## Survey Methodology

## Survey Methodology <br> 49-2



Release date: January 3, 2024

## 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
© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2024
All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.
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

[^0]
# SURVEY METHODOLOGY <br> 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.

## MANAGEMENT BOARD

Chairman E. Rancourt
Past Chairmen C. Julien (2013-2018)
J. Kovar (2009-2013)

| Members | J.-F. Beaumont |
| :--- | :--- |
|  | D. Haziza |
|  | W. Yung |

D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

## EDITORIAL BOARD

Editor J.-F. Beaumont, Statistics Canada
Past Editor W. Yung (2016-2020)
M.A. Hidiroglou (2010-2015)
J. Kovar (2006-2009)
M.P. Singh (1975-2005)

## Associate Editors

- J.M. Brick, Westat Inc.
- K. McConville, Reed College
- P.J. Cantwell, U.S. Census Bureau
- G. Chauvet, École nationale de la statistique et de l'analyse de l'information
- S. Chen, University of Oklahoma Health Sciences Center
- I. Molina, Universidad Complutense de Madrid
- J. Opsomer, Westat Inc
- D. Pfeffermann, University of Southampton
- J.N.K. Rao, Carleton University
- J.L. Eltinge, U.S. Bureau of Labor Statistics
- L.-P. Rivest, Université Laval
- F.J. Scheuren, National Opinion Research Center
- P.L.d.N. Silva, Escola Nacional de Ciências Estatísticas
- P. Smith, University of Southampton
- D. Steel, University of Wollongong
- M. Torabi, University of Manitoba
- D. Toth, U.S. Bureau of Labor Statistics
- J. van den Brakel, Statistics Netherlands
- C. Wu, University of Waterloo
- W. Yung, Statistics Canada
- L.-C. Zhang, University of Southampton


## Assistant Editors C. Bocci, K. Bosa, C. Boulet, S. Matthews, C.O. Nambeu and Y. You, Statistics Canada

## 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 2, December 2023

## Contents

Waksberg invited paper series
Raymond L. Chambers
The missing information principle - A paradigm for analysis of messy sample survey data ..... 219
Special paper in memory of Professor Jean-Claude Deville
Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé
Jean-Claude Deville's contributions to survey theory and official statistics ..... 257
Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" Guillaume Chauvet. ..... 299
Marc Christine ..... 305
Françoise Dupont ..... 313
Camelia Goga and Anne Ruiz-Gazen
Jean-Claude Deville: Mathematics lover, high-flying researcher, and visionary ..... 321
Carl-Erik Särndal ..... 339
Invited papers presented at the $\mathbf{2 0 2 1}$ Colloque francophone sur les sondages
Louis-Paul Rivest
Statistical methods for sampling cross-classified populations under constraints ..... 347
Alina Matei, Paul A. Smith, Marc J.E. Smeets and Jonas Klingwort Targetted double control of burden in multiple surveys ..... 363
Estelle Medous, Camelia Goga, Anne Ruiz-Gazen, Jean-François Beaumont, Alain Dessertaine and Pauline Puech QR prediction for statistical data integration ..... 385
Vincent Loonis
Constructing all determinantal sampling designs. ..... 411
Regular papers
Jerzy Wieczorek Design-based conformal prediction ..... 443
J. Michael Brick and Jill M. DeMatteis
Sample designs and estimators for multimode surveys with face-to-face data collection ..... 475
Yilin Chen, Pengfei Li and Changbao Wu Dealing with undercoverage for non-probability survey samples ..... 497
Balgobin Nandram, Nathan B. Cruze and Andreea L. Erciulescu
Bayesian small area models under inequality constraints with benchmarking and double shrinkage ..... 517
Emily Berg
Small area prediction of general small area parameters for unit-level count data ..... 547
Yanzhe Li, Sander Scholtus and Arnout van Delden
A method for estimating the effect of classification errors on statistics for two domains ..... 569
Don Edwards, Piaomu Liu and Alexandria Delage
Model-based stratification of payment populations in Medicare integrity investigations ..... 607
Acknowledgements ..... 627
Announcements ..... 629
In other journals. ..... 631

## Waksberg Invited Paper Series

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

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2025 Waksberg Award.

This issue of Survey Methodology opens with the $23^{\text {th }}$ paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Jack Gambino (Chair), Maria Giovanna Ranalli, Denise Britz do Nascimento Silva and Kristen Olson for having selected Raymond L. Chambers as the author of 2023 Waksberg Award paper.

