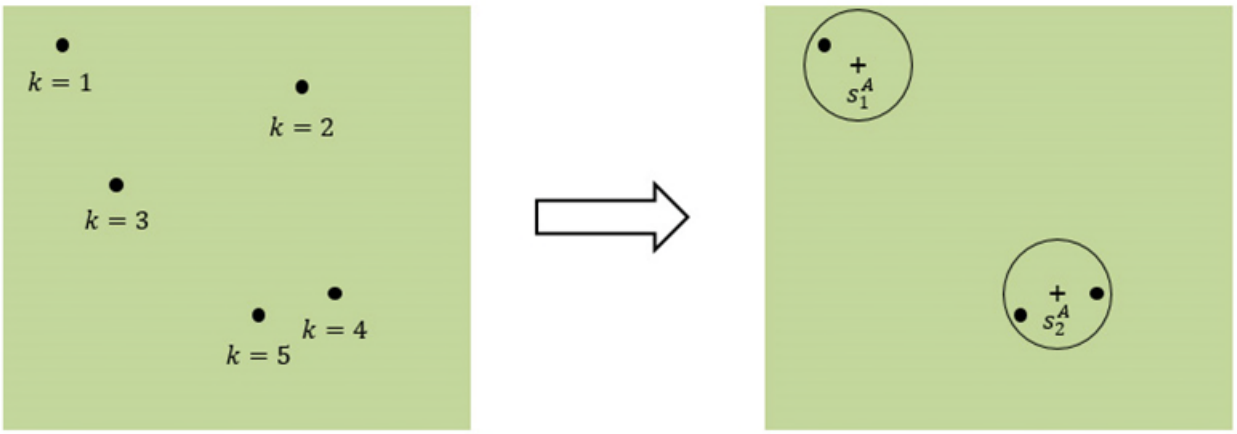
which is the area of the intersection between $\mathcal{U}^A$ and the circle centered on x_k. For any point $x \in \mathcal{U}^A$, the synthetic variable is

$$y^A(x) = \sum_{k \in U^B} \frac{L^{AB}(x,k)\, y_k^B}{M_{+k}^{AB}} = \sum_{x_k \in C_r(x)} \frac{y_k^B}{M_{+k}^{AB}}. \tag{3.19}$$

The solution obtained in this case is equivalent to the aggregation function in Stevens and Urquhart (2000, Section 4.2), to the attribute density in Gregoire and Valentine (2007, Section 10.2) or to the local density in Mandallaz (2007, equation 4.5).

For illustration, we present a toy example in Figure 3.1. We are interested in a rectangular territory $\mathcal{U}^A$ with a global measure $M^A = 7 \times 8 = 56\, m^2$. Inside this area, we have a population U^B of $N^B = 5$ trees. A sample of $n^A = 2$ points is selected by independent drawings with the marginal PDF given in (3.4), which leads to the observation of, say, $s^A = \{s_1^A, s_2^A\}$.

Figure 3.1 A simple example of links between a continuous population $\mathcal{U}^A$ and a discrete population U^B.



For each point $s \in s^A$, the circle $C_r(s)$ centered on s with radius $r = 1$ is drawn, and all the trees k such that their center x_k is inside these circles are surveyed. In the example presented in Figure 3.1, we obtain $s^B = \{1, 4, 5\}$. The values of the variable of interest for the units in s^B are, say:

$$y_1^B = 1,\ y_4^B = 4,\ y_5^B = 3.$$

We compute the quantities M_{+k}^{AB} for the trees $k \in s^B$, applying equation (3.18). For the trees $k = 4$ and 5 the circle $C_r(x_k)$ is included in $\mathcal{U}^A$, resulting in $M_{+k}^{AB} \simeq 3.14$. For the tree $k = 1$, we have $M_{+1}^{AB} \simeq 3.00$. We compute the synthetic variable $y^A(x)$ for the points in s^A, making use of equation (3.19). We obtain

$$y^A(s_1) = \frac{y_1^B}{3.00} \simeq 0.33,$$

$$y^A(s_2) = \frac{y_4^B}{3.14} + \frac{y_5^B}{3.14} \simeq 2.23.$$

Finally, we compute the weights w_k^B for the trees $k \in s^B$ by means of the weight share method, making use of equation (3.16). We obtain

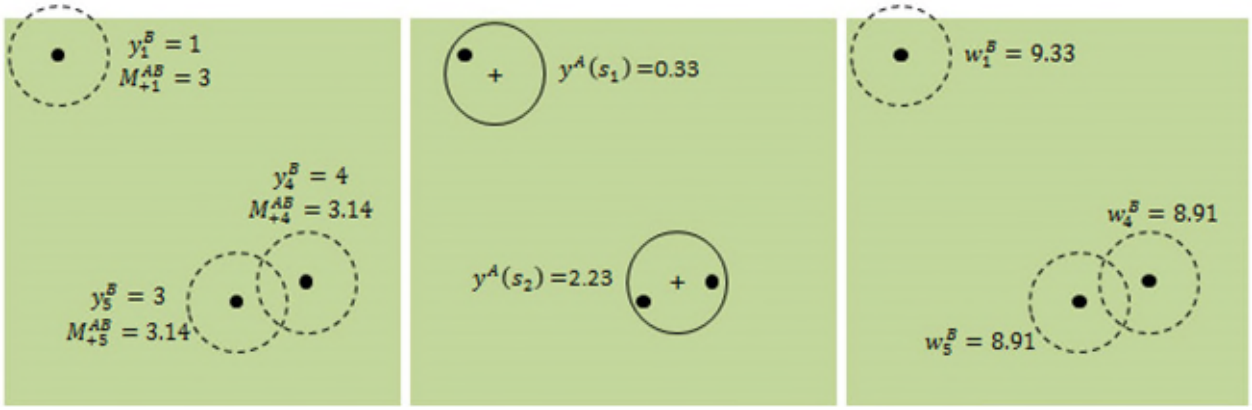
$$w_1^B = \frac{d^A(s_1)}{M_{+1}^{AB}} = \frac{28}{3.00} \simeq 9.33,$$
$$w_4^B = \frac{d^A(s_2)}{M_{+4}^{AB}} = \frac{28}{3.14} \simeq 8.91,$$
$$w_5^B = \frac{d^A(s_2)}{M_{+5}^{AB}} = \frac{28}{3.14} \simeq 8.91.$$

The estimate for the total τ_y^B is therefore

$$\hat{\tau}_y^B(s^B) = \sum_{k \in s^B} w_k^B y_k^B \simeq 71.72.$$

The results are summarized in Figure 3.2.

Figure 3.2 Computation of the synthetic variable $y^A(x)$ and of the weights w_k^B obtained by the weight share method on a simple example.

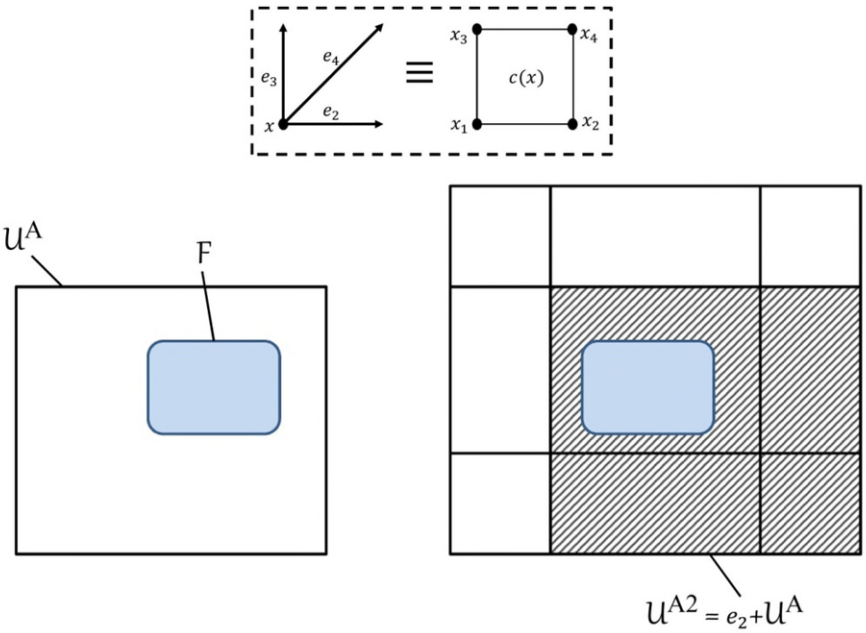


3.4 Application to spatial cluster sampling for forest inventories

When the accessibility to the field is difficult, it is common practice in forest inventories to use clusters of plots (Köhl and Magnussen, 2015). These clusters have a fixed geometric form determined before the survey. For instance, plots may be positioned at the corners of a square of 50m size (see Mandallaz, 2007, Section 4.3).

Suppose that we are again interested in a population U^B of trees located on a forest $\mathcal{F}$. Let $\mathcal{U}^A$ denote a set such that $\mathcal{F} \subset \mathcal{U}^A$, and let $e_1, \ldots, e_L$ denote a set of L vectors in $\mathbf{R}^2$. For any point $x \in \mathcal{U}^A$, the cluster $c(x)$ is defined as the set of points $\{x_l \equiv x + e_l;\ l = 1, \ldots, L\}$. Following the notation by Mandallaz (2007), we take e_1 as the null vector, and $x_1 = x$ is seen as the origin of the cluster. Let us denote by $\mathcal{U}^{Al} = e_l + \mathcal{U}^A$ for $l = 1, \ldots, L$, and $\mathcal{V}^C = \bigcup_{l=1}^{L} \mathcal{U}^{Al}$ their union. It is also supposed that the set $\mathcal{U}^A$ is large enough to ensure that $\mathcal{F} \subset \mathcal{U}^{Al}$ for any $l = 1, \ldots, L$ (see Mandallaz, 2007, Section 4.3). For illustration, an example in the case $L = 4$ is presented in Figure 3.3.

Figure 3.3 An example of cluster sampling in continuous populations.



Lecture note. In the upper panel, an example of cluster of size $L = 4$ originated in the point x is presented. In the left panel, two continuous populations such that $\mathcal{F} \subset \mathcal{U}^A$ are given. In the right panel, the sets $\mathcal{U}_{Al}, l = 1, \ldots, 4$ associated to $\mathcal{U}^A$ are presented, where the hatched set stands for $\mathcal{U}_{A2}$.

Cluster sampling is performed by first selecting a sample S^A of n^A points in $\mathcal{U}^A$, according to a continuous sampling design. For each point $x \in S^A$, we obtain the associated cluster $c(x) = \{x_1, \ldots, x_L\}$ originated in $x_1 = x$, and the associated circles $C_r(x_l)$ are drawn with some predetermined radius r. All the trees $k \in U^B$ such that their center x_k is inside one of the circles $C_r(x_l), l = 1, \ldots, L$ for some $x \in S^A$ are surveyed. An example is presented in Figure 3.4.

For any point $x \in \mathcal{U}^A$, the synthetic variable $y^A(x)$ is obtained in two steps, using $\mathcal{V}^C$ as a pivotal population. We first define a link function between $\mathcal{V}^C$ and U^B as

$$L^{CB}(z,k) = \begin{cases} 1 \text{ if } x_k \in C_r(z), \\ 0 \text{ otherwise,} \end{cases} \tag{3.20}$$

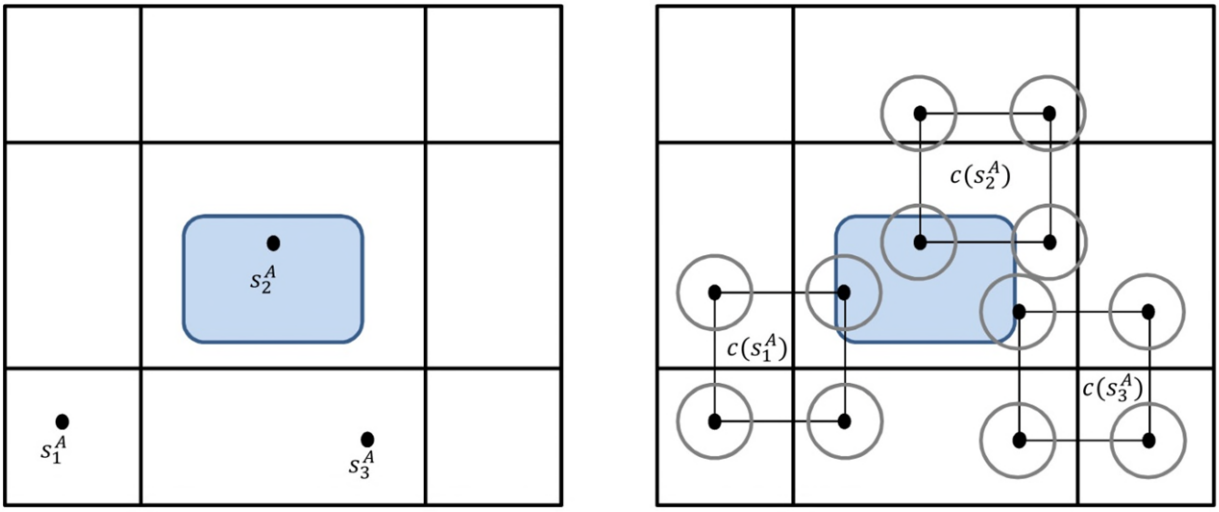
with x_k the center of the tree k and z a point in the union set $\mathcal{V}^C$. This is similar to the link function defined in (3.17), and following the same lines of reasoning we obtain the intermediary synthetic variable

$$y^C(z) = \sum_{x_k \in C_r(z)} \frac{y_k^B}{M_{+k}^{CB}} \quad \text{for any} \quad z \in \mathcal{V}^C, \tag{3.21}$$

where $M_{+k}^{CB} = \int_{z \in \mathcal{V}^C} 1\{z \in C_r(x_k)\}\, dz$ is the area of the intersection between $\mathcal{V}^C$ and the circle centered on x_k. Then, we define a link function between $\mathcal{U}^A$ and $\mathcal{V}^C$ as

$$L^{AC}(x,z) = \begin{cases} 1 \text{ if } z \in c(x) = \{x_1, \ldots, x_L\}, \\ 0 \text{ otherwise.} \end{cases} \tag{3.22}$$

Figure 3.4 An example of cluster sampling where $n^A = 3$ points are initially selected from $\mathcal{U}^A$.



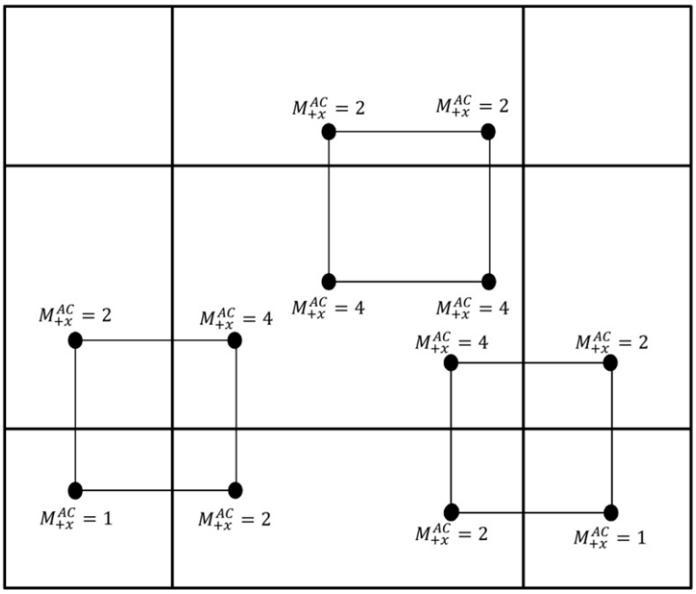
In other words, x and z are linked if z is one of the vertices of the cluster originated in x. The quantity

$$M_{+z}^{AC} = \int_{x \in \mathcal{U}^A} L^{AC}(x, z)\, dx \tag{3.23}$$

is the number of clusters having z as a vertex. We give, in Figure 3.5, the values obtained for the vertices of the clusters in the example initiated in Figure 3.4. We finally obtain the synthetic variable

$$y^A(x) = \sum_{z \in c(x)} \frac{y^C(z)}{M_{+z}^{AC}} = \sum_{z \in c(x)} \frac{1}{M_{+z}^{AC}} \sum_{x_k \in C_r(z)} \frac{y_k^B}{M_{+k}^{CB}} \quad \text{for any} \quad x \in \mathcal{U}^A. \tag{3.24}$$

Figure 3.5 Values of the quantities M_{+x}^{AC} for the vertices in a cluster sample.



As proved in Proposition 2, τ_y^B may be unbiasedly estimated by $\hat{\tau}_y^B$, and an unbiased variance estimator is directly obtained by applying equation (3.10) or equation (3.11). Therefore, the weight share method provides a general solution for both estimation and variance estimation for cluster sampling, under an arbitrary sampling design performed in $\mathcal{U}^A$. On the other hand, the solution described in Mandallaz (2007, equation 4.17) is suitable only when S^A is obtained by independent uniform selections from $\mathcal{U}^A$.

4. Simulation study

In this simulation study, we consider estimation and variance estimation for spatial cluster sampling. We wish to compare the weight share method with the solution described in Mandallaz (2007, equation 4.17), in the situation when the sample S^A is obtained by independent uniform selections from $\mathcal{U}^A$. This is the purpose of the first simulation study described in Section 4.1. We also wish to evaluate the weight share method in the situation when the sample S^A of cluster origins is not selected by independent uniform sampling, and when Mandallaz's technique may therefore not be applied. This is the purpose of the second simulation study described in 4.2.

The continuous population $\mathcal{U}^A$ that we consider is a square of length 1,000 meters. Inside $\mathcal{U}^A$, the forest is located on a square territory $\mathcal{F}$ of length 600 meters, but the location of the forest and its area are seen as unknown prior to the survey. A population of $N^B = 30{,}942$ black pines is generated in the forest $\mathcal{F}$. The main characteristics of the population of pines in terms of volume (y_{1k}^B, cube meters) and breast-height diameter (y_{2k}^B, centimeters) are summarized in Table 4.1.

Table 4.1
Mean, standard deviation, minimum and maximum for the breast-height diameter and volume for the population of black pines

	Mean	Standard deviation	Minimum	Maximum
Breast-height diameter (centimeters)	17.28	6.38	3.57	41.78
Volume (cube meters)	0.17	0.16	0.00	0.99

We are interested in estimating the following parameters: the total number of trees $\tau_{y0}^B = \sum_{k \in U^B} y_{0k}^B$ with $y_{0k}^B = 1$, the total volume of wood $\tau_{y1}^B = \sum_{k \in U^B} y_{1k}^B$, the mean volume of trees $\mu_{y1}^B = \tau_{y1}^B / N^B$, the mean breast-height diameter $\mu_{y2}^B = \tau_{y2}^B / N^B$, the average volume of wood per square meter of forest $\bar{Y}_1^B = \tau_{y1}^B / M^{\mathcal{F}}$, where $M^{\mathcal{F}}$ stands for the area of the forest $\mathcal{F}$.

4.1 Comparison between the weight share method and Mandallaz's estimator for cluster sampling

We first compare the performance of the proposed estimators to those proposed by Mandallaz (2007, Section 4.3), in case of cluster sampling (see Section 3.4) where the sample S^A of cluster origins is selected by independent uniform sampling from the territory $\mathcal{U}^A$. We select a sample S^A of size $n^A = 100, 200$ or 400. A sample of trees is then selected and surveyed by using the cluster sampling

technique described in Section 3.4, where we use square clusters of size $L = 4$ and length 60 meters, and plots with a radius r of 25 meters.

For a given sample, the estimators under the weight share method are obtained as follows. The estimators of the totals τ_{y0}^B and τ_{y1}^B are obtained from equation (3.16) as

$$\hat{\tau}_{y0}^B = \frac{M^A}{n^A} \sum_{x \in S^A} y_0^A(x), \tag{4.1}$$

$$\hat{\tau}_{y1}^B = \frac{M^A}{n^A} \sum_{x \in S^A} y_1^A(x),$$

where y_0^A and y_1^A are obtained by plugging into equation (3.24) the variables y_{0k}^B and y_{1k}^B, respectively, and where M^A is the area of $\mathcal{U}^A$. The estimators of the population means μ_{y1}^B and μ_{y2}^B are

$$\hat{\mu}_{y1}^B = \frac{\hat{\tau}_{y1}^B}{\hat{\tau}_{y0}^B} \quad \text{and} \quad \hat{\mu}_{y2}^B = \frac{\hat{\tau}_{y2}^B}{\hat{\tau}_{y0}^B},$$

respectively, obtained by using the plug-in principle, and where $\hat{\tau}_{y2}^B$ is obtained as described in equation (4.1). The estimator of $\overline{Y}_1^B$ is

$$\hat{\overline{Y}}_1^B = \frac{\hat{\tau}_{y1}^B}{\hat{M}^{\mathcal{F}}},$$

where

$$\hat{M}^{\mathcal{F}} = \frac{M^A}{n^A} \sum_{x \in S^A} y_3^A(x) \quad \text{with} \quad y_3^A(x) = \sum_{z \in c(x)} \frac{1\{z \in \mathcal{F}\}}{M_{+z}^{AC}} \tag{4.2}$$

is an unbiased estimator of $M^{\mathcal{F}}$. The estimators proposed by Mandallaz are obtained as follows. The estimator of $\overline{Y}_1^B$ is

$$\hat{\overline{Y}}_{1,\text{mand}}^B = \frac{\sum_{x \in S^A} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_r(z)} \frac{y_{1k}^B}{M_{+k}^{\mathcal{F}}}}{\sum_{x \in S^A} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\}}$$

with $M_{+k}^{\mathcal{F}}$ the area of the intersection between $\mathcal{F}$ and the circle centered on x_k, see equation (4.17) in Mandallaz (2007). The estimators of the totals τ_{y0}^B and τ_{y1}^B are

$$\hat{\tau}_{y0,\text{mand}}^B = \frac{1}{L} \frac{M^A}{n^A} \sum_{x \in S^A} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_r(z)} \frac{y_{0k}^B}{M_{+k}^{\mathcal{F}}},$$

$$\hat{\tau}_{y1,\text{mand}}^B = \frac{1}{L} \frac{M^A}{n^A} \sum_{x \in S^A} \sum_{z \in c(x)} 1\{z \in \mathcal{F}\} \sum_{k \in C_r(z)} \frac{y_{1k}^B}{M_{+k}^{\mathcal{F}}},$$

see equation (4.16) in Mandallaz (2007). The estimators of the population means μ_{y1}^B and μ_{y2}^B are

$$\hat{\mu}^{B}_{y1,\text{mand}} = \frac{\hat{\tau}^{B}_{y1,\text{mand}}}{\hat{\tau}^{B}_{y0,\text{mand}}} \quad \text{and} \quad \hat{\mu}^{B}_{y2,\text{mand}} = \frac{\hat{\tau}^{B}_{y2,\text{mand}}}{\hat{\tau}^{B}_{y0,\text{mand}}},$$

respectively, obtained by using the plug-in principle.

The sampling and estimation steps are repeated $D = 10{,}000$ times. For an estimator $\hat{\theta}$ of a parameter θ, we compute the Monte Carlo Percent Relative Bias

$$\text{RB}\{\hat{\theta}\} = 100 \times \frac{D^{-1}\sum_{d=1}^{D}\hat{\theta}_d - \theta}{\theta}, \tag{4.3}$$

and the Monte Carlo Mean Square Error

$$\text{MSE}\{\hat{\theta}\} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d - \theta\right)^2. \tag{4.4}$$

The results are given in Table 4.2. For any of the five parameters, both estimation methods lead to virtually unbiased estimators, and the mean square errors are very close. The estimator of Mandallaz performs slightly better for $\bar{Y}_1^B$, while the weight share method performs slightly better for the other parameters.

Table 4.2
Percent relative bias and mean square error for five parameters estimated by means of the weight share method or by estimators proposed by Mandallaz

			τ^B_{y0}	τ^B_{y1}	μ^B_{y1}	μ^B_{y2}	$\bar{Y}_1^B$
$n^A = 100$	Weight share	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$2.04\ 10^{7}$	$6.24\ 10^{5}$	$1.22\ 10^{-6}$	$1.81\ 10^{-3}$	$8.76\ 10^{-7}$
	Mandallaz	RB (%)	-0.02	-0.02	0.00	0.00	-0.17
		MSE	$2.06\ 10^{7}$	$6.29\ 10^{5}$	$1.25\ 10^{-6}$	$1.85\ 10^{-3}$	$8.65\ 10^{-7}$
$n^A = 200$	Weight share	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$9.91\ 10^{6}$	$3.03\ 10^{5}$	$6.07\ 10^{-7}$	$9.09\ 10^{-4}$	$4.15\ 10^{-7}$
	Mandallaz	RB (%)	-0.01	-0.01	0.00	0.00	-0.15
		MSE	$9.99\ 10^{6}$	$3.05\ 10^{5}$	$6.20\ 10^{-7}$	$9.27\ 10^{-4}$	$4.08\ 10^{-7}$
$n^A = 400$	Weight share	RB (%)	-0.02	-0.02	0.00	0.00	-0.03
		MSE	$5.03\ 10^{6}$	$1.54\ 10^{5}$	$2.99\ 10^{-7}$	$4.44\ 10^{-4}$	$2.15\ 10^{-7}$
	Mandallaz	RB (%)	-0.02	-0.03	0.00	0.00	-0.04
		MSE	$5.07\ 10^{6}$	$1.55\ 10^{5}$	$3.05\ 10^{-7}$	$4.52\ 10^{-4}$	$2.12\ 10^{-7}$

Note: Relative bias (RB); Mean square error (MSE).

For the estimators obtained under the weight share method, we also considered variance estimation. We did not perform variance estimation for Mandallaz's estimators, since in Mandallaz (2007, Section 4.3), variance estimators are only proposed for spatial means like $\bar{Y}_1^B$, and not for population totals or population means. Under independent uniform sampling from $\mathcal{U}^A$, applying equation (3.10) leads to the unbiased variance estimator

$$\hat{V}_{\mathrm{HT}}\left(\hat{\tau}_{y0}^{B}\right) = \frac{\left(M^{A}\right)^{2}}{n^{A}} \times \frac{1}{n^{A}-1}\sum_{x\in S^{A}}\left\{y_{0}^{A}(x) - \frac{1}{n^{A}}\sum_{s\in S^{A}} y_{0}^{A}(s)\right\}^{2}, \tag{4.5}$$

for $\hat{\tau}_{y0}^{B}$, and a similar expression for $\hat{\tau}_{y1}^{B}$. An approximately unbiased variance estimator for $\hat{\mu}_{y1}^{B}$ is obtained by replacing in (4.5) the variable $y_{0}^{A}(x)$ by the linearized variable

$$u_{1}^{A}(x) = \frac{1}{\hat{\tau}_{y0}^{B}}\left\{y_{1}^{A}(x) - \hat{\mu}_{y1}^{B}\, y_{0}^{A}(x)\right\}, \tag{4.6}$$

and can similarly be obtained for $\hat{\mu}_{y2}^{B}$. An approximately unbiased variance estimator for $\hat{\bar{Y}}_{1}^{B}$ is obtained by replacing in (4.5) the variable $y_{0}^{A}(x)$ by the linearized variable

$$u_{2}^{A}(x) = \frac{1}{\hat{\tau}_{y3}^{B}}\left\{y_{1}^{A}(x) - \hat{\bar{Y}}_{1}^{B} y_{3}^{A}(x)\right\}. \tag{4.7}$$

The sampling and estimation steps are repeated $E = 1{,}000$ times. To measure the bias of a variance estimator $\hat{V}_{\mathrm{HT}}(\hat{\theta})$, we use the Monte Carlo Percent Relative Bias

$$\mathrm{RB}\left\{\hat{V}_{\mathrm{HT}}(\hat{\theta})\right\} = 100 \times \frac{E^{-1}\sum_{e=1}^{E}\hat{V}_{\mathrm{HT}}\left(\hat{\theta}_{e}\right) - \mathrm{MSE}(\hat{\theta})}{\mathrm{MSE}(\hat{\theta})}, \tag{4.8}$$

where $\mathrm{MSE}(\hat{\theta})$ is obtained independently from the first run of $D = 10{,}000$ simulations, see equation (4.4). The results are presented in Table 4.3. All the variance estimators are approximately unbiased.

Table 4.3
Percent relative bias of a variance estimator for five parameters estimated by means of the weight share method

n^{A}	$\hat{V}\left(\hat{\tau}_{y0}^{B}\right)$	$\hat{V}\left(\hat{\tau}_{y1}^{B}\right)$	$\hat{V}\left(\hat{\mu}_{y1}^{B}\right)$	$\hat{V}\left(\hat{\mu}_{y2}^{B}\right)$	$\hat{V}\left(\hat{\bar{Y}}_{1}^{B}\right)$
100	-1.21	-1.17	-0.48	-0.04	-0.17
200	1.75	1.83	-0.59	-0.76	3.35
400	0.49	0.62	0.40	1.09	-0.29

4.2 Evaluation of the weight share method for cluster sampling with non-uniform sampling of the clusters

We consider a second case of cluster sampling when the sample S^{A} is not selected by independent uniform sampling from $\mathcal{U}^{A}$, so that the estimators proposed by Mandallaz (2007) can not be used. The population $\mathcal{U}^{A}$ is first partitioned into a sub-population $\mathcal{U}_{1}^{A}$ of length 300 meters and height 1,000 meters (west part), and a sub-population $\mathcal{U}_{2}^{A}$ of length 700 meters and height 1,000 meters (east part). The area of $\mathcal{U}_{1}^{A}$ and $\mathcal{U}_{2}^{A}$ are denoted as M_{1}^{A} and M_{2}^{A}, respectively. A sample of size $n_{1} = 75,\ 150$ or 300 points is first selected globally from $\mathcal{U}^{A}$, and a second sample of size $n_{2} = 25,\ 50$ or 100 is then selected from $\mathcal{U}_{2}^{A}$. For example, this case may arise if it is of specific interest to perform estimation on the sub-population $\mathcal{U}_{2}^{A}$, and if an extension sample may be funded to ensure that the global sample selected from

$\mathcal{U}_2^A$ is sufficiently large. The union of these two samples is denoted as S^A. We also let n_1^A denote the (random) size of $S_1^A = S^A \cap \mathcal{U}_1^A$, and n_2^A denote the (random) size of $S_2^A = S^A \cap \mathcal{U}_2^A$.

Conditionally on n_1^A and n_2^A, unbiased estimators for the totals τ_{y0}^B and τ_{y1}^B are

$$\hat{\tau}_{y0,\text{cond}}^B = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_0^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_0^A(x),$$

$$\hat{\tau}_{y1,\text{cond}}^B = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_1^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_1^A(x),$$

where y_0^A and y_1^A are obtained by plugging into equation (3.24) the variables y_{0k}^B and y_{1k}^B, respectively. The estimators of the population means μ_{y1}^B and μ_{y2}^B are

$$\hat{\mu}_{y1,\text{cond}}^B = \frac{\hat{\tau}_{y1,\text{cond}}^B}{\hat{\tau}_{y0,\text{cond}}^B} \quad \text{and} \quad \hat{\mu}_{y2,\text{cond}}^B = \frac{\hat{\tau}_{y2,\text{cond}}^B}{\hat{\tau}_{y0,\text{cond}}^B},$$

respectively, obtained by using the plug-in principle. The estimator of $\bar{Y}_1^B$ is

$$\hat{\bar{Y}}_{1,\text{cond}}^B = \frac{\hat{\tau}_{y1,\text{cond}}^B}{\hat{M}_{\text{cond}}^{\mathcal{F}}},$$

where

$$\hat{M}_{\text{cond}}^{\mathcal{F}} = \frac{M_1^A}{n_1^A} \sum_{x \in S_1^A} y_3^A(x) + \frac{M_2^A}{n_2^A} \sum_{x \in S_2^A} y_3^A(x)$$

is an unbiased estimator of $M^{\mathcal{F}}$, where $y_3^A(x)$ is defined in equation (4.2).

The sampling and estimation steps are repeated $D = 10{,}000$ times. For an estimator $\hat{\theta}$ of a parameter θ, we compute the Monte Carlo Percent Relative Bias given in (4.3) and the Monte Carlo Mean Square Error given in (4.4). The results are given in Table 4.4. For any of the five parameters, the estimators are virtually unbiased. The mean square error decreases as the sample size increases, as could be expected.

Table 4.4
Percent relative bias and Mean square error for five parameters estimated by means of the weight share method

		τ_{y0}^B	τ_{y1}^B	μ_{y1}^B	μ_{y2}^B	$\bar{Y}_1^B$
$n^A = 100$	RB (%)	0.03	0.03	0.00	0.00	-0.02
	MSE	$1.85\ 10^7$	$5.67\ 10^5$	$1.19\ 10^{-6}$	$1.78\ 10^{-3}$	$8.5\ 10^{-7}$
$n^A = 200$	RB (%)	-0.02	-0.02	0.01	0.00	-0.06
	MSE	$9.40\ 10^6$	$2.87\ 10^5$	$5.83\ 10^{-7}$	$8.70\ 10^{-4}$	$4.34\ 10^{-7}$
$n^A = 400$	RB (%)	0.09	0.09	0.00	0.00	0.05
	MSE	$4.67\ 10^6$	$1.43\ 10^5$	$2.90\ 10^{-7}$	$4.35\ 10^{-4}$	$2.18\ 10^{-7}$

Note: Relative bias (RB); Mean square error (MSE).

We now consider variance estimation. Conditionally on n_1^A and n_2^A, an unbiased variance estimator for $\hat{\tau}_{y0,\text{cond}}^B$ is

$$\hat{V}_2\left(\hat{\tau}_{y0}^B\right) = \frac{\left(M_1^A\right)^2}{n_1^A} \times \frac{1}{n_1^A - 1} \sum_{x \in S_1^A} \left\{ y_0^A(x) - \frac{1}{n_1^A} \sum_{s \in S_1^A} y_0^A(s) \right\}^2 + \frac{\left(M_2^A\right)^2}{n_2^A} \times \frac{1}{n_2^A - 1} \sum_{x \in S_2^A} \left\{ y_0^A(x) - \frac{1}{n_2^A} \sum_{s \in S_2^A} y_0^A(s) \right\}^2, \quad (4.9)$$

and can be expressed similarly for $\hat{\tau}_{y1,\text{cond}}^B$. An approximately unbiased variance estimator for $\hat{\mu}_{y1,\text{cond}}^B$ is obtained by replacing in (4.9) the variable $y_0^A(x)$ by the linearized variable given in equation (4.6), and can be obtained similarly for $\hat{\mu}_{y2}^B$. An approximately unbiased variance estimator for $\hat{\bar{Y}}_{1,\text{cond}}^B$ is obtained by replacing in (4.9) the variable $y_0^A(x)$ by the linearized variable given in equation (4.7).

The sampling and estimation steps are repeated $E = 1{,}000$ times. To measure the bias of a variance estimator $\hat{V}_2(\hat{\theta})$, we use the Monte Carlo Percent Relative Bias defined in (4.3). The results are presented in Table 4.5. All the variance estimators are approximately unbiased.

Table 4.5
Percent relative bias of a variance estimator for five parameters estimated by means of the weight share method

n^A	$\hat{V}_2\left(\hat{\tau}_{y0}^B\right)$	$\hat{V}_2\left(\hat{\tau}_{y1}^B\right)$	$\hat{V}_2\left(\hat{\mu}_{y1}^B\right)$	$\hat{V}_2\left(\hat{\mu}_{y2}^B\right)$	$\hat{V}_2\left(\hat{\bar{Y}}_1^B\right)$
100	3.18	3.16	-3.02	-2.96	-0.28
200	0.97	0.92	-1.59	-1.01	-1.88
400	1.12	1.03	0.05	-0.36	-1.46

5. Discussion

There are several reasons in practice why a population has no tractable sampling frame. When the population of interest may be linked to a discrete population for which a sampling frame is available, the usual weight share method (Deville and Lavallée, 2006) makes it possible to obtain probability samples from the population of interest, as well as unbiased estimators and variance estimators for this population. We showed that this approach may be generalized in a natural way when the population of interest is linked to a continuous population, by using a synthetic function on this continuous population which may be interpreted as a local measure of density. In case of spatial cluster sampling with independent uniform selection of the cluster origins, our simulation results show that the weight share method and Mandallaz's estimator perform similarly. Mandallaz's estimator can not be applied when the cluster origins are not selected by independent uniform sampling. In this case, our simulation results confirm that the weight share method leads to unbiased estimators.

In our view, making use of the weight share method has several advantages. Firstly, it enables to easily handle the so-called edge corrections, i.e., the fact that some units have an ancestor subset which intersects

with the area of interest. Using the area of the ancestor subset M_{+k}^{AB} in the estimation weights (see equation (3.16)) leads to exactly unbiased estimators, while alternative edge corrections may be somewhat cumbersome, see Gregoire and Valentine (2007, Section 10.7) or Roesch, Green and Scott (1993), for example. Also, our approach enables us to go back to the population which was indeed sampled. This is necessary to make use of the Horvitz-Thompson estimator, and to compute an unbiased variance estimator. This is possible in full generality, i.e., under an arbitrary continuous sampling design, by using the theory developed by Cordy (1993).

This method is obviously not limited to forest inventory. One example is the survey on crop practices conducted by the French statistical office of the Ministry of Agriculture, until 2006. The sample for this survey consisted of parcels, selected from the points of the survey Ter-Uti (Chapelle-Barry, 2008). The Ter-Uti survey is dedicated to the production of statistics on the land use, and the points where the land use was associated to crop practices are the basis from which to select the parcels. Calculating weights for the parcels was done by considering that each parcel had a probability of being drawn proportional to its surface. In this way, the method produced the same weights than the weight share method, except for edge effects. However, using the weight share method would lead to have a general methodological framework enabling the derivation of a variance estimator. Other topics, linked to environmental issues, could also take advantage from the application of this method. This extension of the weight share method should also be considered for estimators that are not directly Horvitz-Thompson estimators, such as those relative to two-phase sampling, which are often used for environmental issues.

Acknowledgements

The authors would like to thank Pierre Lavallée and Minna Pulkkinen for useful discussions. The authors would also like to thank the referee, the associate editor and the assistant editor for helpful comments on an earlier version of the manuscript.

References

Ardilly, P., and Le Blanc, D. (2001). Sampling and weighting a survey of homeless persons: A French example. *Survey Methodology*, 27, 1, 109-118. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5859-eng.pdf.

Chapelle-Barry, C. (2008). Enquête sur les pratiques culturales en 2006. *Agreste Chiffres et Données*.

Clairin, R., and Brion, P. (1996). *Manuel de Sondages : Applications aux Pays en Développement*. Documents et manuels du CEPED, 3.

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18(5), 353-362.

De Vitiis, C., Falorsi, S. and Inglese, F. (2014). Implementing the first ISTAT survey of homeless population by indirect sampling and weight sharing method, Springer. *Contributions to Sampling Statistics*, 119-138.

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

Gregoire, T.G., and Valentine, H.T. (2007). Sampling Strategies for Natural Resources and the Environment. CRC Press.

Himelein, K., Eckman, S. and Murray, S. (2014). Sampling nomads: A new technique for remote, hard-to-reach, and mobile populations. *Journal of Official Statistics*, 30(2), 191-213.

Köhl, M., and Magnussen, S. (2015). Sampling in forest inventories, Springer. *Tropical Forestry Handbook*, 1-50.

Laporte, A., Vandentorren, S., Détrez, M.-A., Douay, C., Le Strat, Y., Le Méner, E., Chauvin, P. and The Samenta Research Group (2018). Prevalence of mental disorders and addictions among homeless people in the greater Paris area, France. *International Journal of Environmental Research and Public Health*, 15(2), 241.

Lavallée, P. (2009). *Indirect Sampling*. Springer Science & Business Media.

Lawrence, M., McRoberts, R.E., Tomppo, E., Gschwantner, T. and Gabler, K. (2010). Comparisons of national forest inventories, Springer. *National Forest Inventories*, 19-32.

Lohlé-Tart, L., Clairin, R., François, M. and Gendreau, F. (1988). De l'Homme au Chiffre. Réflexions sur l'Observation Démographique en Afrique.

Lorenc, B., Smith, P.A. and Bavdaž, M. (2018). *The Unit Problem and Other Current Topics in Business Survey Methodology*. Cambridge Scholars Publishing.

Mandallaz, D. (2007). Sampling Techniques for Forest Inventories. CRC Press.

Roesch, F.A., Green, E.J. and Scott, C.T. (1993). An alternative view of forest sampling. *Survey Methodology*, 19, 2, 199-204. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14453-eng.pdf.

Stevens, D. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics: The Official Journal of the International Environmetrics Society*, 8(3), 167-195.

Stevens, D.L., and Urquhart, N.S. (2000). Response designs and support regions in sampling continuous domains. *Environmetrics: The Official Journal of the International Environmetrics Society*, 11(1), 13-41.

Xu, X., and Lavallée, P. (2009). Treatments for link nonresponse in indirect sampling. *Survey Methodology*, 35, 2, 153-164. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11038-eng.pdf.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 163-190
Statistics Canada, Catalogue No. 12-001-X

Modelling time change in survey response rates: A Bayesian approach with an application to the Dutch Health Survey

Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek and Barry Schouten[1]

Abstract

Precise and unbiased estimates of response propensities (RPs) play a decisive role in the monitoring, analysis, and adaptation of data collection. In a fixed survey climate, those parameters are stable and their estimates ultimately converge when sufficient historic data is collected. In survey practice, however, response rates gradually vary in time. Understanding time-dependent variation in predicting response rates is key when adapting survey design. This paper illuminates time-dependent variation in response rates through multi-level time-series models. Reliable predictions can be generated by learning from historic time series and updating with new data in a Bayesian framework. As an illustrative case study, we focus on Web response rates in the Dutch Health Survey from 2014 to 2019.

Key Words: Response propensity; Time series; Multilevel model; Bayesian analysis.

1. Introduction

Over the last two decades, responsive and adaptive design (Chun, Heeringa and Schouten, 2018) have attracted considerable interest in assembling survey design features ahead of or during data collection, with an ultimate goal of survey cost-quality optimization by a search for efficient resource allocation. The emergence of Web surveys, the availability of process data, and the increase in survey costs have driven research regarding the monitoring (Kreuter, 2013) and adaptation (Schouten, Peytchev and Wagner, 2017) of data collection. However, a thorough understanding of how design features and time change affect important parameters in response and cost models is imperative to apply adaptation. For example, a critical factor is the likelihood of a participant to engage in a survey, i.e., their response propensity, which can be sensitive to factors both dependent on and independent from the nature of the survey itself. Additionally, the cost of the survey is a complex calculation that covers everything from planning the survey, to performing it and the data workup afterwards, and it can directly impact the type of survey performed, which can in turn influence response propensities (RPs). For this reason, the development of such parameter measurements is necessary before the data collection operation begins.

The last decade has seen a renewed importance in the predictability of RP for responsive and adaptive design. In survey methodology, using propensity scores (Rosenbaum and Rubin, 1983) is the common way to tailor differential features to sampled cases for desired cost- or quality-related goals. In a changing data collection climate, the performance and structure of a survey design hinge heavily on propensity models that may lead to inefficient decisions. For instance, by relying only on process or response data in the early stages of a responsive survey, the estimates of RP may produce biased estimates of the final RP by the end of the data collection (Wagner and Hubbard, 2014). Also, the uncertainty of RP estimates

1. Shiya Wu, Utrecht University, Department of Methodology and Statistics. E-mail: s.wu@uu.nl; 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.

should be incorporated into propensity models in order to avoid suboptimal designs (Burger, Perryck and Schouten, 2017).