## 2023 Waksberg Invited Paper

## Author: Raymond L. Chambers

Ray Chambers is Honorary Professorial Fellow at the National Institute for Applied Statistics Research Australia, University of Wollongong, Australia. He was Leverhulme Professor of Social Statistics and Head of the Department of Social Statistics, University of Southampton as well as inaugural Director of the Southampton Statistical Sciences Institute. He is an elected member of the International Statistical Institute and a Fellow of the American Statistical Association. He was co-Editor in Chief of the International Statistical Review 2015-2019. He was President of the International Association of Survey Statisticians, 2011-2013 and International Representative on the Board of the American Statistical Association, 2011-2014. He has published widely on robust model-based methods for inference from complex survey data, and particularly where this complexity arises through integration of data from multiple sources. With Chris Skinner, he jointly edited Analysis of Survey Data, Wiley, 2003. He co-authored Maximum Likelihood Estimation for Sample Surveys, CRC Press, 2012, with David Steel, Alan Welsh and Suojin Wang, and An Introduction to Model-Based Survey Sampling with Applications, Oxford University Press, 2012, with Robert Clark. With Li-Chun Zhang, he jointly edited Analysis of Integrated Data, CRC Press, 2019. He was elected as a Fellow of the Academy of Social Sciences in Australia in 2021.

## Waksberg Award honorees and their invited papers since 2001

2024

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

2015 Robert Groves, "Towards a quality framework for blends of designed and organic data". Proceedings: Symposium 2016, Growth in Statistical Information: Challenges and Benefits.
2014 Constance Citro, "From multiple modes for surveys to multiple data sources for estimates". Survey Methodology, vol. 40, 2, 137-161.
2013 Ken Brewer, "Three controversies in the history of survey sampling". Survey Methodology, vol. 39, 2, 249262.

2012 Lars Lyberg, "Survey quality". Survey Methodology, vol. 38, 2, 107-130.
2011 Danny Pfeffermann, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". Survey Methodology, vol. 37, 2, 115-136.
2010 Ivan Fellegi, "The organisation of statistical methodology and methodological research in national statistical offices". Survey Methodology, vol. 36, 2, 123-130.
2009 Graham Kalton, "Methods for oversampling rare subpopulations in social surveys". Survey Methodology, vol. 35, 2, 125-141.
2008 Mary Thompson, "International surveys: Motives and methodologies". Survey Methodology, vol. 34, 2, 131141.

2007 Carl-Erik Särndal, "The calibration approach in survey theory and practice". Survey Methodology, vol. 33, 2, 99-119.
2006 Alastair Scott, "Population-based case control studies". Survey Methodology, vol. 32, 2, 123-132.
J.N.K. Rao, "Interplay between sample survey theory and practice: An appraisal". Survey Methodology, vol. 31, 2, 117-138.
2004 Norman Bradburn, "Understanding the question-answer process". Survey Methodology, vol. 30, 1, 5-15.
David Holt, "Methodological issues in the development and use of statistical indicators for international comparisons". Survey Methodology, vol. 29, 1, 5-17.
2002 Wayne Fuller, "Regression estimation for survey samples". Survey Methodology, vol. 28, 1, 5-23.
2001 Gad Nathan, "Telesurvey methodologies for household surveys - A review and some thoughts for the future". Survey Methodology, vol. 27, 1, 7-31.

# The missing information principle - A paradigm for analysis of messy sample survey data 

Raymond L. Chambers ${ }^{1}$


#### Abstract

Sample surveys, as a tool for policy development and evaluation and for scientific, social and economic research, have been employed for over a century. In that time, they have primarily served as tools for collecting data for enumerative purposes. Estimation of these characteristics has been typically based on weighting and repeated sampling, or design-based, inference. However, sample data have also been used for modelling the unobservable processes that gave rise to the finite population data. This type of use has been termed analytic, and often involves integrating the sample data with data from secondary sources.

Alternative approaches to inference in these situations, drawing inspiration from mainstream statistical modelling, have been strongly promoted. The principal focus of these alternatives has been on allowing for informative sampling. Modern survey sampling, though, is more focussed on situations where the sample data are in fact part of a more complex set of data sources all carrying relevant information about the process of interest. When an efficient modelling method such as maximum likelihood is preferred, the issue becomes one of how it should be modified to account for both complex sampling designs and multiple data sources. Here application of the Missing Information Principle provides a clear way forward. In this paper I review how this principle has been applied to resolve so-called "messy" data analysis issues in sampling. I also discuss a scenario that is a consequence of the rapid growth in auxiliary data sources for survey data analysis. This is where sampled records from one accessible source or register are linked to records from another less accessible source, with values of the response variable of interest drawn from this second source, and where a key output is small area estimates for the response variable for domains defined on the first source.


Key Words: Maximum likelihood; Combined data; Informative sampling; Nondeterministic linkage; Small area estimation.