Accurate estimates of RP are thus the crux of survey operations. For this reason, survey researchers apply historic data to estimate the coefficients of a propensity model, and then use those estimated coefficients for the upcoming rounds of a survey. Bayesian analysis (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin, 2013) is a natural approach to utilize both historic and new data for improving predictions. Prior beliefs generated from historic data are evolved into posteriors, which serve as the priors for the subsequent analysis as the upcoming data accumulates. Schouten, Mushkudiani, Shlomo, Durrant, Lundquist and Wagner (2018) were the first to apply a general Bayesian method to analyze RP and cost in the Dutch Health Survey. They discuss that misspecification of the priors may weaken prediction performance. As a result, prior elicitation becomes an influential step. The incorporation of expert beliefs is a prerequisite for such prior elicitation. This has a long history in biometric and medical literature, but the application is in its infancy in the context of surveys. Recent examples have been West, Wagner, Coffey and Elliott (2021), who reviewed empirical evidence for survey propensity prediction, Coffey, West, Wagner and Elliott (2020), who consulted data collection managers about the estimated coefficients, and Wu, Schouten, Meijers and Moerbeek (2022), who used data collection staff as experts for relevant historic leverage under criteria for a new or redesigned survey.

So far, the approaches assume RPs are stable in a relatively short period. In a fixed survey climate, these parameters remain stable and their estimates ultimately converge with the accumulation of historic data. In survey practice however, those parameters change gradually over time, which means that predictions may not converge. For example, seasonal variation and downward trends in response rates can be observed. Thus, the benefit of prior elicitation could potentially be undone when ignoring time change. Recent articles by Mushkudiani and Schouten (2019), and Fang, Burger, Meijers and van Berkel (2020) describe what time-dependent factors significantly affect the parameter estimation accuracy, but the impact on prediction accuracy is still unknown, which is the topic of this paper.

This paper provides new insights into flexible time series models in a structural fashion for RPs in adaptive survey designs. We attempt to interpret time change in survey RPs that correlate significantly with nonresponse biases when nonresponse is subject to time change. Our approach applies to repeated cross-sectional surveys with multiple data collection phases.

Our main objective is to make reliable predictions for RP across relevant population strata (Note that population strata in which response propensities can differ herein can be subpopulations of interest either. They are called strata throughout, even though they do not necessarily coincide with sampling strata.) and to examine the prediction performance so that we can measure how time alters the RP. This general question can be reduced to four concrete aspects:

1) What time-series components contribute most to variation in RPs?
2) What level of RP prediction accuracy can be achieved for the next upcoming time period?
3) How does prediction accuracy vary over population strata?
4) How does prediction accuracy depend on the length of the historic survey time series?

The abundant knowledge of historic survey time series allows us to learn the effects of time-related factors on RP. We consider two levels, time and strata, which make up multiple components involved in a time-series model. The components describe variation over time or strata or over both, and they can be analyzed individually as well as collectively. Several survey methodology studies employ such a multilevel time-series model approach in official statistics; e.g., Boonstra and van den Brakel (2019 and 2022) estimate monthly and quarterly regional unemployment rates using a Bayesian hierarchical model to borrow strength over time, space, and from auxiliary series. Such usage originates from the small area estimation literature (Rao and Molina, 2015).

In this paper, we use the Dutch Health Survey (GEZO) to evaluate our approach regarding the four research questions above. This survey has had a stable design since 2011 and we focus on the time series from 2014 to 2019.

To optimize predictions, we compare a collection of model compositions by different information criteria to obtain a balance between goodness of fit and model complexity. To evaluate the "optimal" model, we will assess its predictive performance and accuracy by its ability to correctly capture the magnitude and variation of RPs. Important to note, we focus on the achievement of reliable inference over time, rather than on minimizing nonresponse error, which is one of the objectives adaptive survey designs pursue.

This paper first introduces several time-related factors of great relevance to variation with a hypothetical illustrative example in Section 2, then goes on to the differential model compositions in the general form of the Bayesian multilevel time-series model in Section 3. Section 4 optimizes the model performance based on an empirical analysis of GEZO. We discuss our findings and end up with the brief overview of future work in Section 5.

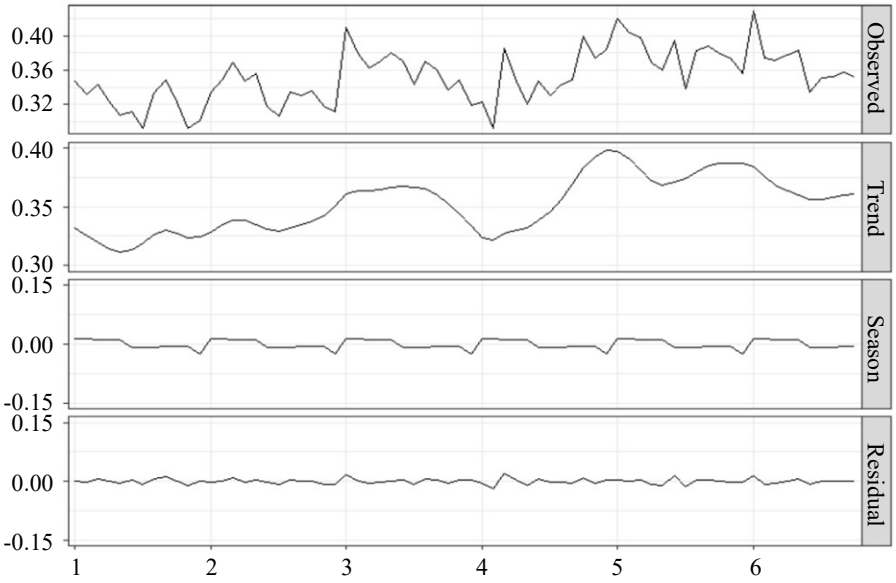
2. Time series components of survey response rates

It is well-known that response propensity (RP) changes gradually in time. Failing to incorporate this temporal dependence in design decisions can lead to ineffective survey designs. In this section, we use an illustrative example for introducing some time-related factors linked with considerable variation in RP.

We focus on population subgroups, or strata, as indexed by $g \in \{1, \ldots, G\}$, since we aim ultimately to let the proposed models inform adaptive design decisions. The strata are formed with the help of auxiliary variables that are linked to the sample and are, thus, available for all sample units. A time-series RP $\rho_{g,t}$ in stratum g and time t is a sequence of random variables. Assuming the availability of historic survey data up to time t, we are interested in measuring variation caused by time-related factors for the most up-to-date RP predictions. To achieve this goal, we first propose potential time-dependent factors. As an illustrative example of a time series divided into the following components: trend, seasonality, and so on, Figure 2.1 compares the overall response rate to the following time-dependent variation:

- *Trend*. The trend describes the long-term movement of the observed time series without the seasonal variation. It shows the general tendency of the population-level response rates over years, which can be linear or nonlinear. Hence, the growth or the fall of the long-term forecasts can be studied by this trend. As seen in Figure 2.1, the long-term direction does not behave like a cyclic fluctuation. Of greater importance for model development is to separate the total trend into a global trend shared by all strata, and local, i.e., stratum-specific trends.
- *Seasonality*. Seasonal variation in the overall responses describe periodic movements that recur regularly and do not influence annual averages. The periodic fluctuations possess a systematic and calendar-related nature that can be predicted and attributed to a fixed season per year. For instance, the response rate would be higher in the early period of the year while relatively lower in the middle year or in December.
- *Residual fluctuation*. The residual variation is the part of the signal obtained after excluding all of the above components. This part is usually modeled as white noise, i.e., as independent normally distributed fluctuations.

Figure 2.1 The observed series of simulated overall response rates over years versus its decomposition.



In addition, there may also be some additional time-dependent components not revealed in Figure 2.1 that nevertheless have a strong impact on the reliability of stratum RP predictions. For this reason, we also consider extra stratum-related time-dependent components:

- *Stratum*. Different subgroups have different response behaviors, such as, young subgroups are more likely to respond to the web survey than old subgroups due to the latter having potentially less access to or unfamiliarity with the internet. This variation in subgroups leads to a

differential stratum-level trend and could potentially also contribute to differential seasonal movement.

- *Sampling variation*. Sampling variation complicates the estimation of RPs, especially for strata with small sample sizes. The sampling variation is taken into account by adopting a binomial likelihood.
- *Unexpected events*. Unexpected events, such as web servers being down temporarily, will appear as outliers and may violate the existing pattern. They correspond to irregular movements during short periods. The resulting variation does not follow a particular model, is unpredictable, and can become influential in predicting future RPs.
- *Intervention*. Design change, such as introducing incentive, is used widely to conduct intervention on purpose, in order to stimulate responses for an improvement in data collection quality, and even to efficiently allocate limited resources for a reduction in survey cost. Intervention has a permanent impact on response propensities. The influence can be predictable, but only can be studied at the expense of wasting the potential value of rich historic data and of a long time period of data collection since then the implement of intervention. The resulting variation is less likely to affect seasonal patterns, while it can bring similar impacts on responses for some strata.

All components together, except for the sampling variation, form the signal, i.e., the latent true but unknown RPs. The mathematical formulations corresponding to each component are introduced in the following section that proposes the structural time series model (See Harvey, 1990 and Durbin and Koopman, 2012 for general background information on those time series components and models).

3. Methods

In this section, we translate the time series components discussed in Section 2 to multilevel time series models and devise the estimation strategy. We adopt a Bayesian approach in order to account for the uncertainty within the historic survey data and to update response propensity (RP) predictions in time. The use of multilevel models is widespread in small area estimation, in which interest focusses on reliable estimation for domains such as geographic areas, time periods, demographic subgroups, or a combination thereof, whose sample sizes are often too small to provide reliable direct estimates, see Rao and Molina (2015) for an overview. Early references to the literature of small area studies using time series multilevel models include Pfeffermann and Burck (1990), Rao and Yu (1994), Datta, Lahiri, Maiti and Lu (1999), You, Rao and Gambino (2003). In most such studies, including Boonstra and van den Brakel (2019), a Gaussian sampling distribution is assumed, possibly after a suitable transformation of the data. A notable difference of our current application to RPs is that we use a binomial sampling distribution, which is a natural distribution to describe the response process given the number of sampled individuals in each

demographic subgroup and time period. Such binomial time series models have been considered by Franco and Bell (2015). Their approach bears a resemblance to our strategy, whereas ours involves more different types of time series components in the model specification, such as seasonality.

We begin the discussion of our method by first introducing the notation used throughout the paper. Next, we describe our model and the strategy used for estimating the RP, and we conclude with outlining the criteria used to evaluate the performance and applicability of prediction models for RPs in the Bayesian framework.

3.1 The multi-level time series model specification

The objective is to predict stratum-level RPs at a certain point in time. The population or a sample is partitioned into strata based on several auxiliary variables, i.e., stratified, equivalent to a cross-classification of selected variables. Here, we assume the stratification is specified prior to fitting the models. The categories of each variable may be merged to ensure sufficient sample sizes.

Let sample size in stratum g at wave t be $n_{g,t}$ and the number of respondents be $r_{g,t}$, where $g \in \{1,\ldots,G\}$ and $t \in \{1,\ldots,T\}$. The number of strata G is typically in the order of 10 to 20, and T refers to survey waves, each of which is a new replication of the survey starting from a fresh sample. We assume that all sampled units are independent in their response behavior within and between strata. For stratum g and time t, response $r_{g,t}$ follows a binomial distribution conditionally on RP $\rho_{g,t}$ and sample size $n_{g,t}$, i.e., $r_{g,t} \mid n_{g,t}, \rho_{g,t} \sim \text{Binom}\left(n_{g,t}, \rho_{g,t}\right)$. Because RP is constrained to fall between 0 and 1, we transform the 0-1 scale to the real line $\mathbb{R}$ by utilizing a logit link function, where other link functions are usable as well. The function provides a nonlinear transformation and produces a latent variable $\theta_{g,t}$, which follows the log-odds function,

$$\theta_{g,t} = \text{logit}\left(\rho_{g,t}\right) = \ln\left(\frac{\rho_{g,t}}{1-\rho_{g,t}}\right).$$

We can reverse the transformation to compute $\rho_{g,t}$,

$$\rho_{g,t} = \frac{\exp\left(\theta_{g,t}\right)}{1+\exp\left(\theta_{g,t}\right)}.$$

For any stratum g and any time t, the linear predictor $\theta_{g,t}$ can take the most general form that can be linear, additive, multilevel and comprised of several time series components. As outlined in Section 2.2, there are demographic variables defining the strata, an overall trend, seasonal variation, stratum-specific trends, and a residual variation. Therefore, the multilevel model becomes:

$$\theta_{g,t} = \boldsymbol{\beta}' \mathbf{x}_g + \gamma t + \boldsymbol{\delta}' \mathbf{s}_t + v_g + u_t + z_{g,t} + w_{g,t}, \tag{3.1}$$

where the p-vector of regression effects $\boldsymbol{\beta}$ is associated with time-independent covariates $\mathbf{x}_g$. In the application we focus on later in this paper, all covariates are binary as we only consider categorical

variables. However, in more general usage, the entries could be ordinal or numerical variables, such as contact attempts, and even they could vary over time.

Scalar γ is the slope parameter for the overall linear time trend. Vector $\boldsymbol{\delta}$ contains seasonal effects with vector $\mathbf{s}_t$ selecting the season corresponding to month t. The seasonal effects are either common to all strata, or they can be stratum-specific. In this paper, we define seasons as a division of months in a calendar year, i.e., sets {1, 2}, {3, 4, 5}, {6, 7, 8}, {9, 10, 11} and {12} as Winter, Spring, Summer, Fall and Christmas.

The first three terms are modelled as fixed effects while the last four terms are modelled as random effects in (3.1). The first of these random terms is the random intercepts for strata assumed to be normally distributed with mean 0 and variance σ_v^2 as

$$v_g \sim N\left(0, \sigma_v^2\right) \tag{3.2}$$

identically and independently for $g = 1, \ldots, G$. Secondly, a global time trend is defined by a random effect vector $\mathbf{u} = \left(u_1, \ldots, u_T\right)$ distributed as

$$\mathbf{u} \sim N\left(0, \sigma_u^2 \mathbf{V}_u\right). \tag{3.3}$$

Covariance matrix $\mathbf{V}_u$ describes the covariance structures between any u_i and u_j. One can assume either a first-order random walk (RW1, known as a local level trend) or a second-order random walk (RW2, the so-called smooth trend). The time-dependence structures are more conveniently expressed by the precision matrix, $\mathbf{Q}_u = \mathbf{V}_u^{-1}$. The precision matrix is preferred over the covariance matrix, since it is sparse and allows for efficient computation for hierarchical posterior inference in a Bayesian analysis, see e.g., Rue and Held (2005). The matrix $\mathbf{Q}_u$ for RW1 and RW2 is a tridiagonal matrix and a pentadiagonal matrix (Assumed a band matrix is $Q = \left(q_{i,j}\right)$, 1 has one non-zero bands along the main diagonal such that $q_{i,j} = 0$ if $\left|i - j\right| > 1$, while 2 has two non-zero bands such that $q_{i,j} = 0$ if $\left|i - j\right| > 2$.) respectively (see Appendix C for their definitions). Note that the precision matrices $\mathbf{Q}_u$ are singular, leading to an improper prior. This is not a problem, as constraints can be imposed on these random effects to ensure that all model coefficients remain identifiable. Under RW1 and RW2 $\mathbf{u}$, the constraint is $\sum_t u_t = 0$. Under RW2 $\mathbf{u}$, the constraint $\sum_t t u_t = 0$ is additionally imposed, so that the corresponding overall level and linear slope are captured by the model's intercept and fixed effect γ.

We also consider distributions other than the normal distribution in (3.3). In particular, we consider Laplace, Student-t and horseshoe priors as alternatives. Such priors can be framed as scale-mixtures of the normal distribution, see West (1987), Carvalho, Polson and Scott (2010) and Polson and Scott (2010).

The third random effect term $\mathbf{z}_g = \left(z_{g,1}, \ldots, z_{g,T}\right)$ denotes stratum-specific trends distributed as

$$\mathbf{z}_g \sim N\left(0, \sigma_z^2 \mathbf{V}_z\right) \tag{3.4}$$

for $g = 1, \ldots, G$. Covariance matrix $\mathbf{V}_z$ describes a RW1 over the months. The corresponding precision matrix is the same as described above, and a sum-to-zero constraint is imposed on each trend vector $\mathbf{z}_g$,

as the stratum-specific levels are already captured by the random intercepts v_g. Important to note is that the trends $\mathbf{z}_g$ share a common covariance parameter σ_z^2. One could consider a separate variance parameter per stratum but we found it resulted in overfitting.

The last term $w_{g,t}$ in (3.1) represents white noise and allows for remaining unstructured variation in RPs over time and strata, i.e., at the most detailed level. For any stratum g and time t, these components are independently and identically distributed as

$$w_{g,t} \sim N\left(0, \sigma_w^2\right), \tag{3.5}$$

using a single variance parameter σ_w^2.

(3.1) describes the most general model considered combining all underlying components. Section 4 investigates this encompassing model as well as models built from various subsets of the components described in (3.2)-(3.5).

3.2 The estimation strategy

In this section, we adopt a hierarchical Bayesian approach to estimate model coefficients and predict RPs. Since the posterior distributions are unavailable in closed form a Gibbs sampler is used as implemented in the *mcmcsae* R package (Boonstra, 2021). We begin this subsection by specifying the priors assigned to the model parameters.

For the fixed effects $\boldsymbol{\beta}$ we assume a weakly informative prior,

$$\boldsymbol{\beta} \sim N\left(0, 100\mathbf{I}_\beta\right),$$

with identity matrix $\mathbf{I}_\beta$. Standard errors for $\boldsymbol{\beta}$ are taken as 10, which is sufficiently large concerning the scale of RPs relative to the covariate scales. Similarly, the linear time trend γ, and seasonal effects δ, are assigned weakly informative priors also, with the same standard error.

For the random-effect components, the variance parameters in (3.2)-(3.5) are assigned inverse χ^2 priors, conditionally on auxiliary parameters ξ, with 1 degree of freedom and a scale parameter ξ^2. For example, $\sigma_v^2 \mid \xi_v \sim \text{Inv}-\chi^2\left(1, \xi_v^2\right)$. The hyperparameters ξ are assigned $N\left(0,1\right)$ priors. Combining the normal ξ with the conditional inverse chi-squared variances results in marginal half-Cauchy priors for each standard deviation parameter σ_v, σ_u, σ_z and σ_w. As Gelman (2006) and Polson and Scott (2010) suggest, the half-Cauchy priors for standard deviations, or the more general half-t family of priors, generally perform better than the commonly used inverse gamma priors for variance parameters, which can be too informative.

The (hyper) parameter vector, denoted by ψ,

$$\psi = \left(\beta, \gamma, \delta, v, u, z, w, \sigma_v^2, \sigma_u^2, \sigma_z^2, \sigma_w^2, \xi_v, \xi_u, \xi_z, \xi_w\right)$$

includes all parameters in (3.1), the variance parameters associated with random effect terms as well as the introduced auxiliary parameters. The likelihood function can be written as

$$p\left(r \mid n,\psi\right) \propto \prod_{g,t} \rho_{g,t}^{r_{g,t}} \left(1-\rho_{g,t}\right)^{n_{g,t}-r_{g,t}}, \tag{3.6}$$

where $\rho = \text{logit}^{-1}\left(\theta\left(\psi\right)\right)$ and θ is the linear predictor function of vector ψ as expressed in (3.1). Based on Bayes' theorem, the posterior of vector ψ is proportional to the product of the prior and the likelihood, i.e., $p\left(\psi \mid n,r\right) \propto p\left(\psi\right) p\left(r|n,\psi\right)$. The Gibbs sampler then generates samples from the joint posterior, and the posterior estimates of RP $\rho_{g,t}$ comes as a by-product of these samples – per sample, RPs can be computed using reversed logit transformation. Repeated samples are drawn from the full conditional posterior of each (hyper) parameter. See Appendix D for more information on the full conditional posterior distributions.

Three Markov Chains are produced by the Gibbs Sampler using the *mcmcsae* package (Boonstra, 2021) programmed in R (R Core Team, 2020). Each chain consists of 1,500 draws that are sequentially generated; however only the last 1,000 draws are kept for the estimation algorithm. Convergence of the MCMC sample is assessed using trace and autocorrelation plots. The Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992) is evaluated to diagnose the mixing of the chains. In particular, the autocorrelation of sequential draws is reduced, as the blocked Gibbs sampler updates all fixed and random coefficients simultaneously. In addition, the approach includes a novel data augmentation approach for sampling from binomial logistic models (Polson, Scott and Windle, 2013) which is known to lead to an efficient and relatively fast converging sampler.

3.3 Performance criteria

To guide the model building using the model components and priors described in Sections 3.1 and 3.2, and to assess the models' adequacy, we employ three criteria for model assessment and one for model predictive performance.

The common and popular selection criteria in Bayesian hierarchical settings are the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010, 2013) and the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin and van der Linde, 2002). They are chosen in the pursuit of a reasonable balance between model fit, model complexity and efficient computation (see Appendix C for their definitions). Models with lower DIC/WAIC are preferred. Next, we use posterior predictive p-values to check model adequacy, i.e., simulating draws from the posterior predictive distribution and comparing them to the observed data, see e.g., Gelman, Meng and Stern (1996). This evaluates whether the multilevel model can reproduce data similar to the observations. The p-values are defined as

$$p \;=\; \Pr\left(S\left(r^{\text{rep}}\right) \;\geq\; S(r) \mid r\right), \tag{3.7}$$

where S is a test statistic and r^{rep} denotes a replicated dataset generated from the posterior predictive distribution based on the fitted model, $p\left(r^{\text{rep}} \mid r\right) = \int p\left(r^{\text{rep}} \mid \rho, n\right) p\left(\rho \mid r, n\right) d\rho$. The p-values are estimated from the MCMC output, and values close to 0 or 1 are indicative of a poor fit regarding statistic S. Here we consider two test statistics:

1. $S\left(r\right) = \bar{r}$, the unweighted mean of the replicate data-vector.

2. $S(r) = \frac{1}{\mathrm{GT}-1}\sum_{g,t}\left(r_{g,t}-\bar{r}\right)^2$, the unweighted variance of the replicate data-vector and $\bar{r}$ is the mean of $r_{g,t}$.

To assess the models' prediction performance, we define a predictive measure: the root mean squared error (RMSE) in stratum g at month t as the square root of the sum of two terms: 1) the quadratic differences between the posterior means of $\rho_{g,t}$ and the observed response rate (RR), and 2) the posterior variances of $\rho_{g,t}$. The general form of the expression in stratum g at month t is

$$\mathrm{RMSE}(g,t) = \sqrt{\left(E_{\pi_t}\left(\rho_{g,t}\right)-\hat{\rho}_{g,t}\right)^2 + \mathrm{var}_{\pi_t}\left(\rho_{g,t}\right)}, \tag{3.8}$$

where $\hat{\rho}_{g,t}$ is the realized value of RP and estimated by the observed RR, and π_t is the posterior predictive distribution of the RPs, when employing historic data up to and including $t-1$ and new data in t for RP prediction. For ease of notation, the two terms under the square root in (3.8) are referred to as the bias term $\left(B(g,t)\right)$ and the standard deviation $\left(\mathrm{SD}(g,t)\right)$. The bias term in (3.8) will, in general, be larger than zero due to random variation in the sampling of strata and in the response of sample units. For this reason, we benchmark the RMSE against an empirical lower bound denoted by $\mathrm{RMSE}_{\min}$. The lower bound estimate is called the Monte Carlo approximation to the posterior mean of the binomial standard deviations, which is a function of the k^{th} iteration from the posterior draws of $\rho_{g,t}$,

$$\mathrm{RMSE}_{\min}(g,t) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}, \tag{3.9}$$

where k runs over MCMC draws and $n_{g,t}$ is the size of stratum g sample in month t. (3.8) and (3.9) give one-month assessments per stratum g. They need to be aggregated across strata and in time to get meaningful overall assessments.

In any particular month, a stratum with a larger sample size should impose more weight on the reliable predictions. The weights $d_{g,t}$ are defined as the sample proportion, i.e.,

$$d_{g,t} = \frac{n_{g,t}}{\sum_g n_{g,t}} \quad \text{subject to} \quad \sum_g d_{g,t} = 1.$$

Thus, the sub-terms

$$B(t) = \sqrt{\sum_g d_{g,t}\left(E_{\pi_t}\left(\rho_{g,t}\right)-\hat{\rho}_{g,t}\right)^2}$$

and

$$\mathrm{SD}(t) = \sqrt{\sum_g d_{g,t}\,\mathrm{var}_{\pi_t}\left(\rho_{g,t}\right)}$$

in month t should be the square root of the sum of the weighted individual measures $B(g,t)$ and $\mathrm{SD}(g,t)$ by $d_{g,t}$ over strata, while the lower bound over strata in time t becomes

$$\text{RMSE}_{\min}(t) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\sum_g d_{g,t}\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

Also, the stratum-specific sub-terms

$$B(g,T) = \frac{1}{T}\sum_t\sqrt{\left(E_{\pi_t}\left(\rho_{g,t}\right)-\hat{\rho}_{g,t}\right)^2}$$

and

$$\text{SD}(g,T) = \frac{1}{T}\sum_t\sqrt{\text{var}_{\pi_t}\left(\rho_{g,t}\right)}$$

in a time period $T=\left\{t \mid t_1,\ldots,t_T\right\}$ are the average of the individual measures $B(g,t)$ and $\text{SD}(g,t)$ over months where t indicates a month, while stratum-specific lower bound over time period T becomes the average of the individual measures $\text{RMSE}_{\min}(g,t)$, i.e.,

$$\text{RMSE}_{\min}(g,T) = \frac{1}{T}\sum_t\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

Furthermore, the overall sub-terms or term in a time period T becomes the average of the weighted sub-terms $B(t)$, $\text{SD}(t)$ and $\text{RMSE}_{\min}(t)$ over months, i.e.,

$$B(T) = \frac{1}{T}\sum_t\sqrt{\sum_g d_{g,t}\left(E_{\pi_t}\left(\rho_{g,t}\right)-\hat{\rho}_{g,t}\right)^2},$$

$$\text{SD}(T) = \frac{1}{T}\sqrt{\sum_g d_{g,t}\text{var}_{\pi_t}\left(\rho_{g,t}\right)}$$

and

$$\text{RMSE}_{\min}(T) = \frac{1}{T}\sum_t\frac{1}{K}\sum_{k=1}^{K}\sqrt{\sum_g d_{g,t}\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

4. Analysis of results

In this section, we introduce the Dutch Health Survey (GEZO) as a case study to demonstrate how the multi-level time series models can be built and how we update RPs in time. We address the four research questions in corresponding subsections.

4.1 The Dutch Health Survey

The GEZO has been conducted annually since 1981 by Statistics Netherlands as a repeated cross-sectional survey in which a sample of households was interviewed with the aim of providing an overview of developments in the health, medical consumption, lifestyle and preventive behavior of the Dutch population. The sampling frame is formed by first drawing a sample from municipalities and then from all

people who live in the selected municipalities. As of 2010, the survey changed to a mixed-mode survey involving an initial web and the follow-up telephone (or face-to-face) interview. Non-respondents to web were contacted via telephone if their telephone numbers were known at the register, and otherwise a face-to-face interview was arranged. Over these years, the sample size was increased to 15,000 and the overall response rate was increased by 25%. From 2014 onwards, the mix of the follow-ups was changed to a face-to-face interview. In 2018, however, a part of the web non-responses was approached via a face-to-face interview in a more effective way. The propensity to respond to personal interviews in a time series strongly depends on web response outcomes, so in a sense modeling follow-up propensity is conditional on web RP model. This issue needs consideration more than interpreting time change in web RP and is beyond the main aim of this paper. For the sake of simplicity, our concern is to model web response propensity in this paper as a fundamental start, hence modeling follow-up RP in a time series is more suited to future research. As an important note here, only the web GEZO data from 2014-01 up to and including 2019-10 are analyzed in this study. We employ three auxiliary variables that stem from administrative frame or registers. The prescribed auxiliary variables are age, gender and ethnicity, which divides the population or its sample into 20 disjoint strata (see Appendix A for more information).

The GEZO conducted over many years is a relatively consistent survey design. This feature makes exploring time-dependence in RPs valid because of the abundant time series. Our interest focuses on monthly response data, i.e., sample size and the number of respondents of each stratum. Predictions are made monthly but also can be aggregated quarterly or annually.

4.2 What time series components contribute most to variation in RPs?

We address this question in two steps: First, we go through model combinations and then we compare their performance. The comparison of multiple models is made from two views: (1) “what combination fits best to response data?” and (2) “what combination makes the most reliable predictions?” We use information criteria and posterior predictive p-values to measure the performance of each model, and thus search for an “optimal” model. The model is preferred when it has lower information criteria and predictive p-values closer to 0.5.

Since trying all combinations of components in (3.1) places a heavy burden on computation, it is important to apply an efficient search for the “optimal” model. To do so, we fit the models to response data using the following strategy:

1. Start with the baseline model (auxiliary variables only).
2. Add fixed effects sequentially, linear time trend and seasonal trends, to the baseline.
3. Investigate whether the model in 2 continues to improve with global time effects or global seasonality.
4. Investigate whether stratum-specific time trends or seasonal effects further improve the model.
5. Determine whether a white noise term for unexplained variation is needed.
6. Explore robustness for outliers through different prior specifications or time-dependent structure of global random intercepts over time.
7. Evaluate the model using a number of diagnostics.

Table 4.1 shows the selection results. The fixed-effect models (M1 to M3) behave worse than the mixed effect models relative to the trade-off between fitness and complexity, as the latter ones yield lower ICs (DIC, WAIC). Comparing M2/M3 to M1 implies that time slope λ or seasonality δ causes a decrease in ICs. However, the model further improves by introducing global trend u_t, as a significant decrease in ICs in M4 relative to M3 is observed. As M5 and M6 show, the improvement continues with the addition of random intercepts for strata v_g and stratum-specific time trends $z_{g,t}$. Although white noise $w_{g,t}$ seems to add only very little in M7 overall, the posterior predictive p-values for variances imply that it is worth to include white noise. Further, we found that using a local level trend (RW1) or smooth trend (RW2) as the global trend u_t makes hardly any difference concerning ICs for models M6 to M11.

Finally, the 4th column of Table 4.1 shows the prior distribution used for the global trend coefficients u_t. The non-normal priors that have been attempted do not further improve ICs, but because of heavier tails they help to combat an outlier in the data, an exceptional issue in February 2017.

Table 4.1
Summary of the multilevel time-series models considered

Model	Fixed	Random	Prior	DIC	pDIC	WAIC	pWAIC	PPP	
								Mean	Variance
M1	β	-	-	7,511	7	7,518	13	0.501	0.006
M2	β, λ	-	-	7,415	8	7,421	15	0.503	0.051
M3	β, λ, δ	-	-	7,368	12	7,378	22	0.498	0.092
M4	β, δ	u_t	Normal	7,255	43	7,280	68	0.484	0.168
M5	β, δ	u_t, v_g	Normal	6,916	56	6,925	65	0.491	0.172
M6	β, δ	$u_t, v_g, z_{g,t}$	Normal	6,790	98	6,781	90	0.494	0.356
M7	β, δ	$u_t, v_g, z_{g,t}, w_{g,t}$	Normal	6,790	131	6,769	110	0.517	0.425
M8	β, δ	$u_t, v_g, z_{g,t}, w_{g,t}$	Laplace	6,790	133	6,768	111	0.503	0.397
M9	β, δ	$u_t, v_g, z_{g,t}, w_{g,t}$	T-distributed	6,790	130	6,769	110	0.492	0.391
M10	β, δ	$u_t, v_g, z_{g,t}, w_{g,t}$	Horseshoe	6,790	131	6,769	110	0.518	0.411
M11	β, λ, δ	$u_t, v_g, z_{g,t}, w_{g,t}$	Laplace	6,806	150	6,779	122	0.519	0.413

Notes: "-" indicates no random effects or prior.
Deviance Information Criterion (DIC); Widely Applicable Information Criterion (WAIC); Posterior predictive p-values (PPP).

T-distributed and horseshoe priors are likely to accommodate and be robust against the outlier better than normal and Laplace priors, as shown by comparing their posterior means of global trend u_t in Appendix B. Besides, the local level trend of M8 seems to outweigh slightly the smooth trends of M11. P-values of the mean of M8 bring the value closer to 0.5 than M9 and M11.

To determine which model is flexible to the outlier and which one generates reliable estimation throughout the series, we look further into the discrepancy between observations and model-based estimates, specially M8, M9 and M11. The comparison demonstrates that the three models have limited ability to combat the outlier. The lower quantiles attempt to reach to the outlier but cannot cover it. In addition, the Laplace prior has slightly smaller uncertainty about the posterior estimates than the T-distributed prior, but it has similar size in uncertainty to the smooth trend model (see Appendix B).

4.3 What level of response propensity prediction accuracy can be achieved for the next upcoming new time period?

To answer the second research question, we estimate the level of and variation in overall response predictions for the forthcoming data collection wave. The estimated level is the deviation of the expected posterior propensity prediction from the realized response rate, while the estimated variation refers to the prediction accuracy in the overall RP. Also, we measure the balance between the level and variation and compare it with the benchmark in (3.9). The assessment allows us to validate if gains can be achieved from our method. Actions can be taken to adapt/maintain data collection in the following wave once the gain is known upon historic series.

We stress that the analysis is made based on the "optimal" model, M8. For all strata in a new sample per month, the months since January 2014 up to but not including the present month are viewed as the historic time series, which are used for training M8. Then we use the fresh sample from the present month for the estimated predictive criteria. The historic time series is accumulated and predictive criteria are updated with the new wave. The rolling assessment ends with 2019-09 as one month must be left for the prediction exercise because 2019-10 is the last month of data available. To lend robustness to the impact of historic size on predictive performance, we let historic time series start with 60 months (from 2014-01 to 2018-12) as the default initial trial.

Table 4.2 shows that the posterior uncertainty in the overall RP predictions decrease steadily but slowly and converge to around 0.027. Because of the sampling variation that is inherent to the bias term, the pattern for bias is erratic and shows at best a modest decrease. Relative to the realized response rates, the greatest deviation of posterior means is around 0.07 in January and June, and the smallest deviation around 0.04 in March, May, August and October. The RMSE results vary along with the bias term across months, as the estimated SDs are much smaller than the estimated biases. The RMSE has a maximum value of 0.084 in January, and is likely caused by the outlier months in early 2017. Although the model reacts to this disruption, it has a negative impact on the performance of the resulting predictions in this month. Aside from January, in some months the estimated RMSE is close to the benchmark $RMSE_{min}$. It can be concluded that the estimated accuracy lies relatively close to the maximal possible accuracy.

Table 4.2
One month ahead prediction of three measures of RPs over strata: bias, standard deviation (SD), and the root mean square error (RMSE) compared to the benchmark ($RMSE_{min}$)

	2019									
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**
bias	0.078	0.064	0.045	0.062	0.046	0.077	0.063	0.049	0.058	0.048
SD	0.030	0.031	0.031	0.029	0.028	0.029	0.028	0.027	0.028	0.027
RMSE	0.084	0.071	0.055	0.069	0.054	0.082	0.068	0.056	0.065	0.055
$RMSE_{min}$	0.055	0.056	0.055	0.055	0.055	0.055	0.048	0.048	0.049	0.049

Notes : The column indicates the present month for evaluating prediction performance.
Response propensities (RPs).

4.4 How does prediction accuracy vary over population strata?

This research question concerns the different strata and how well the model performs in predicting RP per stratum. For this purpose, we consider the stratum-level RMSE as well as its two components, i.e.,

bias and standard deviation. The evaluation measures are taken as the average over the ten months ahead predictions. Month 2019-10 is the last month available. Looking ahead by almost a year allows data collection staff to plan adaptive designs well ahead of time.

Similar to Section 4.3, we limit the analysis to the assumptions. The preferred model is selected from Section 4.1, and the historic time series is fixed to be 60 months (2014 to 2018). For each stratum, the model is fully trained by the fixed historic data and makes inference on predictions in the remaining months in 2019.

Table 4.3 shows prediction criteria for each stratum together with the benchmark. The estimated bias terms vary widely between strata. The greatest departure of posterior expectation from the realized response rates occurs in stratum 8, 10, 12, and 18, all with biases larger than 0.1. Compared to biases, there is a relatively smooth change in the estimated SDs around 0.03, where stratum 4 has smallest uncertainty about the posterior estimates with 0.018.

Some strata with greater biases may have less accuracy in posterior estimates of RPs than strata with less biased propensity. Similarly, the more biased the prediction is, the greater the RMSE is estimated to be. This is because the estimated biases are much greater than the estimated SDs. It is not surprising that stratum 10 has the greatest value in RMSE, where the prediction is the most biased and has the least precision. The RMSE results can catch up with, and even can be comparable/superior to the benchmark. For example, when the model generates prediction for stratum 20, more significant gains can be achieved than other strata.

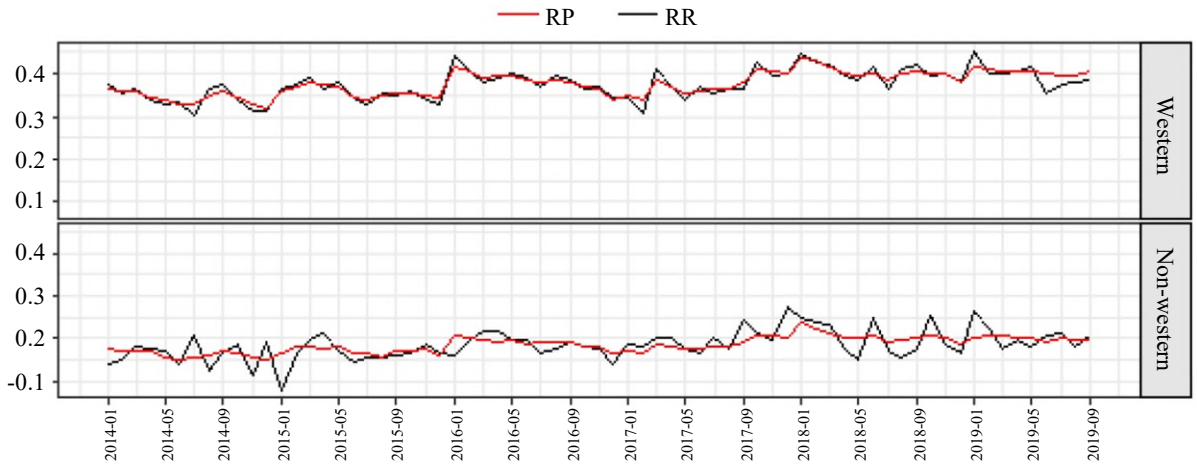
Table 4.3
The average of ten months ahead prediction of three measures in each stratum: bias, standard deviation (SD), and the root mean square error (RMSE) that is compared to the benchmark ($RMSE_{min}$)

	bias	SD	RMSE	$RMSE_{min}$
1	0.045	0.030	0.060	0.046
2	0.066	0.030	0.077	0.094
3	0.028	0.025	0.039	0.039
4	0.049	0.018	0.053	0.061
5	0.035	0.026	0.045	0.035
6	0.047	0.021	0.053	0.064
7	0.062	0.032	0.073	0.054
8	0.105	0.030	0.111	0.154
9	0.047	0.031	0.060	0.045
10	0.165	0.035	0.173	0.160
11	0.044	0.031	0.057	0.048
12	0.134	0.030	0.138	0.092
13	0.030	0.027	0.044	0.042
14	0.081	0.022	0.086	0.074
15	0.044	0.028	0.056	0.038
16	0.067	0.022	0.072	0.072
17	0.030	0.031	0.048	0.053
18	0.114	0.029	0.120	0.146
19	0.031	0.029	0.046	0.041
20	0.095	0.030	0.105	0.172

The predictive performance shows a significant difference between strata when there is only one different characteristic. For example, stratum 20 RMSE is 0.06 lower than stratum 10 RMSE. This seems

to imply that female groups may have smaller bias or variance than male groups when non-western people above the age of 64 are considered. Given the age and ethnicity of groups and compared with non-western groups (even rows), RMSE results are much lower in western groups (odd rows). To validate this supposition, some strata are combined into subgroups with less detailed characteristics. As Figure 4.1 shows, the model yields better predictions for western group than non-western groups, as expected posterior estimates reach mostly the observed response per month. The comparative performance for age/gender groups are in Appendix B.

Figure 4.1 Monthly posterior means of RP aggregated over Ethnic groups versus observed response rates (RR) of Ethnic groups.



Note: Month 2014-01 to 2018-12 for the estimated model and Month 2019-01 to 2019-10 for RP predictions.

4.5 How does prediction accuracy depend on the length of the historic survey time series?

The primary concern of this question is to find out how robust the prediction performance is to the amount of historic time series that is used for model training and predicting. For this purpose, we continue with the average of three-month ahead predictions of RMSE and its two terms, bias and SD, at the overall level at any given time point. We call this length-based average the quarterly average. To explore the impact of historic data size, we perform 3-split time series cross validation on dataset, i.e., successively add three months of new data to the training dataset used for model-based predictions. This analysis is iterated on a rolling basis and the step-by-step strategy is laid down as follows,

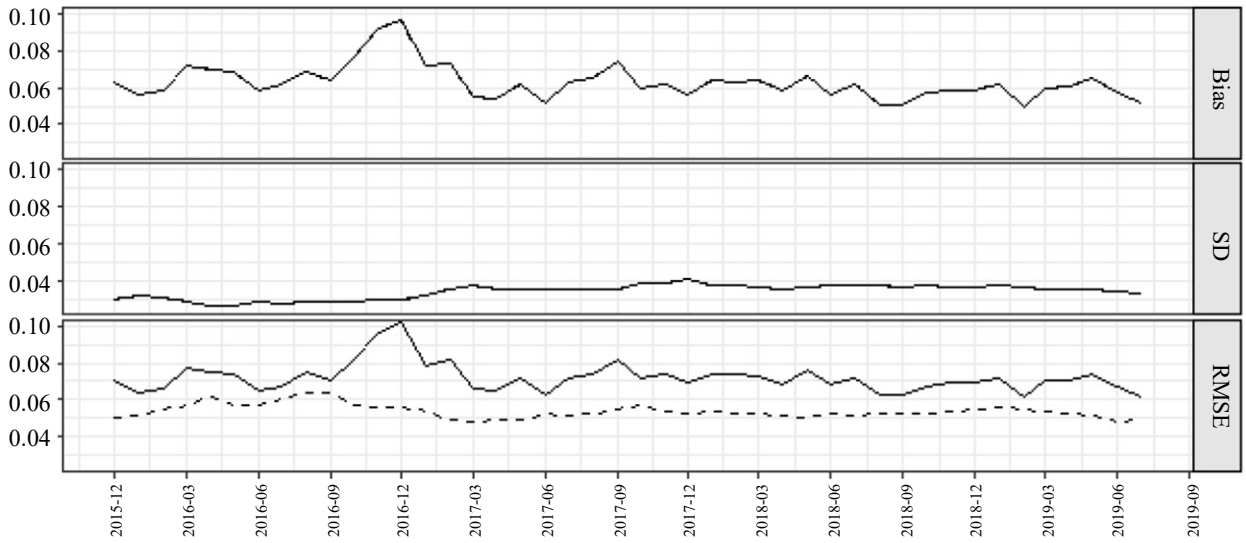
1. Select the model components based on the whole time series.
2. Select the baseline set of historic time periods of length t. Partition the window into the training set D_o of the first $t-3$ time periods and the test set D_t of the last three periods.
3. Data D_o trains the selected model, by simulating from the posterior distribution of all model parameters, given D_o.

4. Based on the simulated model in 3, posterior predictive means and variances are computed for the RPs for each stratum and each time point in the test set D_t.
5. Based on individual RPs predictions in 4, compute overall Bias, and SD, by using the sample proportion $d_{g,t}$ in stratum g at time t as weights, then equation (3.8) computes the quarterly average of RMSE.
6. Expand the window to $t+1$, moving one time period forward. Repeat 3-5 for updating the predictions of RMSE and its two terms.
7. Repeat 6 until length t is the last available time period.

We stress that it is necessary to use at least two years as the initial training length when seasonal components are included. In our case, time periods are months and the time series runs until 2019-07.