## 1. Introduction

### 1.1 Descriptive and analytical inference with multiple data sources

Over the last century, sample surveys have become the primary method by which data are collected for analysis of social and economic processes, and empirical analysis of survey data is often the way theories are developed and investigated. While there has been a large increase in recent years in the data available from registers and administrative and business systems, these data are often limited. The number of variables for which information is gathered is often small, the definition of the variables may not be what is required, the data may be out-of-date, they may be available only in aggregate form and coverage of the population may be limited. A survey can be used to collect information on many variables at the individual person or business level, using relevant definitions in a consistent manner. This allows great flexibility in the estimates produced and the analyses possible.

How the sample is defined in a sample survey can vary considerably. Probability-based samples based on complex designs that reflect the heterogeneity and complex structures of the population of interest
represent one extreme. These designs may use auxiliary data available concerning the population at the time of selection. At the other extreme are convenience samples (Galloway, 2005) whose relationship to the population they are supposed to represent is problematic. In many cases external information, both on the process of sample selection as well as the comparability of the selected sample with the target population in terms of some known characteristics, is available to the analyst. Traditionally this auxiliary information has been used to improve sample design and estimation. With the increasing amount of data available in administrative and business databases, the sample survey now has a significant new role in supplementing such data sources so that they can be fully exploited. In particular, we now regularly face situations where data from multiple data sources need to be integrated for inference.

In order to address this issue of integrated inference, I will distinguish between statistical analysis aimed at estimating the value of an observable population quantity (e.g., the population average value of a variable) and analysis aimed at summarising the relationship between population variables in terms of a statistical model (e.g., a regression model). In the former case it is clear that the value of the population quantity becomes more and more "known" as the sample size increases, with the value known precisely (at least in theory) when the population is completely enumerated. This type of analysis is referred to variously as enumerative, predictive, descriptive or finite population inference. In the latter case the model is an abstract concept, corresponding to an idealisation of how the values of the different variables in the model relate to one another across the entire population. The "true" model is never known precisely, irrespective of how large a sample is used in the survey. This type of analysis is referred to as analytic inference. It is usually carried out by fitting the assumed statistical model to the survey data, with the nature and strength of the population relationship then summarised from the estimated values of the model parameters (e.g., the estimated regression coefficients).

Unfortunately, two quite distinct modes of inference exist for these two cases. If the target is a finite population quantity (e.g., a population average) the inferential framework is based on repeated sampling of the population, i.e., the population values for the variable of interest are held fixed. This is often referred to as design-based inference. On the other hand, if the target is a parameter of a statistical model for the population of interest (e.g., a regression parameter), then inference is typically model-based, i.e., it is with respect to potential population values that could have arisen under the true model. Little (2012) has described this state of affairs as "inferential schizophrenia" since the distinction between a target of inference that corresponds to a finite population quantity (e.g., the small area mean of a variable $Y$ ) and one that is the parameter of a model for the finite population values (e.g., the model expectation of $Y$ ) is very blurred. Model-based prediction theory (both frequentist and Bayesian) overcome this by explicitly allowing for the impact of the sample design in model-based inference.

In this paper I will adopt the frequentist interpretation of this framework, focussing on the use of the Missing Information Principle (the MIP) for model-based analytic inference in "messy" data situations where data from multiple sources are available. Other Waksberg Award papers have also discussed modelbased analytic inference from sample survey data (Scott, 2006; Rao, 2005; Pfeffermann, 2011), but none have zeroed in on the MIP and its use in the "messy" data that often arise in an integrated data context.

### 1.2 Weighting and complex sample designs

Most surveys use complex sample designs, reflecting and exploiting the heterogeneity of their target populations. The complexities introduced by the interaction of the survey design with these sources of heterogeneity are often difficult to handle using standard statistical methods. One particular issue that often arises is the role of sample selection probabilities in the analysis of the survey data. Complex sample designs typically result in unequal selection probabilities, one consequence of which is that the distributions observed in the sample can be very different from those in the population from which the sample was selected. There is confusion amongst survey users, as well as among non-survey statisticians (Gelman, 2007), about whether one should use selection probabilities in analysis, and if so, how this should be done in order to effectively "capture" the information about the sample design, and its effects, that they contain. A standard strategy is to weight for unequal selection probabilities when analysing survey data. The purpose of such weighting is to compensate for differences between the sample and population introduced by the sampling scheme (Pfeffermann, 1993). Using weighted summation of the sample data is attractive because it lends itself naturally to the estimation of linear parameters such as averages and totals, which are the primary objectives of many sample surveys, and also because linear estimators are very straightforward to build into survey estimation systems.

There are two main approaches to constructing weights. Often this is via the reciprocals of the selection probabilities. Such weights have a long history in descriptive surveys but may also be incorporated in model fitting, for example by pseudo-likelihood methods (Skinner, Holt and Smith, 1989). In these methods estimating equations that assume simple random sampling are modified to incorporate the survey weights. Second, weights may incorporate auxiliary information concerning the population, for example in poststratification and regression estimation (Bethlehem and Keller, 1987; Chambers, 1996). In this case a multiple regression model for predicting survey variables by auxiliary variables is used to define the weights, with a widely used example being Generalized Regression (GREG) estimation (Särndal, Swensson and Wretman, 1992). A variant of this approach is calibration estimation (Deville and Särndal, 1992; Särndal, 2007), where weights are constructed so that they are close to the inverses of the selection probabilities while at the same time allowing the weighted estimates to agree with selected population moments of key auxiliary variables.

The other principal concern with analysis involving sample survey data is how to make efficient use of external or auxiliary information when carrying out this analysis. Often, auxiliary information available about the population from a variety of sources (census, administrative registers, other surveys) is used to produce benchmarks that are used to constrain the survey estimates. Benchmarks are values of population characteristics or external estimates of these characteristics that are more reliable than unconstrained estimates derived from the sample data. Although many population benchmarks are often available, wellknown results on model over-fitting indicate that the number of benchmarks used in constraining survey weights should be limited to prevent instability of the resulting estimates. This leads to the conclusion that one should limit the number of calibration constraints imposed on the weights. The issue is particularly
important in multipurpose surveys, where the amount of related population information, and hence number of constraints, can be large (Bardsley and Chambers, 1984; Chambers, 1996). There are important practical advantages if the same weights are used for different estimates, and this can be achieved in model-based weighting if the same benchmarks are used for different survey variables. But this comes at a price. In particular, the resulting estimates can be inefficient, because the weights become very variable due to model over-parameterisation. Optimal methods for selecting the appropriate amount of external information to use in weighting have been discussed in the literature, but are limited to linear situations (Silva and Skinner, 1997; Clark and Chambers, 2008). Little is known about the extension to nonlinear situations or to the role of calibration information in analytic inference.

### 1.3 Analysis of complex survey data

Much attention has been devoted to the analysis of complex data over the last three decades. For example, in the 1990s the UK ESRC Research Programme on Analysis of Large and Complex Datasets focussed on the development of methods for the statistical modelling of the complex data collected in social science investigations. Since then, there has been a rise in interest in the theoretical foundations of inference based on sample survey data (Krieger and Pfeffermann, 1992; Breckling, Chambers, Dorfman, Tam and Welsh, 1994; Dorfman, Chambers and Wang, 2002; Little, 2003; Chambers and Skinner, 2003; Pfeffermann, 2011; Chambers, Steel, Wang and Welsh, 2012; Little, 2022). In particular, it is now accepted that statistical methods that assume that the distribution of the sample data and the distribution of the population data are identical generally lead to biased inference, since they take no account of either the complex sample design or the availability of auxiliary data.

There are three frameworks for frequentist inference that are generally used to deal with this problem.

- Pseudo-likelihood: This is a hybrid approach, with the unknown sufficient statistics in the population level likelihood estimating equations replaced by sample-weighted estimators (Kish and Frankel, 1974; Binder, 1983; Godambe and Thompson, 1986). The role of sample weights is therefore to adjust for differences between the sample distribution and the finite population distribution (Pfeffermann, 1993). Such weights usually have no connection with the variance structure of the data and so can lead to considerable inefficiency under the model. Furthermore, the weights used in practice are themselves adjusted, sometimes substantially, in order to integrate external population information (e.g., via calibration). However, weighted methods are very simple to implement and so are widely used.
- Sample likelihood: An explicit model for the distribution of the sample data, based on the use of Bayes theorem to integrate the population model and the sampling procedure, is used to develop a likelihood (Krieger and Pfeffermann, 1992; Pfeffermann, Krieger and Rinott, 1998; Pfeffermann and Sverchkov, 1999, 2003). This sample likelihood approach is typically more efficient than pseudo-likelihood. However, since it focuses on the distribution of the sample data as the basis
for inference, rather than on the population, it ignores non-sample information from auxiliary integrated data, making it less than fully efficient.
- Maximum likelihood: This is a fully efficient approach, where the auxiliary information and the sampling design are directly accounted for in the likelihood, itself defined by a joint model for the survey variables, the sampling and non-response processes and the auxiliary information. The basic methodology is set out in Breckling et al. (1994), while Chambers et al. (2012) provides a comprehensive development of maximum likelihood ideas in sample surveys. Adopting a maximum likelihood approach is conceptually appealing, but applying it to complex data requires care. Complex models are needed for the type of economic and social populations surveyed in practice, and the likelihood has to incorporate information about the sampling scheme and any auxiliary information about the population. A further difficulty arises in secondary analysis. Here the analyst does not always have access to information on how the data were obtained, and so suitable approximations need to be derived. Chambers, Dorfman and Wang (1998) consider likelihood-based analysis where sample design information is not provided, and in Section 5 of this paper I discuss the case where the analysis data set contains linked records but the analyst does not have access to the original data sets used in the linkage process.