In Figure 4.2 RMSE, Bias and SD estimates are length-dependent and computed over strata for three months ahead. The baseline window of training data for fitting the optimal model are 2014-01 to 2015-12. Along the time series window, bias results are approximately greater than two times the SDs. Consequently, RMSE results are dominated by biases and show the same volatility. At the end, their estimates are approximately 0.06, whereas SD estimates undergo a slight increase. The latter is somewhat surprising, as one would normally expect that using a longer time series to estimate the model would decrease the posterior standard errors of prediction. It turns out, however, that two events had a large impact on the prediction performance. First, early in 2017, data collection experienced an interruption caused by technical issues with the web server. This incident had a large immediate impact on RPs and consequently also on model prediction performances. Second, in 2018, conditional incentives were introduced and the survey questionnaire was made smartphone proof. This design intervention had a more gradual and longer lasting impact.

Figure 4.2 One-step forward moving average of quarterly Bias (upper panel), quarterly SD (middle panel) and quarterly RMSE (bottom panel). RMSE (solid) against benchmark RMSE (dashed) when the length of training data set moves on x axis.



The Bias and RMSE results undergo a big increase from 2016-10 to 2017-02. When the training window arrives at time point 2016-10, the test window starts to include 2017-01 data where RPs dropped. Their climbing curves continues and reach maxima around 0.1, when the training window moves to 2016-12 and the test window first moves to the "stable" month 2017-03 where the Bias and RMSE curves drop back around 0.05. During these months a slight, gradual increase in SD can be observed as well.

The inclusion of these outlier months affects prediction accuracy during 2017. Between month 2017-03 and 2017-12 the Bias and RMSE curves are more volatile, and decline only after 2018-01. The SDs have a rising tendency from 2017-03 to 2017-12 and hardly decrease.

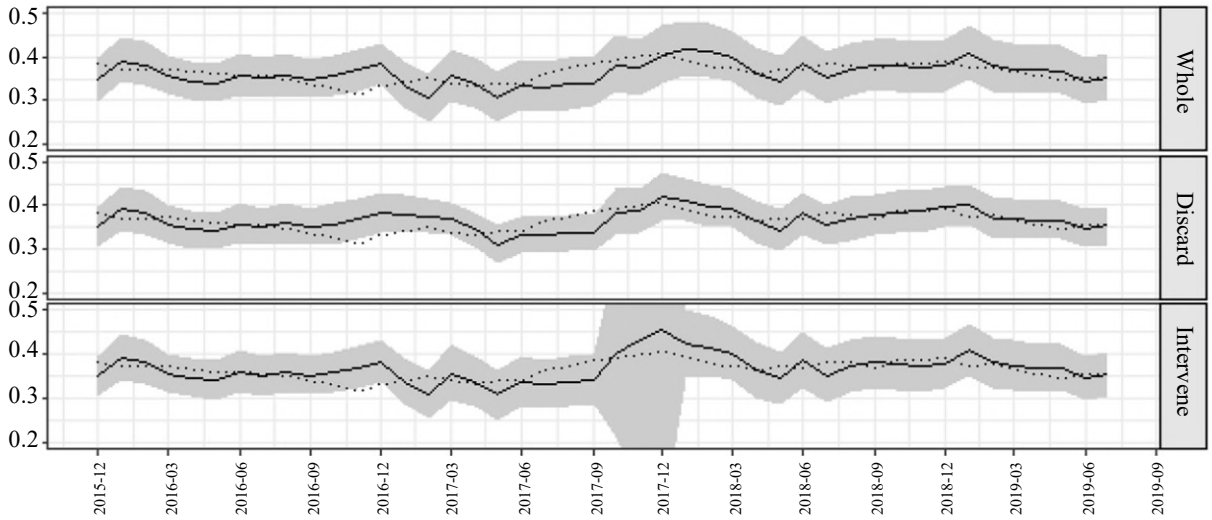
In 2018, the impact of the design intervention on RPs was much more modest than in 2017, but since it is structural it does affect prediction accuracy in 2018 and 2019. The two events, one technical incident and one design change, are realistic in survey practice and when ignored can have a devastating impact. We set an example of how one might deal with them. The extra efforts are:

- *Discard method*. For the original data, clear the response numbers $r_{g,t}$ in 2017-01 and -02 and treat them as missing data. Impute these missing $r_{g,t}$ by the posterior means of simulated responses from the posterior predictive distribution. Note that in Section 4.2 we argued that using specific non-normal priors for time series components can also limit the effect of outliers. If an outlier is quite extreme and known to occur at a specific time, it may however be better to discard it.
- *Intervention method*. Include an intervention term in the model orcapturing the possible structural change. Add intervention binary variables to the original data series and let them be 0-1, where in our case they would take the value 1 and become active from 2018-01. The potential intervention-related effects could be either a single fixed effect, stratum-specific random effect or both.

Results from applying these two methods separately are shown in Figure 4.3.

The two methods have a clear effect on predictions. In the period of 2016-12 to 2017-05 where the training time series window stops, the posterior means of the discard method show a declining trend, relative to the original model's posterior means ("Whole" in Figure 4.3). However, from 2017-10 to 2018-03 the difference in the mode-based means and observations becomes small for the discard method. Also, the discard method decreases uncertainty about posterior means as the credible band becomes narrower since from 2016-12. The intervention-related impact on overall RP cannot be estimated well using just a few new months of data.

Figure 4.3 One-step forward moving average of quarterly posterior estimates of overall RP under three scenarios: (1) the original time series (top panel); (2) the new time series by discarding early 2017 data (middle panel), and (3) the new time series by adding intervention-related effects (bottom panel).



Note: Compared to the moving average of observed response rates quarterly (dashed), the model-based estimates are summarized as the posterior means (solid) with 95% credible region (band in grey). X axis labels the length of training data from 2014 to that time point.

While, it was not our intention to provide a detailed account of modelling options for incidental and structural changes, the time series model we propose can be modified in a relatively straightforward and flexible way. Replication with long survey time series is warranted to get a sense of what options are superior.

5. Discussion

Accurate and reliable prediction of response propensities (RPs) is the key to improving and optimizing adaptive survey designs. Such inference can be complicated due to seasonal variation and time-related trends that may be specific to population strata. In this paper, we introduce a Bayesian multilevel time series model for stratum-level RP predictions. The model is flexible enough to include seasonal variation, various forms of trends, design changes and stratum-dependency, so that it can facilitate preparation for adaptive survey design in a changing survey climate. They are elicited from historic survey data and updated with new survey data.

In this paper, we apply the method to a general population repeated survey, the Dutch Health Survey at Statistics Netherlands, to provide empirical support in a realistic case. The major focus is on improving stratum-level RP predictions that are subject to time-related factors. Based on various model combinations made of these factors, one of our concrete objectives is to search for the highest-performance model that makes a trade-off between model fitness and computational ease. The optimal model is selected based on criteria that assess both performance (high IC, p-value ≈ 0.5) and predictive ability (low RMSE_{RP}).

These measures provide valuable insight into the relative gain achieved by adding new factors. This flexible approach allows other survey researchers to consider different time-related factors and ultimately choose the preferred model in their settings.

The remaining objectives of this paper center on evaluating the prediction performance of stratum-level/overall RPs based on the preferred model. We use predictive metrics, specifically the root-mean-square error (RMSE), to assess uncertainty in predictions. This allows us to directly compare: (1) overall predicted response in first forthcoming data, (2) annual-averages of predicted response for each stratum, and (3) quarter-averages of overall predicted response. We evaluate the role of length of the historic survey time series in both the ultimately preferred model and a model that is re-optimized when data come in. Doing so we can find out when is a suitable time to start implementing an adaptive survey design. Note that when the survey design is made adaptive, it becomes less evident how to learn about the time change in model parameters. Also, the time series model itself may need to be updated depending on the type of survey adaptation.

While our attempt is a first step to adaptive survey designs, there are, however, various methodological and practical considerations that should be addressed. First, our approach is applied to a frequently-repeated cross-sectional survey. Historic data in such surveys has rich resources for relatively robust estimates of model coefficients and for making reliable predictions. When a survey is novel or conducted infrequently at a statistics bureau, our approach may be less powerful. Second, we assume that stratification is done through a fully-saturated model, i.e., strata are pre-specified by some auxiliary variables that are strong predictors for web responses. How does the prediction performance change when adding less influential auxiliary variables? It is important to assess the sensitivity of reliable predictions to the choice of auxiliary variables. Also, we assume that strata are fixed throughout the time series. In survey practice, the selected auxiliary data may gradually change over time, and thus also the relevance of certain strata. Hence, it is essential to consider auxiliary-related change in stratification when predicting responses. Third, we assume the design of a survey should be the same over time, i.e., the model assumptions must be valid over the whole time series. If an intervention or another self-reported mode (e.g., smartphone) is introduced, variation in responses caused by this must be included explicitly. The advanced method is needed because there is no prior historic knowledge for a design change before it happens. A large jump can be caused by the inclusion of such a change in the model and, before the model can be informative about the effects of the change on RPs, a sufficiently long historic sampling must first be acquired.

We see also some limitations to the proposed methodology. In one particular year of the Dutch Health Survey data, we find a sudden increase in the standard deviations of predicted response propensities and overall quality indicators. The increase was the result of the intervention (smartphones were introduced as devices as well as conditional incentives). The results show that the model can be sensitive to design change. Hence, accounting for design changes is necessary and will temporarily reduce prediction performance.

Future research needs to address conditional response predictions in mixed-mode survey designs. In this paper, we focus only on single mode response predictions. Such considerations are worthwhile for optimizing decisions of adaptive survey designs, for example, whether to switch to a cheap or expensive mode given the budget. Our method paves the way for the development of such conditional models.

Currently, the proposed model is designed for repeated cross-sectional surveys, but one may extend to other survey and sampling designs such as rotating panels. Such an extension would imply that panel response/attrition propensities are added to the model vector, and that the correlation structure among the propensities needs to be revisited.

Acknowledgements

This project was generously supported by Chinese Scholarship Council (CSC), Statistics Netherlands (CBS) and Utrecht University. We would like to thank Anouk Roberts for fruitful discussion.

Appendix A

Table A.1
Auxiliary variables form 20 strata and season is considered as an influential factor to predict response propensities

		Category
Auxiliary Variable	Gender	Male
		Female
	Age	Youth (<= 17)
		Young (18-34)
		Middle-aged (35-54)
		Old (55-64)
		Retired (>= 65)
	Ethnicity	Western (incl. native, first and second western generation)
		Non-western (incl. first and second non-western generation)
Variable	Season	Winter (January-February)
		Spring (March-May)
		Summer (June-August)
		Autumn (September-November)
		Christmas (December)

Appendix B

Figure B.1 The posterior means of global time trends u_t under M7 to M10.

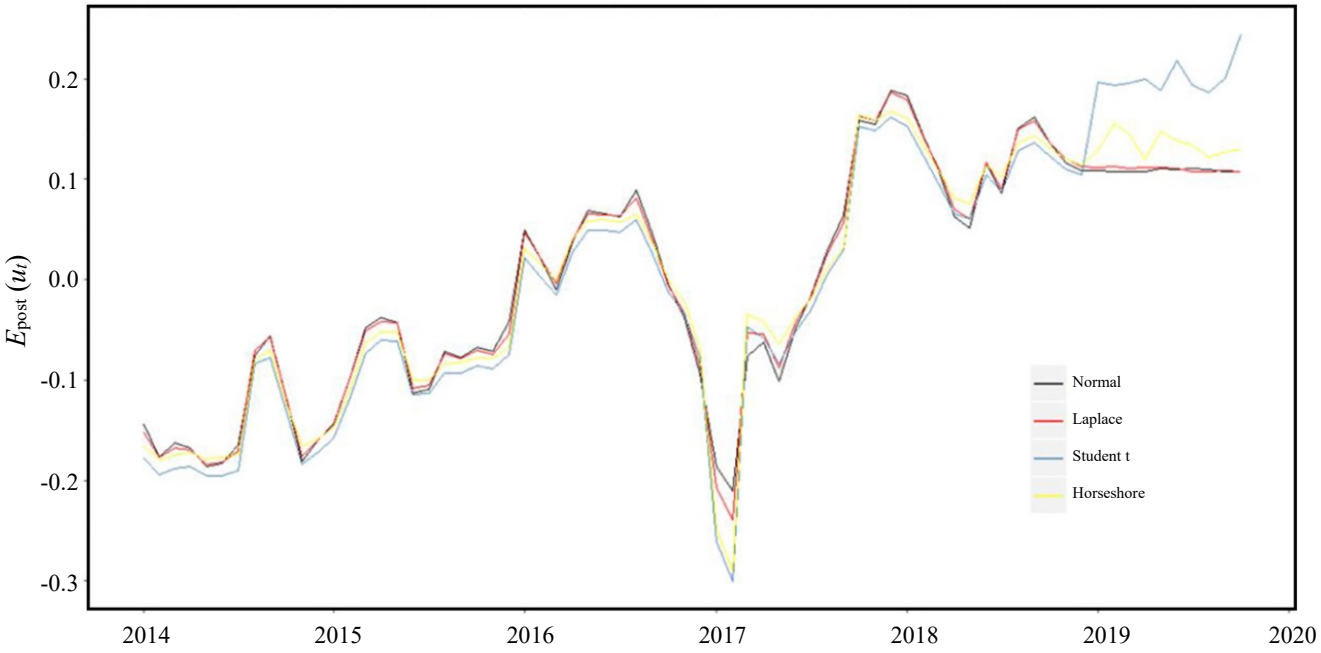
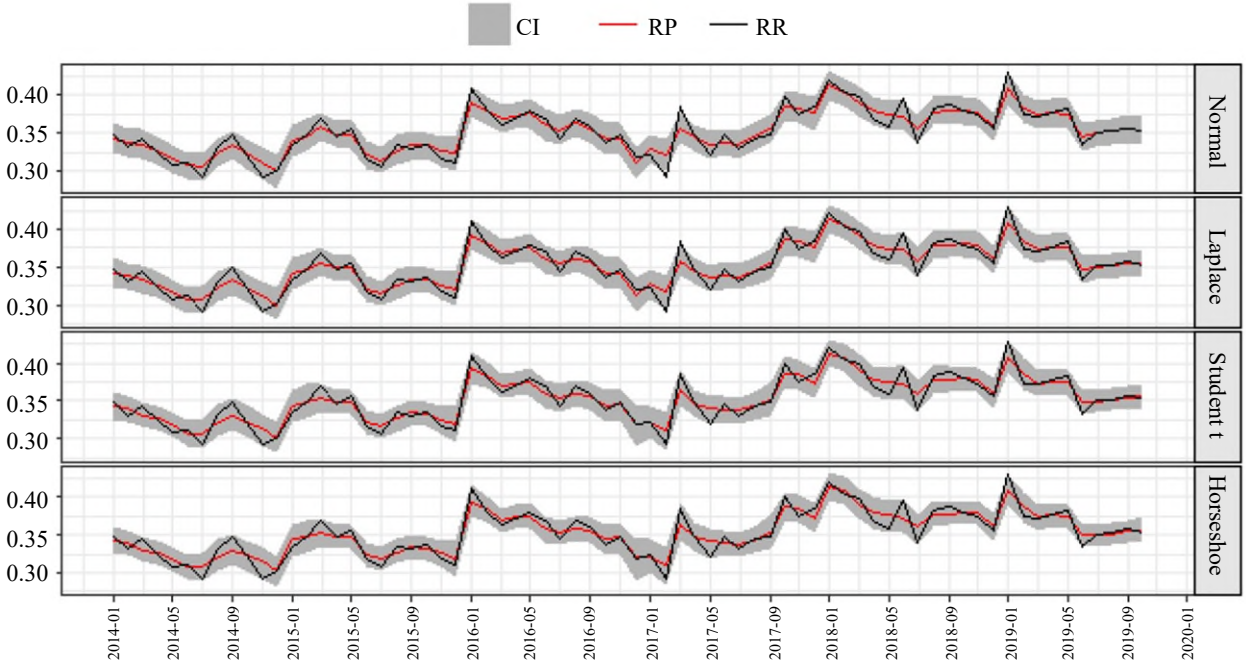
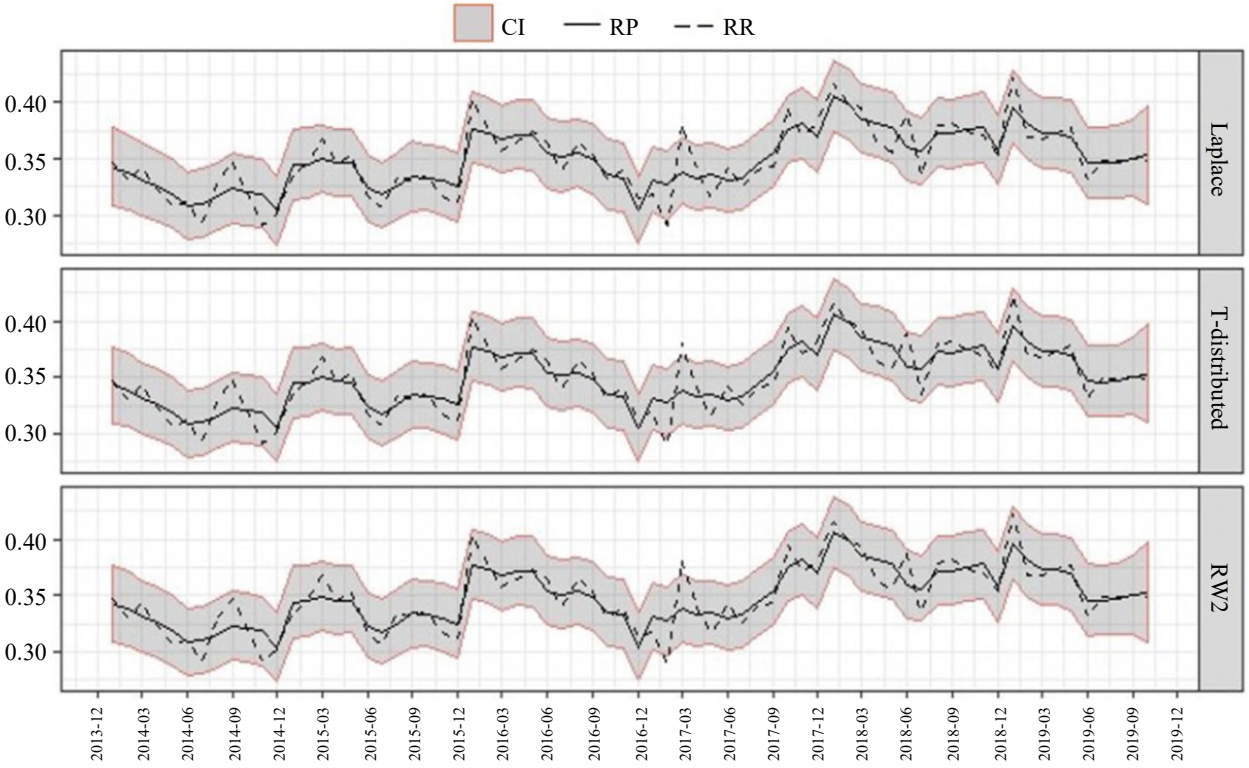


Figure B.2 Compare the posterior predictions of RP over strata made by four models (M7 to M10) to the observed RR and make a choice on the most compatible model with the observed outliers.



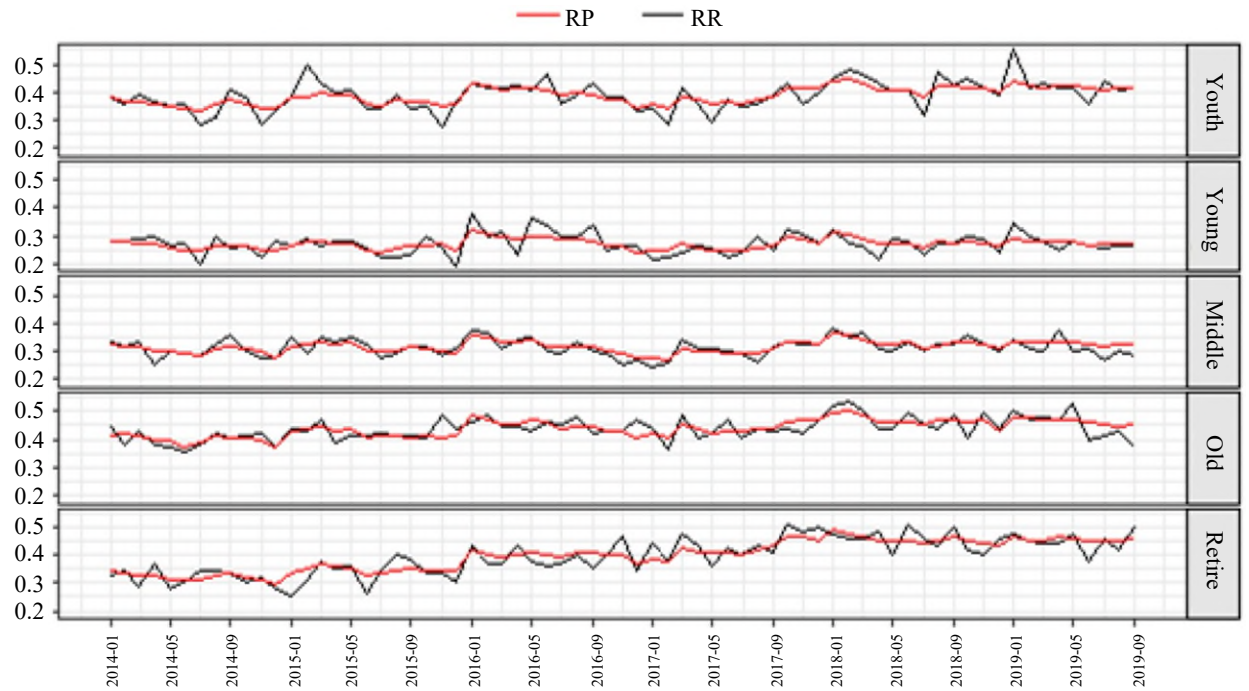
Note: The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

Figure B.3 The posterior predictions of RP over strata against the observed RR under M8, M9 and M11.



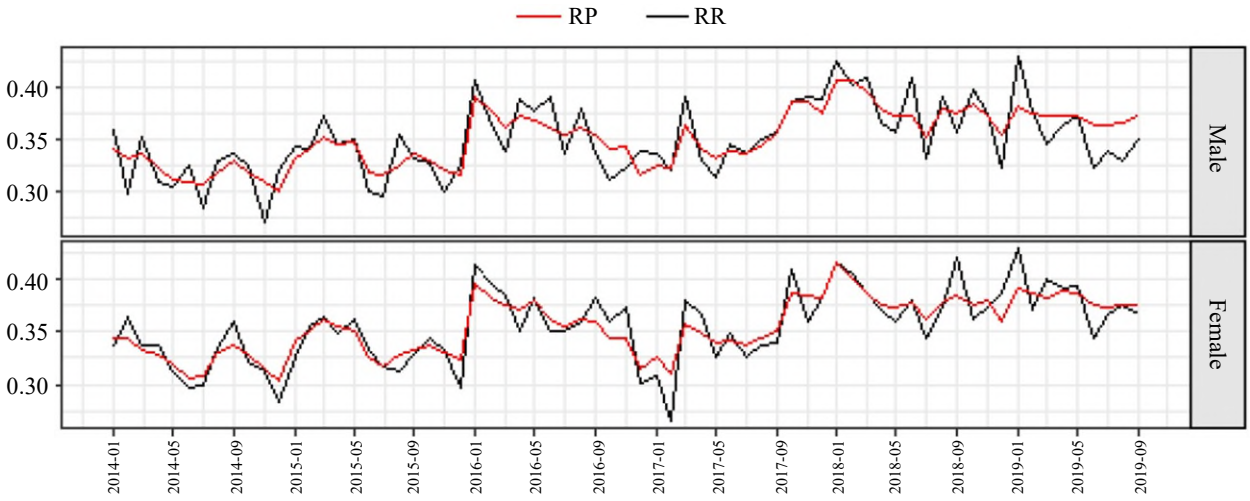
Note: The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

Figure B.4 Monthly posterior means of RP of Age groups versus observed response rates (RR) of Age groups.



Note: Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

Figure B.5 Monthly posterior means of RP of Gender groups versus observed response rates (RR) of Gender groups.



Note: Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

Appendix C

The precision matrix Q_u contains the neighbor structure of the trend innovations (e.g., Rue and Held, 2005). For first order random walk it is

$$\mathbf{Q}_u = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 & \end{pmatrix},$$

and for second order random walk is

$$\mathbf{Q}_u = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 5 & -2 & \\ & & & & & 1 & -2 & 1 & \end{pmatrix}.$$

The DIC is defined as

$$\text{DIC} = -2\left(\log p\left(r \mid n, E_{\text{post}}\,\theta\right) - p_{\text{DIC}}\right),$$

$$p_{\text{DIC}} = 2\left(\log p\left(r \mid n, E_{\text{post}}\,\theta\right) - E_{\text{post}} \log p\left(r \mid n, \theta\right)\right),$$

where $E_{\text{post}}(\theta)$ is the posterior mean of the latent parameter, $p\left(r \mid n, E_{\text{post}}(\theta)\right)$ is the likelihood evaluated at the posterior mean of θ and p_{DIC} is an estimate of the effective number of model parameters. Models with lower DIC values are preferred.

The WAIC is defined as

$$\text{WAIC} = -2 \sum_{i \in [1,\,\text{GT}]} \log E_{\text{post}} p\left(r_i \mid n_i, \theta\right) + 2p_{\text{WAIC}},$$

$$p_{\text{WAIC}} = 2\left(\sum_{i \in [1,\,\text{GT}]} \left(\log E_{\text{post}} p\left(r_i \mid n_i, \theta\right) - E_{\text{post}} \log p\left(r_i \mid n_i, \theta\right)\right)\right).$$

Here $p\left(r_i \mid n_i, \theta\right)$ is the pointwise-likelihood for stratum-by-time combination i. Similar to DIC, models with lower WAIC values are preferred.

Appendix D

The binomial multi-level time-series models are fit using a Gibbs sampler. For the derivation of the set of full conditional distributions we refer to the (appendix of the) technical report version of Boonstra and van den Brakel (2022). There, the Gibbs sampler has been worked out for a general class of multilevel models, which encompasses the set of models discussed here, except for the fact that here we employ a binomial instead of Gaussian data distribution. Fortunately, the use of the scale-mixture data augmentation approach for binomial-logistic likelihoods (Polson, Scott and Windle, 2013) ensures that the same closed-form full conditional distributions as in the Gaussian case can be used with only minimal changes to their parameters, along with an additional full conditional distribution for the auxiliary latent scale factors. To start with the latter, the full conditional for scale factor ω_i is given by

$$p\left(\omega_i \mid r, .\right) = \text{PG}\left(\omega_i \mid n_i, \theta_i\right)$$

independently for all i. For notational simplicity we use index i instead of the double index g, t used in the main text, and r denotes the full observed response vector. Here θ_i is the linear predictor, and $\text{PG}\left(\omega_i \mid n_i, \theta_i\right)$ denotes the Pólya-Gamma distribution with parameters n_i and θ_i, see Polson, Scott and Windle (2013). The coefficients' full conditionals change only in their parameters. For example, in the full conditional for a general random effects component, equation (A.28) in the technical report, the precision matrix Σ^{-1} becomes $\Sigma^{-1} = \text{diag}(\omega)$ and the response vector y gets replaced by "working response" $\frac{r-n/2}{\omega}$. The same holds true for the full conditionals of the fixed effects and auxiliary parameters ξ. All other full conditionals remain unchanged.

References

Boonstra, H. (2021). mcmcsae: MCMC small area estimation. *R Package Version 0.6.0*. https://cran.r-project.org/web/packages/mcmcsae/index.html.

Boonstra, H.J., and van den Brakel, J.A. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 3, 395-425. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00005-eng.pdf.

Boonstra, H.J., and van den Brakel, J.A. (2022). Multilevel time-series models for small area estimation at different frequencies and domain levels. Accepted for publication in *Annals of Applied Statistics*. Technical Report 2018-12, https://www.cbs.nl/en-gb/background/2018/50/models-for-estimation-at-various-aggregation-levels, Statistics Netherlands.

Burger, J., Perryck, K. and Schouten, J.G. (2017). Robustness of adaptive survey designs to inaccuracy of design parameters. *Journal of Official Statistics*, 33(3), 687-708. https://doi.org/10.1515/jos-2017-0032.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480. https://doi.org/10.1093/biomet/asq017.

Chun, A.Y., Heeringa, S. and Schouten, J.G. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34(3), 581-597. 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*, 14(2), 159-194. https://doi.org/10.12758/mda.2020.05.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the US. *Journal of the American statistical Association*, 94(448), 1074-1082.

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Fang, Q., Burger, J., Meijers, R. and van Berkel, K. (2020). The role of time, weather and google trends in understanding and predicting Web survey response. *Survey Research Methods*, 15(1), 1-25. https://doi.org/10.18148/srm/2021.v15i1.7633.

Franco, C., and Bell, W.R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. *Statistics in Transition New Series*, 16(4), 563-584. https://sciendo.com/article/10.21307/stattrans-2015-033.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533. https://doi.org/10.1214/06-ba117a.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472. https://doi.org/10.1214/ss/1177011136.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733-760.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*. CRC press.

Harvey, A.C. (1990). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Kreuter, F. (2013). Facing the nonresponse challenge. *Annals of the American Academy of Political and Social Science*, 645(1), 23-35. https://doi.org/10.1177/0002716212456815.

Mushkudiani, N., and Schouten, B. (2019). Time-Dependent Survey Design Parameters: Choosing the Length of Historic Survey Data in a Bayesian Analysis, Application to the Dutch Health Survey. Workshop paper for advances in adaptive and responsive survey design.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf.

Polson, N.G., and Scott, J.G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9, 501-538. https://doi.org/10.1093/acprof:oso/9780199694587.003.0017.

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(504), 1339-1349. https://doi.org/10.1080/01621459.2013.829001.

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

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley StatsRef: Statistics Reference Online, 1-8.

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. https://doi.org/10.2307/3315407.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. https://doi.org/10.1093/biomet/70.1.41.

Rue, H., and Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications. CRC press.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*. CRC press.

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*, 6(4), 431-464. https://doi.org/10.1093/jssam/smy012.

Spiegelhalter, D.J., Best, N.G., Carlin, B.R. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64, 583-616. https://doi.org/10.1111/1467-9868.00353.

Wagner, J., and Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, 2(3), 323-342. https://doi.org/10.1093/issam/smu009.

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. http://jmlr.org/papers/v11/watanabe10a.html.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3), 646-648. https://doi.org/10.1093/biomet/74.3.646.

West, B.T., Wagner, J., Coffey, S. and Elliott, M.R. (2021). 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*. https://doi.org/10.1093/jssam/smab036.

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. https://sciendo.com/es/article/10.2478/jos-2022-0028.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 191-211
Statistics Canada, Catalogue No. 12-001-X

Sampling with adaptive drawing probabilities

Bardia Panahbehagh, Yves Tillé and Azad Khanzadi[1]

Abstract

In this paper, with and without-replacement versions of adaptive proportional to size sampling are presented. Unbiased estimators are developed for these methods and their properties are studied. In the two versions, the drawing probabilities are adapted during the sampling process based on the observations already selected. To this end, in the version with-replacement, after each draw and observation of the variable of interest, the vector of the auxiliary variable will be updated using the observed values of the variable of interest to approximate the exact selection probability proportional to size. For the without-replacement version, first, using an initial sample, we model the relationship between the variable of interest and the auxiliary variable. Then, utilizing this relationship, we estimate the unknown (unobserved) population units. Finally, on these estimated population units, we select a new sample proportional to size without-replacement. These approaches can significantly improve the efficiency of designs not only in the case of a positive linear relationship, but also in the case of a non-linear or negative linear relationship between the variables. We investigate the efficiencies of the designs through simulations and real case studies on medicinal flowers, social and economic data.

Key Words: Adaptive sampling; Efficiency; Regression models; Sampling design.

1. Introduction

In probability proportional to size sampling (PS), the sample units are selected proportional to size of an auxiliary variable. The sampling design with unequal probabilities with-replacement, PPS, is first introduced by Hansen and Hurwitz (1943). Madow (1949), Narain (1951) and Horvitz and Thompson (1952) proposed without-replacement versions of PPS as ΠPS. Many different schemes have been proposed for ΠPS of which 50 of them are listed in Brewer and Hanif (1983) and Tillé (2006, 2020). Almost all of these methods use the π-estimator (Narain, 1951; Horvitz and Thompson, 1952) to derive an unbiased estimator of the population total and its variance estimator. Generally ΠPS is more efficient than PPS, however PPS offers advantages over ΠPS with respect to simplicity of the sample selection and the variance estimator calculations.

Our goal is to improve PPS and ΠPS designs based on an adaptive approach. The word "adaptive" refers to the use of information from sampled units in the sampling process (Seber and Salehi, 2013). In adaptive designs, it is not possible to select the final sample before starting the sampling process. The concept of an adaptive design is to use the information from the observed sample units to obtain as much information as possible about the population. The proposed approaches are easy to implement. In Section 2, adaptive PPS and, in Section 3, adaptive ΠPS sampling are presented. Section 4 and 5 contain simulations and real case studies to evaluate the effectiveness of PPS sampling and ΠPS sampling, respectively. Conclusions are drawn in Section 6.

1. Bardia Panahbehagh, Institute of Statistics, University of Neuchatel, Neuchatel, Switzerland and Department of Mathematics, Kharazmi University, Tehran, Iran. E-mail: panahbehagh@khu.ac.ir; Yves Tillé, Institute of Statistics, University of Neuchatel, Neuchatel, Switzerland; Azad Khanzadi, Department of Economics, Razi University, Kermanshah, Iran.

2. Adaptive PPS (APPS) sampling

Assume that we have a finite population whose set of labels is denoted by $U=\{1,\ldots,k,\ldots,N\}$. The variable of interest is $\mathbf{y}=(y_1,\ldots,y_k,\ldots,y_N)^\top$ and the auxiliary variable is $\mathbf{x}=(x_1,\ldots,x_k,\ldots,x_N)^\top$. Both variables are assumed to be positive and non-zero, i.e., $\mathbf{y},\mathbf{x}\in\mathbb{R}^N_{>0}$. Suppose that the parameter of interest is the total of the variable of interest,

$$t_y = \sum_{k\in U} y_k.$$

The total of the auxiliary variable is denoted by

$$t_x = \sum_{k\in U} x_k.$$

Also, for any subset A of U with cardinality N_A, we define

$$t_{yA} = \sum_{k\in A} y_k,\quad \bar{y}_A = \frac{1}{N_A}\sum_{k\in A} y_k \quad \text{and} \quad t_{xA} = \sum_{k\in A} x_k.$$

The basic idea behind APPS is to update the vector of auxiliary variables based on the information of the observed variable of interest after each draw. To take an APPS sample of size n, we proceed as described in Algorithm 1.

Algorithm 1. Adaptive PPS (APPS)

Define

$$\mathbf{p}_1 = \frac{\mathbf{x}}{t_x} = (p_{11},\ldots,p_{1k},\ldots,p_{1N})^\top = \left(\frac{x_1}{t_x},\ldots,\frac{x_k}{t_x},\ldots,\frac{x_N}{t_x}\right)^\top. \tag{2.1}$$

Define $s_0=\{\ \}$

For $i=1,\ldots,n$ do

- Select a unit (say j) in U with probabilities $\mathbf{p}_i=(p_{i1},\ldots,p_{ik},\ldots,p_{iN})^\top$.
- Define $s_i=s_{i-1}\cup\{j\}$.
- Compute $\mathbf{p}_{i+1}=(p_{(i+1)1},\ldots,p_{(i+1)k},\ldots,p_{(i+1)N})^\top$, where

$$p_{(i+1)k} = \begin{cases} \dfrac{y_k t_{xs_i}}{t_x t_{ys_i}} & \text{if } k\in s_i \\ p_{1k} & \text{if } k\notin s_i \end{cases}, \quad \text{for all} \quad k\in U. \tag{2.2}$$

In Algorithm 1, the first two units are selected with-replacement using $\mathbf{p}_1$ and $\mathbf{p}_2\left(=\mathbf{p}_1\right)$ respectively and we observe their y values. Indeed, according to (2.2) in Algorithm 1, after observing the y value of the first sample unit (say j) we have

$$p_{2k} = \begin{cases} \dfrac{y_j x_j}{t_x y_j} = \dfrac{x_j}{t_x} = p_{1j} & \text{if } k = j \\ p_{1k} & \text{if } k \neq j \end{cases}, \quad \text{for all} \quad k \in U,$$

or briefly $\mathbf{p}_2 = \mathbf{p}_1$. Therefore at least the y values of two different units are required to update the drawing probabilities vector $(\mathbf{p})$. For the third unit onwards, based on the observed y values, we update the vector of drawing probabilities. Each unit is then selected using a different drawing probabilities vector.

Result 1. *In APPS, for each* $i = 1, \ldots, n$,

$$\sum_{k \in U} p_{ik} = 1,$$

and

$$\hat{t}_{\text{APPS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{k_i}}{p_{ik_i}}$$

is an unbiased estimator of t_y *with variance*

$$V\left(\hat{t}_{\text{APPS}}\right) = E\left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{N} \left(\frac{y_k}{p_{ik}} - t_y\right)^2 p_{ik}\right].$$

An unbiased estimator of the variance is given by:

$$\hat{V}\left(\hat{t}_{\text{APPS}}\right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_{k_i}}{p_{ik_i}} - \hat{t}_{\text{APPS}}\right)^2,$$

where p_{ik_i} *is* k_i^{th} *unit of* $\mathbf{p}_i = \left(p_{i1}, p_{i2}, \ldots, p_{iN}\right)^{\top}$.

For the proof of Result 1, see Appendix A.

Setting the drawing probabilities exactly proportional to size of y, i.e., $p_k = y_k / t_y$, will lead to an unbiased estimator for t_y with zero variance,

$$\hat{t}_{\text{APPS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{k_i}}{p_{ik_i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{k_i}}{y_{k_i} / t_y} = t_y.$$

By following the procedure of Algorithm 1 step by step, the drawing probability for unit k approaches the ideal probability proportional to size based on y. As evidence, consider that if units k and ℓ have

been selected at least once in steps up to and including i, then in all the steps after step i, the ratio of their drawing probabilities is equal to y_k/y_ℓ, which is the same as the ideal case,

$$\frac{p_{jk}}{p_{j\ell}} = \left(\frac{y_\ell\, t_{xs_j}}{t_x\, t_{ys_j}}\right)^{-1} \frac{y_k\, t_{xs_j}}{t_x\, t_{ys_j}} = \frac{y_k/t_y}{y_\ell/t_y}, \quad j > i.$$

3. Adaptive ΠPS (AΠPS) sampling

In general, without-replacement designs are more efficient than with-replacement designs of the same size due to the inclusion of unduplicated information. AΠPS is a kind of adaptive version of ΠPS. To take a AΠPS sample of size n, we proceed as described in Algorithm 2.

Algorithm 2. Adaptive ΠPS (AΠPS)

1. Based on a conventional design (like Simple Random Sampling without-replacement (SRSWOR) or ΠPS) an initial sample s_0 of size n_0 will be selected.
2. Using s_0, y is modeled, for example, by a polynomial of order M of x to detect the potentially non-linear relationship between x and y. In other words, we assume a superpopulation model as

$$y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \ldots + \beta_M x_k^M + \varepsilon_k, \; k \in U,$$

where ε_k is a random variable independent of x_k with $E(\varepsilon_k) = 0$ and then

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2 + \ldots + \hat{\beta}_M x_k^M, \; k \in U^* = U \setminus s_0,$$

where $\hat{\beta}_m, m = 1, \ldots, M,$ can be estimated using the least square method in finite population sampling. If $\hat{y}_k$ is negative or null, it is replaced by a small positive value of the y, so as not to have zero inclusion probabilities.

3. Based on $\hat{y}_k, k \in U^*$ we select a ΠPS of size $n^* = n - n_0$, say s^*.

In Algorithm 2, one can obviously use any parametric or non-parametric model instead of a linear model to obtain a forecast of y_k. Our sampling method will be all the more efficient if the prediction of y_k is accurate. The predicted values must be positive. With an AΠPS sampling design, we can estimate the population total by

$$\hat{t}_{A\Pi\text{PS}} = \sum_{k \in s^*} \frac{y_k}{\Pi_k^*} + \sum_{k \in s_0} y_k,$$

with

$$\Pi_k^* = E\left(I_k \mid s_0\right) = \min\left(c^* \hat{y}_k, 1\right),$$
$$\text{where constant } c^* \text{ is defined by } \sum_{k \in U^*} \min\left(c^* \hat{y}_k, 1\right) = n^*, \tag{3.1}$$

where I_k is an indicator function which takes 1 if unit k is selected as a unit of s^*. With inclusion probabilities exactly proportional to the size of y as in (3.1), provided that

$$0 < \frac{n^* y_k}{t_{yU^*}} \leq 1, \quad \text{for all} \quad k \in U^*,$$

we will have

$$\Pi_k^* = \frac{n^* y_k}{t_{yU^*}}, \quad \text{for all} \quad k \in U^*,$$

which will lead to an unbiased estimator for t_y with zero variance,

$$\hat{t}_{A\Pi\text{PS}} = \sum_{k \in s^*} \frac{y_k}{\Pi_k^*} + \sum_{k \in s_0} y_k = \sum_{k \in s^*} \frac{y_k}{n^* y_k / t_{yU^*}} + \sum_{k \in s_0} y_k = t_{yU^*} + t_{ys_0} = t_y.$$

Then, if we can estimate the y values with high accuracy based on Algorithm 2 and using initial sample s_0, we can estimate t_y with high efficiency.

Result 2. *In $A\Pi PS$,*

$$\hat{t}_{A\Pi\text{PS}} = \sum_{k \in s^*} \frac{y_k}{\Pi_k^*} + \sum_{k \in s_0} y_k,$$

is an unbiased estimator of t_y with variance

$$V\left(\hat{t}_{A\Pi\text{PS}}\right) = E\left(\sum_{k \in U^*} \sum_{\ell \in U^*} \frac{y_k}{\Pi_k^*} \frac{y_\ell}{\Pi_\ell^*} \Delta_{k\ell}^*\right),$$

and provided that all the $\Pi_{k\ell}^*$ are strictly positive (which depends on the sampling design used in U^*), an unbiased estimator of variance is

$$\hat{V}\left(\hat{t}_{A\Pi\text{PS}}\right) = \sum_{k \in s^*} \sum_{\ell \in s^*} \frac{y_k}{\Pi_k^*} \frac{y_\ell}{\Pi_\ell^*} \frac{\Delta_{k\ell}^*}{\Pi_{k\ell}^*},$$

where $\Delta_{k\ell}^* = \Pi_{k\ell}^* - \Pi_k^* \Pi_\ell^*$ and $\Pi_{k\ell}^* = E\left(I_k I_\ell \mid s_0\right)$.

For the proof of Result 2, see Appendix B.