### 1.4 Multiple surveys and auxiliary data

Extension of these approaches to multiple surveys and multiple auxiliary data sources is a relatively unexplored area of research, although the problem is well known, going back to the pioneering work of Patterson (1950) on composite estimation. Merkouris (2004) develops an integrated set of calibrated weights for use with the combined data from two independent surveys that measure the same variable of interest but use different types of auxiliary information. Elliott and Davis (2005) also consider the problem of weighting combined data on the same variable collected in two independent surveys, but allow one of these surveys to have a "higher quality" measurement process than the other, leading to a propensity-based adjustment to the original sample weights for the records from the second, "lower quality", survey. In contrast to both these approaches, which are aimed at re-weighting the combined data set, and hence concerned with marginal analysis of the same variable using this combined data set, Strauss, Carroll, Bortnick, Menkedick, and Schultz (2001) consider how one would go about modelling a joint distribution using the combined data from two independent surveys. In particular, these authors focus on the situation where each survey contributes a different variable to this joint distribution, but there exists a third, typically much smaller, survey with information on both that can be used to create an estimate of this joint distribution by combining the joint information in the small survey with the marginal information in the two larger surveys.

All of the references in the preceding paragraph address real issues that arise with integration of external auxiliary information with data from sample surveys. However, the approaches taken are problem specific and do not fit within a common inferential framework. When combined with efficient prediction, the MIP provides such a framework and can be used to develop solutions to a wide range of combining surveys
issues. In this paper I aim to show why this is the case in three important, but relatively straightforward, areas of application, and in doing so provide the reader with insight about a useful tool for tackling inference using what may be referred to as "messy" data.

### 1.5 Summary of the paper

This paper is meant to provide an overview of the MIP and its application, rather than a detailed methodological development. Consequently, I provide an informal definition of the MIP in Section 2 and in Section 3 I illustrate its use in combining data from two sources for the purpose of estimating a population regression relationship. In Section 4 I discuss the concept of informative sampling and use two simple examples to show how the MIP provides an appropriate framework for modelling sample data in this situation, while in Section 5 I show how the MIP can be used to suggest an efficient way of modelling linked data from two sources when the linkages can contain errors. At the close of this section I go on to show how these methods can then be used for small area estimation when records in different small area can be erroneously linked. Finally, in Section 6 I provide an overview and discussion of other applications where using the MIP has proved useful as well as some potential generalisations.

## 2. The missing information principle and its use

### 2.1 Messy data structures

What are "the data"? From a classical statistical perspective the answer to this question might be typically characterized as a transparent "window" on the population of interest. But the real world is messier. There are multiple sources of data with varying levels of aggregation, suggesting that a more accurate characterization is a distorted window on the target population, plus (perhaps) clearer windows on related populations. To illustrate this, consider some examples:

Example 1: Values of $Y$ from register A and values of $X$ from register B plus values of both variables from a sample of records taken from a linked version of the two registers. The aim is to use the sample data plus the data from the two registers to model the $Y-X$ relationship at register level (Imbens and Lancaster, 1994; Handcock, Rendall and Cheadle, 2005).

Example 2: Values of $Y$ plus auxiliary variables $X$ and $C$ from survey A plus values of the same variable $Y$ plus auxiliary variables $Z$ and $C$ from survey B. Estimates of the population total of $Y$ based on a combined sample are required (Merkouris, 2004).

Example 3: Values of "accurately measured" variables $Y$ and $X$ from a small survey A and values of a "rough approximation" to $X$ from a much larger survey B are available. This information is to be used to calculate small area estimates of $Y$ (Elliott and Davis, 2005).

Example 4: The analyst wishes to fit a model relating variables $Y, X$ and $Z$ at a national level. She has access to values of variables $Y$ and $Z$ from a large national survey plus values of correlated variable $X$ and the same variable $Z$ from another, distinct, large national survey plus values of $Y, X$ and $Z$ from a small, non-representative, third survey (Strauss et al., 2001).

### 2.2 Using the Missing Information Principle to combine data sources

The examples in the previous subsection illustrate pooling of multiple data sources, and all present problems for the analyst. However, they can be tackled by application of the Missing Information Principle or MIP. In particular, suppose that we can identify a model for the distribution of $Y$ in the target population, and this model is characterized by a parameter $\theta$. Suppose further that the data that we have for estimating this parameter are a mix of individual $Y$-values, values of other, related, variables, summary statistics, metadata (e.g., data definitions), paradata (e.g., information about how the data were obtained, sample weights, auxiliary data for the target population), related data from other surveys and other populations and so on. Opposed to this reality, the data we'd like to have for likelihood inference are data that define an ideal "rectangular" dataset containing representative data for the target population and related populations.

A naïve approach in this situation is to assume that the population model for $Y$ also applies to the sample values of this variable, and so the maximum likelihood estimate for $\theta$ can be calculated by maximising the sample contribution to the population likelihood. The corresponding "face value" maximum likelihood estimate for $\theta$ is generally incorrect since the sampling method underpinning the population model (typically simple random sampling) will not be the one underpinning the sample data. Furthermore, the available data includes data from other sources that also contain information about $\theta$. The appropriate likelihood should therefore also take account of this information in order to arrive at the "full information" maximum likelihood estimate for $\theta$.

The MIP provides a route for going from a simple likelihood analysis based on the ideal dataset to the correct likelihood analysis for the actual data that are available. In particular, it states that likelihood-based inference using a "messy" observed dataset $\mathbf{D}_{s}$ can be achieved by carrying out likelihood-based inference using a larger "ideal" dataset $\mathbf{D}_{U}$ with the likelihood estimating equations defined by $\mathbf{D}_{U}$ replaced by their expected values given $\mathbf{D}_{s}$. Note that it doesn't matter what $\mathbf{D}_{U}$ is here. The only requirements are that $\mathbf{D}_{s}$ (the data we have) is a subset of $\mathbf{D}_{U}$ (the data we would like to have), and that likelihood inference using $\mathbf{D}_{U}$ is straightforward. The MIP was first articulated by Orchard and Woodbury (1972) in the context of inference with missing data, and is closely related to the widely used EM algorithm (Dempster, Laird and Rubin, 1977). Its application to analysis of survey data was first described in Breckling et al. (1994). In our subsequent book, Chambers et al. (2012), we provide a comprehensive examination of how working within a MIP-based inferential framework leads to the maximum likelihood estimator (MLE) in a wide variety of messy data situations. In particular, the discussion in Sections 3 and 4 below summarises key aspects of this development by showing how the MIP can be used to fit simple population models to combined survey data.

In Section 5 I expand on this by showing how a difficult to compute MIP-based solution for one particular problem can be approximated by a much easier to compute MIP-based solution to a closely related problem.

In order to apply the MIP, we work with the population distribution of all the data available to the survey analyst. This can be illustrated by considering the simple scenario where there is a single survey sample with non-response, linked to an auxiliary variables dataset, with a single analysis variable $Y$. We use upper case to denote population quantities and lower case to denote sample quantities. Let $\mathbf{y}_{\text {resp }}$ denote the vector of survey respondents' values of $Y$ and let $\mathbf{r}_{s}$ denote the vector of response indicators for the sampled population units. We use $\mathbf{S}_{U}$ to denote the vector of sample inclusion indicators for the surveyed population. The matrix of population values of the auxiliary variables, which can include cluster or PSU indicators, is denoted $\mathbf{Z}_{U}$. Let $\kappa$ denote a vector of known population summary statistics. The available data are $\mathbf{D}_{s}=\left\{\mathbf{y}_{\text {resp }}, \mathbf{r}_{s}, \mathbf{S}_{U}, \mathbf{Z}_{U}, \kappa\right\}$. In addition to $\mathbf{y}_{\text {resp }}$, the quantities $\mathbf{r}_{s}, \mathbf{S}_{U}, \mathbf{Z}_{U}, \kappa$ potentially also contain information about $\theta$. In contrast, the ideal "rectangular" data are $\mathbf{D}_{U}=\left\{\mathbf{Y}_{U}, \mathbf{R}_{U}, \mathbf{S}_{U}, \mathbf{Z}_{U}, \kappa\right\}$ with a density $f\left(\mathbf{D}_{U} ; \Theta\right)$ that is straightforward to write down, and $\theta$ is then either a component of $\Theta$ or defined by a $1-1$ transformation of the components of $\Theta$. In either case if we can compute the MLE for $\Theta$, we can write down the MLE for $\theta$. Note that the likelihood generated by $\mathbf{D}_{U}$ is much easier to write down if $\mathbf{R}_{U}$ or $\mathbf{S}_{U}$ (or both) are ancillary for $\theta$ given $\mathbf{Z}_{U}$ and $\kappa$. That is, the distribution of $\mathbf{Y}_{U}$ and that of $\mathbf{R}_{U}$ and $\mathbf{S}_{U}$ are mutually independent given $\mathbf{Z}_{U}$ and $\kappa$.