In $\hat{t}_{A\Pi PS}$, for extreme cases where the size of s^* is too small, we may exaggerate the role of s^* in estimation relative to s_0. Then we can adjust the estimator by adding a weighting parameter, say $0 \le \alpha \le 1$ as follows:

$$\hat{t}_{A\Pi PS\alpha} = N\left[\alpha \frac{1}{N-n_0}\sum_{k\in s^*}\frac{y_k}{\Pi_k^*} + (1-\alpha)\frac{1}{n_0}\sum_{k\in s_0} y_k\right].$$

Result 3. *In* $A\Pi PS\alpha$, *if we select* s_0 *by SRSWOR, then*

(i) $\hat{t}_{A\Pi PS\alpha} = \hat{t}_{A\Pi PS}$, *for* $\alpha = (N-n_0)/N$,

(ii) $\hat{t}_{A\Pi PS\alpha}$ *is unbiased,* $E\left(\hat{t}_{A\Pi PS\alpha}\right) = t_y$,

(iii) *with the following variance*

$$V\left(\hat{t}_{A\Pi PS\alpha}\right) = N^2\frac{\alpha^2}{(N-n_0)^2}E\left(\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right) + N^2\left(1-\frac{\alpha}{1-f_0}\right)^2\left(\frac{1-f_0}{n_0}S_y^2\right),$$

where $f_0 = n_0/N$,

(iv) *and an unbiased estimator of the variance, provided that all the* $\Pi_{k\ell}^*$ *are strictly positive, is*

$$\hat{V}\left(\hat{t}_{A\Pi PS\alpha}\right) = N^2\frac{\alpha^2}{(N-n_0)^2}\sum_{k\in s^*}\sum_{\ell\in s^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\frac{\Delta_{k\ell}^*}{\Pi_{k\ell}^*} + N^2\left(1-\frac{\alpha}{1-f_0}\right)^2\left(\frac{1-f_0}{n_0}s_{0y}^2\right),$$

where

$$S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(y_k-\bar{y}_U\right)^2 \quad \textit{and} \quad s_{0y}^2 = \frac{1}{n_0-1}\sum_{k\in s_0}\left(y_k-\bar{y}_{s_0}\right)^2,$$

(v) *The optimal value for* α *to minimize the variance of the estimator is*

$$\alpha^* = (1-f_0)\frac{\frac{1-f_0}{n_0}S_y^2}{E\left(\frac{1}{N^2}\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right)+\frac{1-f_0}{n_0}S_y^2}. \tag{3.2}$$

4. Simulations for APPS Sampling

In order to evaluate the designs, we have run simulations on real data. All the simulations in Section 4 and Section 5 have been implemented using Monte Carlo methods with 2,000 iterations. We use a real

case study of medicinal flowers and real data from the statistical center of Iran between 2015-2016 (https://www.amar.org.ir) to evaluate the results of Section 2. To compare the designs, the efficiency is defined by

$$\text{Efficiency} = F_{\bullet} = \frac{V\left(N\bar{y}_s\right)}{V\left(\hat{t}_{\bullet}\right)}, \tag{4.1}$$

where $\bar{y}_s$ is the sample mean in Simple Random Sampling with-replacement (SRSWR) with size n and $\hat{t}_{\bullet}$ indicates the Hansen-Hurwitz estimator in PPS, APPS with n draws or π -estimator in ΠPS. In each case, we indicate the variable of interest and the auxiliary variable. Drawing probabilities for PPS and APPS are calculated based on (2.1) and (2.2) respectively. Also inclusion probabilities for ΠPS are calculated based on the auxiliary variable using (3.1). For inclusion probabilities in (3.1) we used U and m instead of U^* and n^* respectively. As APPS and ΠPS are with and without-replacement designs respectively, in order to have a fair comparison, the cost of the sample needs to be as equal as possible for all of the designs. For this purpose, in each iteration an APPS is implemented first, and then for ΠPS, the sample size, is set to the number of distinct units obtained with n draws in APPS. To implement the ΠPS in this section we used the eliminatory method based on `UPtille` function available in the `R` package `sampling` (Tillé and Matei, 2015).

For the simulations, we considered two kinds of data:

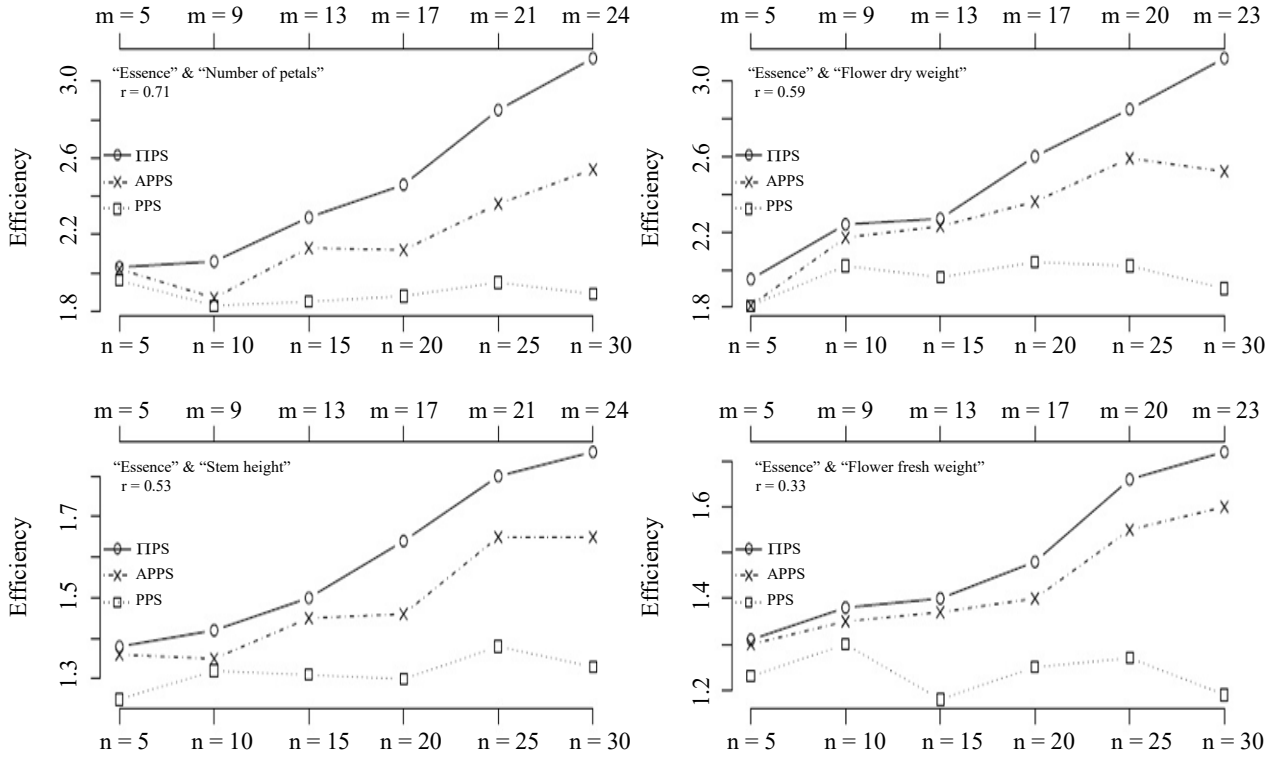
1. *Medicinal Flowers*: The data come from a real case study on chamomile flower (Panahbehagh, Bruggemann, Parvardeh, Salehi and Sabzalian, 2018) as the medicinal use of flowers. The population mean of the "Essence" is the parameter of interest with $t_y = 44.4$ and $N = 60$. In practice, the variable of interest is not know prior to sampling so we use four readily available auxiliary variables with various correlations with the variable of interest. The four auxiliary variables and correlations are "Flower fresh weight" with 0.33, "Flower dry weight" with 0.59, "Stem height" with 0.53 and "Number of petals" with 0.71. The results are presented in Figure 4.1, where the correlations are denoted by r.
2. *Social Data*: These data are from the Statistical Center of Iran gathered from 31 provinces of Iran in 2015-2016 (www.amar.org.ir). Marriage-Divorce and Academic degrees data are official statistics covering all target populations, based on the "National Organization for Civil Registration" and the "Ministry of Science, Research and Technology" respectively. In addition, the provincial population sizes are based on the 2016 census in Iran. We considered four situations having an auxiliary and a variable of interest:
   - The registered number of "Divorce less than 1 year" and "Marriage" as the variable of interest and the auxiliary variable respectively,
   - The registered number of "Divorce less than 1 year" and "Divorce" as the variable of interest and the auxiliary variable respectively,

– The registered number of “Bachelors” and “Diplomas” as the variable of interest and the auxiliary variable respectively,
– The registered number of “Masters and higher” and “Diplomas” as the variable of interest and the auxiliary variable respectively.

The results are presented in Figure 4.2.

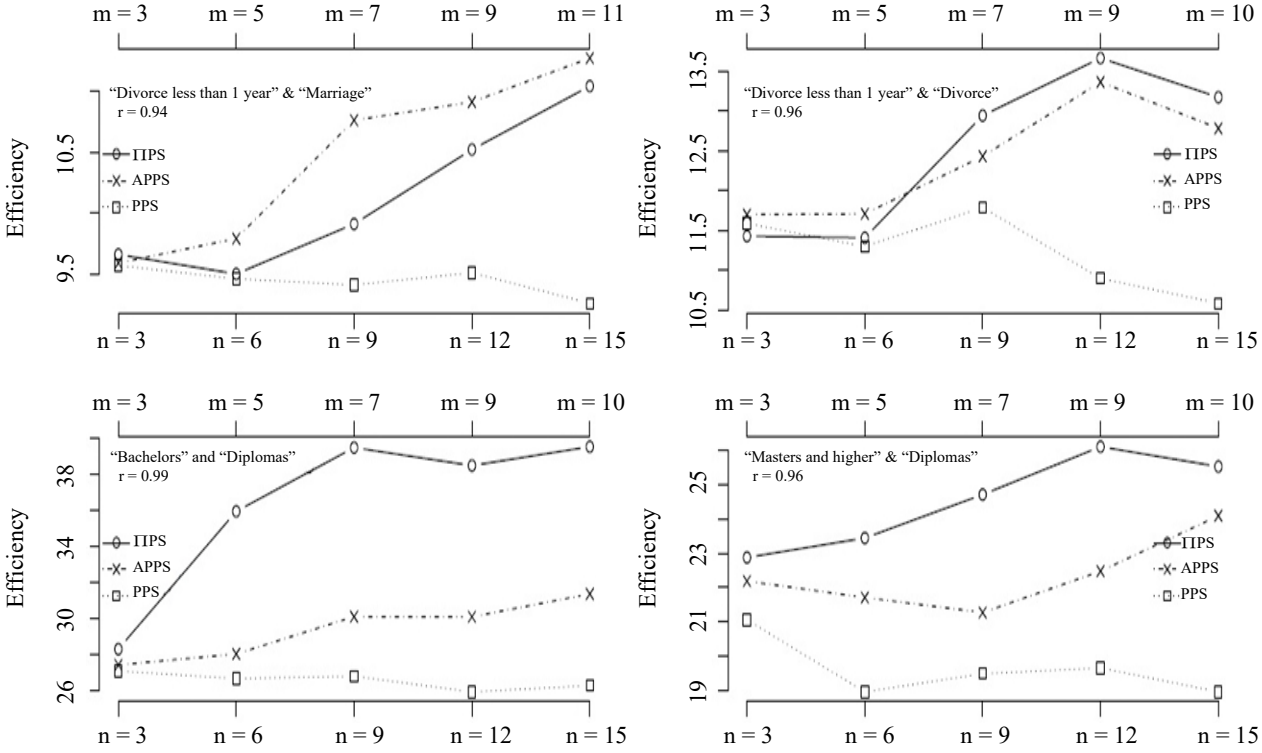
As can be seen in Figure 4.1, comparing the vertical axis, generally the higher the correlation, the higher the efficiency. By comparing Figure 4.2 and Figure 4.1, the efficiency increases dramatically for the social data compare to the medicinal flowers, which was predictable given the correlations of more than 0.90 in the former. A positive relationship between correlation and efficiency was expected because when the correlation is high, the drawing probabilities vector approximates the exact sampling probability proportional to size more accurately.

Figure 4.1 Efficiencies of ΠPS, APPS and PPS relative to SRSWR for the medicinal flowers data with different auxiliary variables. *m* is the size of ΠPS which is the Monte Carlo expectation of the number of distinct units in the respective with-replacement designs (PPS and APPS) of size *n*. At the top-left of each plot, the variable of interest and the auxiliary variable are indicated, with the respective Pearson correlation, indicated by *r*.



Note : Simple random sampling with-replacement (SRSWR); Proportional to size (PS); Adaptive probability proportional to size sampling (APPS); Probability proportional to size sampling (PPS).

Figure 4.2 Efficiencies of ΠPS, APPS and PPS relative to SRSWR for the social data with different auxiliary variables and different correlations. *m* is the size of ΠPS which is the Monte Carlo expectation of the number of distinct units in the respective with-replacement designs (PPS and APPS) of size *n*. At the top-left of each plot, the variable of interest and the auxiliary variable are indicated, with the respective Pearson correlation, indicated by *r*.



Note : Simple random sampling with-replacement (SRSWR); Proportional to size (PS); Adaptive probability proportional to size sampling (APPS); Probability proportional to size sampling (PPS).

In Figure 4.1, for the medicinal flowers data, APPS is more efficient than PPS in all cases. The efficiency of PPS fluctuates slightly with the variation of *n*, which shows that increasing the sample size, improves both SRSWR and PPS at the same level. But at the same time, the efficiency of APPS generally increases with increasing sample size. In fact, in APPS, the larger the sample size, the more updated the auxiliary variable units, and therefore the more accurate the exact proportional to size approximation. Furthermore, although the efficiency of APPS is much closer to ΠPS compared to PPS in most cases, the efficiency of ΠPS, particularly for large sample sizes, is higher than the other two in all Figure 4.1 cases and discrepancy increases with increasing *n*. As a final point in Figure 4.1, the efficiency of APPS is often about the same as ΠPS if the sample size is less than around 15% of the population size (a reasonable sample size).

Most of the results in Figure 4.2 are similar to the results in Figure 4.1. For the social data, like the medicinal flowers data, APPS is more efficient than PPS in all cases and the efficiency of APPS generally increases (with some fluctuations) as the sample size increases. But interestingly, unlike the medicinal flowers, APPS is more efficient than ΠPS in some cases. This is interesting because it is much easier to implement and calculate the estimators in APPS than in ΠPS. Eventually, the efficiency of PPS fluctuates again, but this time it tends to fall slightly as *n* increases.

5. Simulations for AΠPS

Following the notation used in (4.1), "•" indicates the particular strategy, ΠPS, AΠPS or AΠPSα and $\bar{y}_s$ is the sample mean in SRSWOR with size n. Regarding the note in step 2 of Algorithm 2 related to negative values of $\hat{y}$, we replace them with 0.0001 in the simulations.

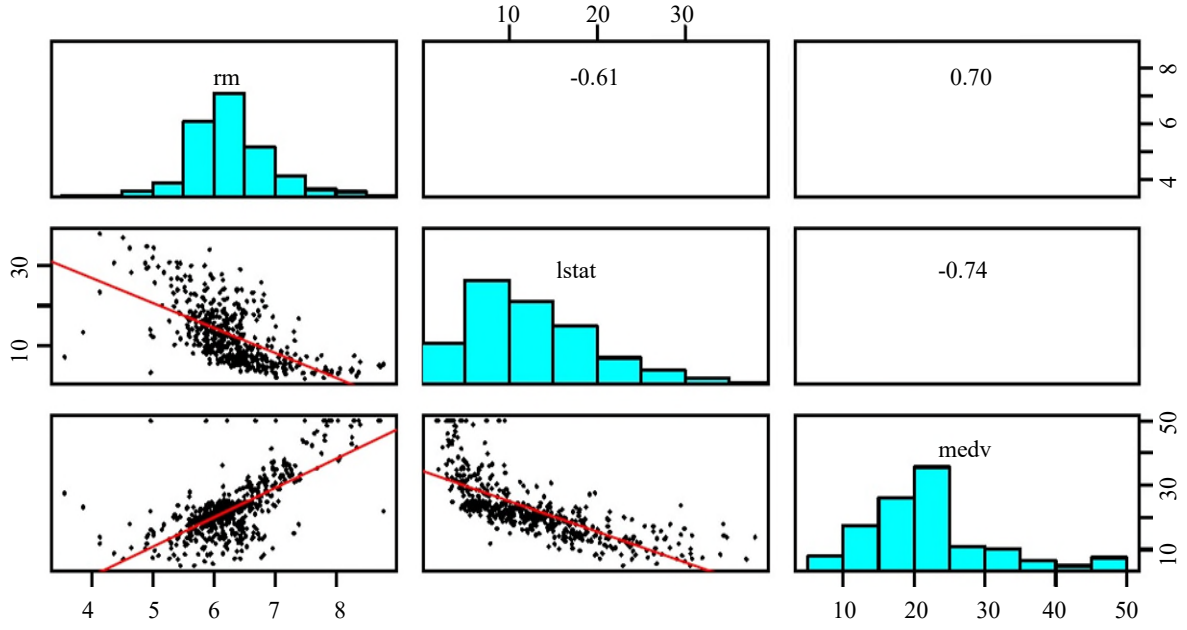
Also for ΠPS and step 3 of Algorithm 2 in AΠPS, we used the maximum entropy design based on `UPmaxentropy` function available in the `R` package `sampling` (Tillé and Matei, 2015).

5.1 Boston data

In this subsection, we analyze a dataset for the city of Boston (see Figure 5.1). Three different housing value variables for suburbs of Boston (Harrison and Rubinfeld, 1978; Belsley, Kuh and Welsch, 1980) are available in `R` package `MASS` as:

- rm: Average number of rooms per dwelling,
- lstat: Percentage of population in weak and deprived economic situation in Boston Suburbs,
- medv: Median value of owner-occupied homes in 1,000$s.

Figure 5.1 The relationship among three variables of interest, rm, lstat and medv, for the Boston data. In this 3×3 matrix of plots, the lower off-diagonal draws scatter plots with fitted linear least squares regressions, the diagonal represents histograms with the name of the variables and the upper off-diagonal reports the Pearson correlations.



Note : rm: Average number of rooms per dwelling; lstat: Percentage of population in weak and deprived economic situation in Boston Suburbs; medv: Median value of owner-occupied homes in 1,000$s.

In urban and residential areas, the larger the dimensions of a house, the more rooms one can expect to have. Also, the larger the dimensions of the house, the higher the value of the house. Therefore, there is a positive relationship between the dimensions of houses and the average number of rooms in each house and a positive relationship between the average number of rooms and the value of the house. In addition, economically disadvantaged people typically live in smaller houses, so the higher the proportion of disadvantaged people in a residential area, the greater the demand for small houses, and therefore the average number of rooms per house in that area will be lower. It follows that in a residential area there will be a negative relationship between the proportion of disadvantaged people and the average number of rooms in each house.

To model the variable of interest y on the auxiliary variable x, we used

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2 + \hat{\beta}_3 x_k^3, \tag{5.1}$$

where the coefficients are estimated based on the least squares error method. Here based on Result 3, the estimator of the optimal value of α given in (3.2) is used, where $\hat{S}_y^2 = s_{0y}^2$ and

$$\hat{E}\left(\sum_{k\in U^*} \sum_{\ell\in U^*} \frac{y_k}{\Pi_k^*} \frac{y_\ell}{\Pi_\ell^*} \Delta_{k\ell}^* \right) = \sum_{k\in s^*} \sum_{\ell\in s^*} \frac{y_k}{\Pi_k^*} \frac{y_\ell}{\Pi_\ell^*} \frac{\Delta_{k\ell}^*}{\Pi_{k\ell}^*}.$$

The results of the simulations on the Boston data are presented in Table 5.1. In all cases, AΠPSα is better than AΠPS and, in almost all cases, is more efficient than SRSWOR (except for some cases with small n and n_0). Also, for AΠPS and AΠPSα the efficiency generally increases with increasing n and n_0. In each model, the R^2's for different cases fluctuated slightly around a certain value, and predictably, the values appear to be independent of the initial sample size. Then, we have only reported the median values of the R^2's for different cases in the table.

As expected, the higher the absolute value of the correlation between x and y, the higher the R^2. Consequently, as AΠPS and AΠPSα use the regression model to predict y values, the higher the R^2, the higher the efficiencies of AΠPS and AΠPSα. Furthermore, ΠPS is better than SRSWOR only for rm-medv, which has a positive and almost linear relationship with some outliers, and ΠPS is less efficient than SRSWOR for the other two models with a negative (albeit strong) correlations. Due to the use of a regression model, AΠPS and AΠPSα are not affected by the sign of the correlations. In the rm-medv model, ΠPS is better than the others but for large sample size, AΠPSα could approach ΠPS.

Looking into Monte Carlo's results in detail, AΠPS, by exaggerating the role of s^* (as discussed in Section 3) in certain iterations, results in very biased estimates for the parameter. Since, it cannot be as efficient as SRSWOR for medv-rm and lstat-rm model, in the next simulation on economic data, we simply compare the efficiencies of AΠPSα and ΠPS designs.

Table 5.1
Efficiencies of ΠPS, AΠPS and AΠPSα with $M = 3$ for Boston data. For each case, the variable of interest y and the auxiliary variable x are specified. Initial and final sample sizes are denoted by n_0 and n respectively, F indicates efficiency and R^2 is R-squared of model $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2 + \hat{\beta}_3 x_k^3$

	n	n_0	R^2	$F_{\Pi PS}$	$F_{A\Pi PS}$	$F_{A\Pi PS\alpha}$
x = medv, y = rm, N = 506	50	15	0.63	1.65	0.59	0.83
		20			0.70	1.05
		25			0.62	1.22
	75	20		1.65	0.81	1.01
		30			0.84	1.27
		40			0.89	1.43
	100	25		1.50	0.96	1.15
		30			0.78	1.34
		50			0.89	1.36
x = stat, y = rm, N = 506	50	15	0.49	0.72	0.62	0.91
		20			0.40	1.03
		25			0.63	1.05
	75	20		0.71	0.77	0.97
		30			0.77	1.10
		40			0.71	1.19
	100	25		0.76	0.93	1.15
		30			0.86	1.23
		50			0.79	1.18
x = medv, y = lstat, N = 506	50	15	0.70	0.11	1.20	1.51
		20			1.12	1.51
		25			1.08	1.64
	75	20		0.11	1.43	1.78
		30			1.41	1.93
		40			1.23	1.76
	100	25		0.10	1.40	1.91
		30			1.42	1.79
		50			1.41	1.93

Note : rm: Average number of rooms per dwelling; lstat: Percentage of population in weak and deprived economic situation in Boston Suburbs; medv: Median value of owner-occupied homes in 1,000$s.