There are two basic quantities used in likelihood inference. They are the score function, i.e., the derivative with respect to $\theta$ of the logarithm of the likelihood function, and the information function, i.e., the negative of the derivative of the score function with respect to $\theta$. The MLE is typically defined as a zero of the score function, while an estimate of the variance of MLE is the inverse of the information function evaluated at the MLE.

Let $\partial_{x} f$ denote a vector of first order partial derivatives with respect to $x$ and let $\partial_{x x} f$ denote the matrix of second order partial derivatives with respect to $x$. Then the MIP corresponds to two identities, proofs of which are set out in Lemma 2.1 of Chambers et al. (2012).

The score identity: Provided the ideal data $\mathbf{D}_{U}$ include the available data $\mathbf{D}_{s}$, the available data score $\mathrm{sc}_{s}$ for the parameter $\Theta$ of the distribution of $\mathbf{D}_{U}$ is the conditional expectation, given these data, of the ideal data score $\mathrm{sc}_{U}$ for $\Theta$, i.e.,

$$
\mathrm{sc}_{s}=E\left\{\partial_{\Theta} \log f\left(\mathbf{D}_{U} ; \Theta\right) \mid \mathbf{D}_{s}\right\}=E_{s}\left(\mathrm{sc}_{U}\right)
$$

The information identity: The available data information $\operatorname{info}_{s}$ for $\Theta$ is the negative of the matrix of partial derivatives for the components of the available data score $\mathrm{sc}_{s}$. This matrix can be written as the conditional expectation, given the available data, of the ideal data information $\operatorname{info}_{U}$ for $\Theta$ minus the corresponding conditional variance of the ideal data score $\mathrm{sc}_{U}$, i.e.,

$$
\operatorname{info}_{s}=E\left\{-\partial_{\Theta \Theta} \log f\left(\mathbf{D}_{U} ; \Theta\right) \mid \mathbf{D}_{s}\right\}-\operatorname{Var}\left\{\partial_{\Theta} \log f\left(\mathbf{D}_{U} ; \Theta\right) \mid \mathbf{D}_{s}\right\}=E_{s}\left(\operatorname{info}_{U}\right)-\operatorname{Var}_{s}\left(\mathrm{sc}_{U}\right)
$$

Note that the conditional expectations and variance in the score and information identities above are with respect to the distribution of the ideal data $\mathbf{D}_{U}$. Also, in many applications $\mathrm{sc}_{s}$ (and hence info ${ }_{s}$ ) turns out to be a function of $\mathbf{D}_{s}^{\text {obs }}=\left\{\mathbf{y}_{\text {resp }}, \mathbf{r}_{s}, \mathbf{S}_{U}, \kappa\right\}$ rather than of $\mathbf{D}_{s}=\left\{\mathbf{y}_{\text {resp }}, \mathbf{r}_{s}, \mathbf{S}_{U}, \mathbf{Z}_{U}, \kappa\right\}$. In such cases it is not difficult to see that the score and information identities still hold, but with $\mathbf{D}_{s}$ replaced by $\mathbf{D}_{s}^{\text {obs }}$. See Result 2 in Chambers et al. (1998).

The MIP is sometimes taken as referring to the information identity only, since the conditional variance term in this identity corresponds to the loss of information about $\Theta$ due to observing the available data $\mathbf{D}_{s}$ and not the ideal data $\mathbf{D}_{U}$. In this paper I take the MIP as being defined by both identities. However, it is the score identity that I find the most useful when faced with a messy data situation since it leads to parameter estimates for a population level model. The information identity can be used to obtain uncertainty estimates by inverting the observed information, but these estimates can be obtained in a variety of other ways including direct differentiation of $\mathrm{sc}_{s}$ as well as via bootstrap simulation of the fitted population level model.

In effect, it is the score identity that specifies the MLE based on the available data, while it is the information identity that shows us how much information about the parameter of interest we actually have given the available data. This is not dissimilar to the way $\mathrm{sc}_{U}$ is used to define a pseudo-likelihood estimator while the "observed information" about this estimator in the available data is given by the inverse of its estimated design variance.