5.2 Economic data

Data from four different economic variables for 180 countries, partially available from 1980 to 2006, were used to evaluate the results of Section 3. The data are collected on the website of the World Bank (2021). The four variables considered in this simulation are:

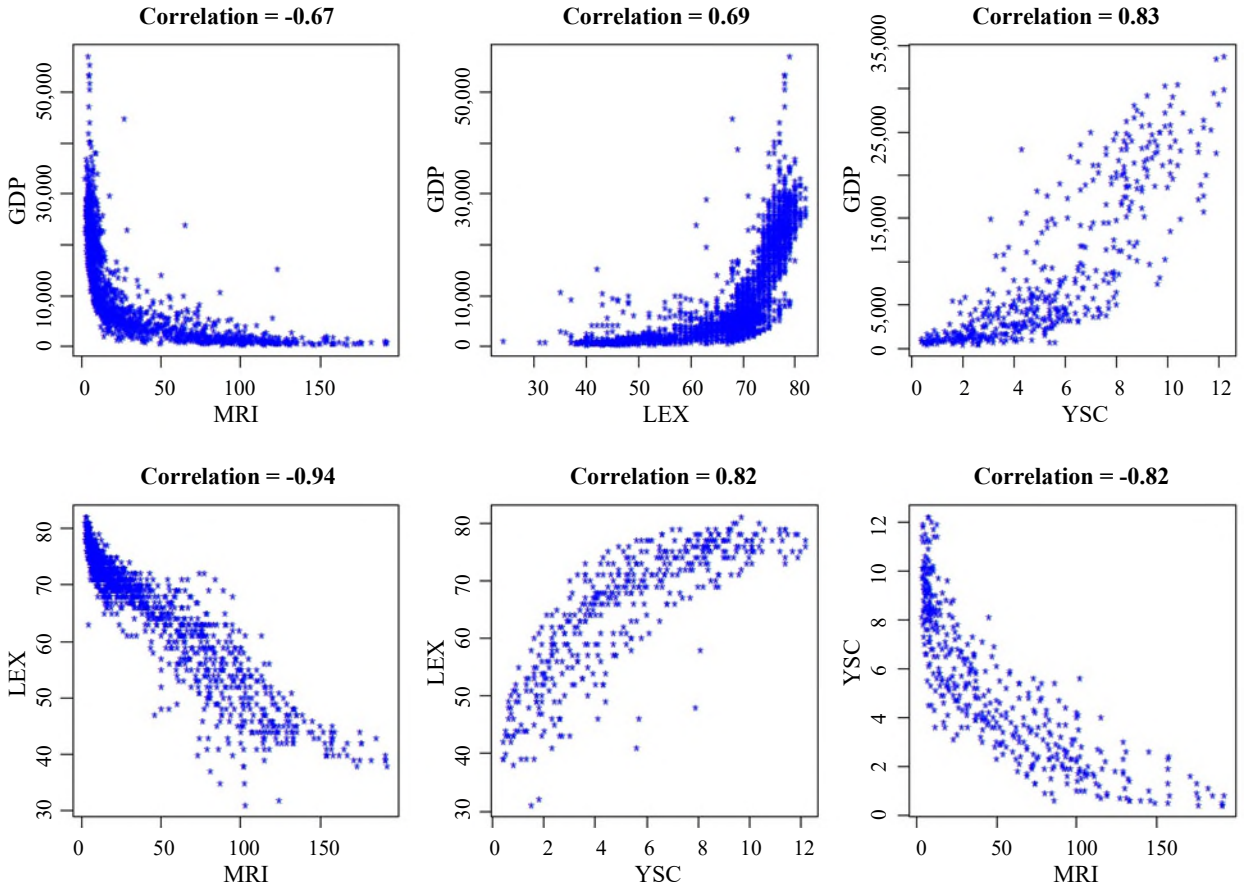
- GDP: Gross domestic product per capita based on purchasing power parity. GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2,000 international dollars.
- MRI: Mortality rate of infants per 1,000 births is the number of infants dying before reaching one year of age.

- LEX: Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
- YSC: Average schooling years in the total population aged over 25.

One of the factors of production is human resources, and the higher the quality and quantity of human resources, the higher the productivity and output of the economy. The quality of human resources can be enhanced by improving their health and well-being. Improving healthcare leads to increased life expectancy and reduced mortality. In addition, the training of human resources leads to their promotion in the fields of science and technology. Therefore, live expectancy and average years of schooling have a positive relationship with GDP per capita, and mortality rate has a negative relationship with GDP per capita.

The relationship among the four variables are presented in Figure 5.2. The population size N varies for different pairs due to the exclusion of missing data.

Figure 5.2 The relationship among the four variables in economic data. Scatter plots for the variables are shown with the Pearson correlations for the two variables at the top of each plot.



Note : Mortality rate of infants (MRI); Life expectancy at birth (LEX); Average schooling years (YSC); Gross domestic product (GDP).

The results presented in Table 5.2 can be summarized as follows:

- $A\Pi PS\alpha$ is more efficient than SRSWOR in all cases, but ΠPS is very inefficient for cases with non-linear or negative relationships. In all cases, except for model YSC-GDP, $A\Pi PS$ is more efficient than Π PS.
- MRI-GDP and LEX-GDP show almost the same pattern but with different signs. $A\Pi PS\alpha$ is efficient in both of them and is more efficient in the model with higher absolute correlation. But ΠPS is efficient for the positive relationship (LEX-GDP) and very inefficient in the negative relationship (MRI-GDP).
- For YSC-LEX, although the relationship is positive and almost linear (with $R^2 = 0.82^2 = 0.67$ for $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$), but contrary to $A\Pi PS\alpha$ which is an efficient design, ΠPS is very inefficient compared to SRSWOR.
- The correlations in both models YSC-LEX and MRI-YSC are the same, but according to R^2, it seems that regression equation (5.1) can predict $\hat{y}$ in the latter model better than the former model. Therefore the mean of the efficiencies in model MRI-YSC (2.93) is higher than model YSC-LEX (2.47).
- MRI-LEX has the highest R^2, and for large initial and final sample sizes, the efficiency of $A\Pi PS\alpha$ is the highest compared to other relationships with the same initial and final sample sizes.
- In general, increasing the sample size leads to an increase in the efficiency of $A\Pi PS\alpha$.
- For $A\Pi PS\alpha$, in all cases (except model MRI-GDP), the highest efficiency (on average) is for the largest sample size $(n = 150)$.

Table 5.2
Efficiencies of ΠPS and $A\Pi PS$ with $M = 3$ for Economic data. For each case, the variable of interest y and the auxiliary variable x are specified. Initial and final sample sizes are denoted by n_0 and n respectively, F indicates efficiency and R^2 is R-squared of model $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2 + \hat{\beta}_3 x_k^3$. The population size N, is different for different pairs due to the exclusion of missing data

	n	n_0	R^2	$F_{\Pi PS}$	$F_{A\Pi PS\alpha}$
x = MRI, y = GDP, N = 1,522	80	20	0.70	0.07	2.18
		30			2.30
		40			2.35
	100	30		0.08	2.39
		40			2.12
		50			2.13
	150	40		0.07	1.88
		60			1.94
		80			2.22
x = MRI, y = LEX, N = 1,619	80	20	0.85	0.01	1.32
		30			2.67
		40			2.94
	100	30		0.01	2.02
		40			2.81
		50			2.89
	150	40		0.01	2.62
		60			4.17
		80			3.22

Note : Mortality rate of infants (MRI); Gross domestic product (GDP); Life expectancy at birth (LEX); Average schooling years (YSC).

Table 5.2 (continued)
Efficiencies of ΠPS and AΠPS with $M = 3$ for Economic data. For each case, the variable of interest y and the auxiliary variable x are specified. Initial and final sample sizes are denoted by n_0 and n respectively, F indicates efficiency and R^2 is R-squared of model $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2 + \hat{\beta}_3 x_k^3$. The population size N, is different for different pairs due to the exclusion of missing data

	n	n_0	R^2	$F_{\Pi PS}$	$F_{A\Pi PS\alpha}$
x = LEX, y = GDP, N = 2,357	80	20	0.76	1.31	2.46
		30			2.68
		40			2.38
	100	30		1.29	2.63
		40			2.57
		50			2.14
	150	40		1.41	2.86
		60			2.58
		80			2.51
x = YSC, y = LEX, N = 452	80	20	0.75	0.08	2.29
		30			2.71
		40			2.30
	100	30		0.08	2.66
		40			2.34
		50			2.50
	150	40		0.07	2.82
		60			2.54
		80			2.10
x = YSC, y = GDP, N = 487	80	20	0.71	3.46	2.30
		30			2.33
		40			2.23
	100	30		4.10	2.87
		40			2.97
		50			2.72
	150	40		7.04	4.49
		60			3.88
		80			2.99
x = MRI, y = YSC, N = 428	80	20	0.78	0.05	2.82
		30			2.89
		40			2.56
	100	30		0.04	2.73
		40			2.94
		50			2.63
	150	40		0.04	3.65
		60			3.26
		80			2.90

Note : Mortality rate of infants (MRI); Gross domestic product (GDP); Life expectancy at birth (LEX); Average schooling years (YSC).

6. Conclusions

Two adaptive versions of PS, with- and without- replacement have been presented. Both versions are based on information observed in the process of sampling, and help the sampler to obtain a more efficient sample, leading to more accurate estimates. Compared to the conventional versions, these adaptive versions of PS require no additional information and only need time to analyze the initial sample to decide on the next steps in the sampling process.

APPS is easy to implement, more efficient than its conventional version, PPS, and sometimes more efficient than ΠPS. The simulations show that APPS is always more efficient than the PPS. In addition, increasing the sample size gives APPS the ability to update more units of the auxiliary variable, resulting

in increased efficiency. Besides these advantages, APPS has two weaknesses: the design is with-replacement and the sample units must be selected one by one in order.

On the other hand, AΠPS is a without-replacement design that must be implemented in two phases. In the first phase, an initial sample is selected and y is modeled on x and in the second phase, the final sample is selected based on the predicted y values. The relationship between x and y is modeled using the sample information based on Taylor expansion theory in the first phase of sampling. Next a proportional to size scheme is used in the second phase of sampling. The simulations confirm that AΠPS is an efficient and reliable design.

Acknowledgements

The authors would like to thank two referees and the associate editor for constructive and very insightful comments that helped them improve this article.

Appendix A

Proof of Result 1

To prove that $\sum_{k\in U} p_{ik} = 1,$ we have

$$\begin{aligned}\sum_{k\in U} p_{ik} &= \sum_{k\in s_{i-1}} p_{ik} + \sum_{k\in U\setminus s_{i-1}} p_{ik} = \sum_{k\in s_{i-1}} \frac{y_k\, t_{xs_i}}{t_x\, t_{ys_i}} + \sum_{k\in U\setminus s_{i-1}} \frac{x_k}{t_x} \\ &= \frac{t_{xs_i}}{t_x t_{ys_i}} t_{ys_i} + \frac{t_{xU\setminus s_i}}{t_x} = \frac{t_{xs_i}}{t_x} + \frac{t_{xU\setminus s_i}}{t_x} = 1.\end{aligned}$$

For unbiasedness of $\hat{t}_{\text{APPS}},$ we have

$$\hat{t}_{\text{APPS}} = \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

where $Z_i = y_{k_i} / p_{ik_i}$ and $Z_i \,|\, s_{i-1} \sim f_i(z),$ with

$$f_i(z) = \begin{cases} p_{ik} & \text{if } z = \dfrac{y_k}{p_{ik}}, k = 1, \ldots, N \\ 0 & \text{otherwise.}\end{cases}$$

Then

$$E(Z_i) = E\Big[E\big(Z_i \,|\, s_{i-1}\big)\Big] = E\left(\sum_{k=1}^{N} \frac{y_k}{p_{ik}} p_{ik}\right) = E(t_y) = t_y,$$

and therefore

$$E\left(\hat{t}_{\text{APPS}}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left[E\left(Z_i \middle| s_{i-1}\right)\right] = t_y.$$

For the variance, we have

$$V\left(\hat{t}_{\text{APPS}}\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V\left(Z_i\right) + \sum_{i=1}^{n}\sum_{j\neq i=1}^{n} C\left(Z_i, Z_j\right)\right],$$

where C is a symbol for covariance function. Then

$$\begin{aligned} V\left(\hat{t}_{\text{APPS}}\right) &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\left[EV\left(Z_i \middle| s_{i-1}\right) + VE\left(Z_i \middle| s_{i-1}\right)\right]\right. \\ &\quad \left. + 2\sum_{i=1}^{n}\sum_{j<i}^{n}\left(EC\left(Z_i, Z_j \middle| s_{i-1}\right) + C\left[E\left(Z_i \middle| s_{i-1}\right), E\left(Z_j \middle| s_{i-1}\right)\right]\right)\right\}. \quad \text{(A.1)} \end{aligned}$$

Now, we have

$$\begin{aligned} V\left[E\left(Z_i \middle| s_{i-1}\right)\right] &= V\left(t_y\right) = 0, \\ C\left[E\left(Z_i \middle| s_{i-1}\right), E\left(Z_j \middle| s_{i-1}\right)\right] &= C\left(t_y, Z_j\right) = 0, \end{aligned}$$

and

$$E\left[V\left(Z_i \middle| s_{i-1}\right)\right] = E\left[\sum_{k=1}^{N}\left(\frac{y_k}{p_{ik}} - t_y\right)^2 p_{ik}\right].$$

Also for $i > j$, Z_j is constant given s_{i-1}, and therefore $C\left(Z_i, Z_j \middle| s_{i-1}\right) = 0$. Finally, using (A.1) we have

$$V\left(\hat{t}_{\text{APPS}}\right) = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{N}\left(\frac{y_k}{p_{ik}} - t_y\right)^2 p_{ik}\right].$$

Moreover,

$$\hat{V}\left(\hat{t}_{\text{APPS}}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{ik_i}} - \hat{t}_{\text{APPS}}\right)^2,$$

is an unbiased estimator for $V\left(\hat{t}_{\text{APPS}}\right)$, because

$$\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{ik_i}} - \hat{t}_{\text{APPS}}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{ik_i}}\right)^2 - n\hat{t}^2_{\text{APPS}},$$

and

$$\begin{aligned} E\left(\frac{y_{k_i}}{p_{ik_i}}\right)^2 &= V\left(\frac{y_{k_i}}{p_{ik_i}}\right) + E^2\left(\frac{y_{k_i}}{p_{ik_i}}\right) \\ &= EV\left(\frac{y_{k_i}}{p_{ik_i}} \mid s_{i-1}\right) + VE\left(\frac{y_{k_i}}{p_{ik_i}} \mid s_{i-1}\right) + t_y^2 \\ &= E\left[\sum_{k=1}^{N}\left(\frac{y_k}{p_{ik}} - t_y\right)^2 p_{ik}\right] + 0 + t_y^2, \\ E\left(\hat{t}^2_{\text{APPS}}\right) &= V\left(\hat{t}_{\text{APPS}}\right) + E^2\left(\hat{t}_{\text{APPS}}\right) = V\left(\hat{t}_{\text{APPS}}\right) + t_y^2, \end{aligned}$$

then

$$\begin{aligned} E\left[\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{ik_i}} - \hat{t}_{\text{APPS}}\right)^2\right] &= E\left[\sum_{i=1}^{n}\sum_{k=1}^{N}\left(\frac{y_k}{p_{ik}} - t_y\right)^2 p_{ik}\right] + nt_y^2 - nV\left(\hat{t}_{\text{APPS}}\right) - nt_y^2 \\ &= n^2 V\left(\hat{t}_{\text{APPS}}\right) - nV\left(\hat{t}_{\text{APPS}}\right) = n(n-1)V\left(\hat{t}_{\text{APPS}}\right), \end{aligned}$$

and therefore $E\hat{V}\left(\hat{t}_{\text{APPS}}\right) = V\left(\hat{t}_{\text{APPS}}\right)$.

Appendix B

Proof of Result 2

To prove $\hat{t}_{A\Pi\text{PS}}$ is unbiased and driving its variance, we have

$$\begin{aligned} E\left(\hat{t}_{A\Pi\text{PS}}\right) &= E\left[E\left(\hat{t}_{A\Pi\text{PS}} \mid s_0\right)\right] = E\left(\sum_{k\in U^*}\frac{y_k}{\Pi_k^*}E\left(I_k^* \mid s_0\right) + \sum_{k\in s_0} y_k\right) \\ &= E\left(\sum_{k\in U^*} y_k + \sum_{k\in s_0} y_k\right) = E\left(t_y\right) = t_y, \end{aligned}$$

and

$$\begin{aligned} V\left(\hat{t}_{A\Pi\text{PS}}\right) &= EV\left(\hat{t}_{A\Pi\text{PS}} \mid s_0\right) + VE\left(\hat{t}_{A\Pi\text{PS}} \mid s_0\right) \\ &= E\left(\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right) + V\left(t_y\right) \\ &= E\left(\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right), \end{aligned}$$

respectively. Note that U^*, Π_k^* and $\Delta_{k\ell}^*$ are random based on the design.

Also, to prove that $\hat{V}\left(\hat{t}_{A\Pi\text{PS}}\right)$ is an unbiased estimator of the variance, we have

$$
\begin{aligned}
E\left[\hat{V}\left(\hat{t}_{A\Pi\text{PS}}\right)\right] &= E\left[E\left(\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\frac{\Delta_{k\ell}^*}{\Pi_{k\ell}^*}I_kI_\ell \mid s_0\right)\right] \\
&= E\left[\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\frac{\Delta_{k\ell}^*}{\Pi_{k\ell}^*}E\left(I_kI_\ell \mid s_0\right)\right] \\
&= E\left(\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right).
\end{aligned}
$$

Appendix C

Proof of Result 3

(i) This part of Result 3 will be easily proved by replacing α with $\left(N-n_0\right)/N$.

(ii) To prove that $E\left(\hat{t}_{A\Pi\text{PS}\alpha}\right)=t_y$, it is enough to note that when s_0 is a SRSWOR of size n_0, then $U\setminus s_0$ is a SRSWOR of size $N-n_0$. Therefore

$$
\begin{aligned}
E\left(\hat{t}_{A\Pi\text{PS}\alpha}\right) &= E\left[E\left(\hat{t}_{A\Pi\text{PS}\alpha} \mid s_0\right)\right] \\
&= NE\left(\frac{\alpha}{N-n_0}\sum_{k\in U\setminus s_0}\frac{y_k}{\Pi_k^*}E\left(I_k^* \mid s_0\right)+\frac{1-\alpha}{n_0}\sum_{k\in s_0}y_k\right) \\
&= N\left[\alpha E\left(\bar{y}_{U\setminus s_0}\right)+\left(1-\alpha\right)E\left(\bar{y}_{s_0}\right)\right]=N\bar{y}_U=t_y.
\end{aligned}
$$

(iii) For calculating the variance of $\hat{t}_{A\Pi\text{PS}\alpha}$ we have

$$
\begin{aligned}
V\left(\hat{t}_{A\Pi\text{PS}\alpha}\right) &= EV\left(\hat{t}_{A\Pi\text{PS}\alpha} \mid s_0\right)+VE\left(\hat{t}_{A\Pi\text{PS}\alpha} \mid s_0\right) \\
&= N^2E\left(\frac{\alpha^2}{\left(N-n_0\right)^2}\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right)+N^2V\left(\alpha\bar{y}_{U\setminus s_0}+\left(1-\alpha\right)\bar{y}_{s_0}\right) \\
&= N^2E\left[\frac{\alpha^2}{\left(N-n_0\right)^2}\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right]+N^2\left(1-\frac{\alpha}{1-f_0}\right)^2\left(\frac{1-f_0}{n_0}S_y^2\right).
\end{aligned}
$$

(iv) The proof of this part is the same as Result 2 and the fact that in SRSWOR, $E\left(s_{0y}^2\right)=S_y^2$.

(v) The optimal value α^* is obtained by calculating the derivative of $V\left(\hat{t}_{A\Pi\text{PS}\alpha}\right)$ with respect to α and solving the resulting equation

$$
\frac{dV\left(\hat{t}_{A\Pi\text{PS}\alpha}\right)}{d\alpha} = 0,
$$

which leads to

$$
\frac{\alpha}{\left(N-n_0\right)^2}E\left[\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right]-\frac{1}{1-f_0}\left(1-\frac{\alpha}{1-f_0}\right)\left(\frac{1-f_0}{n_0}S_y^2\right) = 0.
$$

Thus

$$\alpha = \left(1-f_0\right)\frac{\frac{1-f_0}{n_0}S_y^2}{E\left(\frac{1}{N^2}\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right)+\frac{1-f_0}{n_0}S_y^2}.$$

In order to show that the calculated α minimizes the variance, it is easy to show that the second derivative of $V\left(\hat{t}_{A\Pi PS\alpha}\right)$, given $f_0<1$ and $S_y^2>0$, is strictly positive:

$$\frac{d^2V\left(\hat{t}_{A\Pi PS\alpha}\right)}{d\alpha^2} = \frac{1}{\left(N-n_0\right)^2}E\left[\sum_{k\in U^*}\sum_{\ell\in U^*}\frac{y_k}{\Pi_k^*}\frac{y_\ell}{\Pi_\ell^*}\Delta_{k\ell}^*\right] + \frac{1}{\left(1-f_0\right)^2}\left(\frac{1-f_0}{n_0}S_y^2\right)>0.$$

References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.

Brewer, K.R.W., and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Harrison, D., and Rubinfeld, D. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81-102.

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.

Madow, W.G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20, 333-354.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Panahbehagh, B., Bruggemann, R., Parvardeh, A., Salehi, M.M. and Sabzalian, M.R. (2018). An unbalanced ranked set sampling to get more than one sample from each set. *Journal of Survey Statistics and Methodology*, 6, 256-305.

Seber, G.A.F., and Salehi, M.M. (2013). *Adaptive Sampling Designs*. Springer, Berlin, Heidelberg.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations*. Wiley, Hoboken.

Tillé, Y., and Matei, A. (2015). *Sampling: Survey Sampling*. R package version 2.7.

World Bank (2021). World bank open database (free and open access to global development data). https://scholar.harvard.edu/files/shleifer/files/data_friedman.xls. Accessed: 2021-11-10.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 38, No. 4, December 2022

Special Issue on Respondent Burden

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 39, No. 1, March 2023

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 50, No. 3, September/septembre 2022

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 50, No. 4, December/décembre 2022

Special Issue:50th anniversary of CJS/50e anniversaire de la RCS

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.