The use of the score identity in the MIP to obtain the MLEs given the available data is an example of the application of what may be termed the "Prediction Principle", which is based on the fact that the minimum mean squared error predictor of the value of an unobserved random variable (say $Y$ ) given the value of another random variable (say $X$ ) is $E(Y \mid X)$. This principle underpins the model-based approach to sampling inference, starting with the seminal contributions of Royall (1970) and Royall (1976). In the score identity $Y$ corresponds to $\mathrm{sc}_{U}$ and $X$ corresponds to $\mathbf{D}_{s}$. So the best predictor of the solution to $\mathrm{sc}_{U}=0$ is the solution to $\mathrm{sc}_{s}=E\left\{\mathrm{sc}_{U} \mid \mathbf{D}_{s}\right\}=0$. Furthermore, since the ideal data score $\mathrm{sc}_{U}$ is a function of the sufficient statistics for $\Theta$ defined by $\mathbf{D}_{U}$, the score identity also tells us that when $\mathrm{sc}_{U}$ is a linear function of these sufficient statistics the best approximation to $\mathrm{sc}_{U}$ given the available data $\mathbf{D}_{s}$ is obtained by replacing these population sufficient statistics in $\mathrm{sc}_{U}$ by their expected values given $\mathbf{D}_{s}$.

## 3. Combining survey data and marginal population information Comparing the MIP with calibrated weighting

### 3.1 Calibration weighting in surveys

Likelihood analysis based on the MIP is a general and powerful way of incorporating external information into inference. However, its usefulness depends on our ability to construct models that capture the dependence between the survey variables and this external information at some "ideal" level and that
also allow straightforward conditioning on the available data. Calibrated weighting is a widely used method of survey estimation that also allows external information to be used, typically in the form of calibration constraints that ensure the weighted survey data are capable of exactly reproducing known finite population quantities. In most cases, the population quantities of interest are totals associated with auxiliary variables and so we consider constraints of the form $\mathbf{w}_{s}^{\prime} \mathbf{Z}_{s}=\mathbf{1}_{U}^{\prime} \mathbf{Z}_{U}$ where $\mathbf{Z}_{U}$ is the matrix of the $N$ population values of a set of survey variables with known population totals, $\mathbf{Z}_{s}$ is the corresponding matrix of the $n$ sample values, $\mathbf{w}_{s}$ is a vector of sample weights and $\mathbf{1}_{U}$ is a unitary $N$-vector.

Deville and Särndal (1992) introduced the idea of using calibrated sample weights $w_{i}$ that are closest to the expansion weights $\pi_{i}^{-1}$ where $\pi_{i}$ denotes the sample inclusion probability of population unit $i$. There are a variety of metrics for measuring closeness that can be used for this purpose, but the most popular is the chi square metric $Q=\left(\mathbf{w}_{s}-\boldsymbol{\pi}_{s}^{-1}\right)^{\prime} \Omega\left(\mathbf{w}_{s}-\boldsymbol{\pi}_{s}^{-1}\right)$ where $\boldsymbol{\pi}_{s}^{-1}$ is the vector of expansion weights and $\Omega$ is a positive definite matrix chosen by the analyst to reflect heteroskedasticity in the population values of the survey variables. Minimising $Q$ subject to calibration leads to weights

$$
\mathbf{w}_{s}^{\mathrm{cal}}=\boldsymbol{\pi}_{s}^{-1}+\Omega^{-1} \mathbf{Z}_{s}\left(\mathbf{Z}_{s}^{\prime} \Omega^{-1} \mathbf{Z}_{s}\right)^{-1}\left(\mathbf{Z}_{U}^{\prime} \mathbf{1}_{U}-\mathbf{Z}_{s}^{\prime} \boldsymbol{\pi}_{s}^{-1}\right)
$$

An alternative take on calibration is to view it as ensuring model-unbiased linear prediction of population totals (Valliant, Dorfman and Royall, 2000; Chambers and Clark, 2012). This is because weighting implies the use of a linear predictor $\mathbf{w}_{s}^{\prime} \mathbf{y}_{s}$ for the population total $\mathbf{1}_{U}^{\prime} \mathbf{Y}_{U}$, and if $\mathbf{Y}_{U}=\mathbf{Z}_{U} \beta+\mathbf{e}_{U}$ with $E\left(\mathbf{e}_{U} \mid \mathbf{Z}_{U}\right)=$ $\mathbf{0}_{U}$, then under non-informative sampling given $\mathbf{Z}_{U}$ any set of weights that are calibrated on $\mathbf{Z}_{U}$ will also define an unbiased predictor of $\mathbf{1}_{U}^{\prime} \mathbf{Y}_{U}$ under this linear model. So calibration is a good thing - provided the linear model assumption is valid.

### 3.2 Application to data from two sources

Consider the case where the population $U$ is such that the values $y_{i}$ and $x_{i}$ of two scalar variables, $Y$ and $X$ are stored on separate registers, each of size $N$. A simple random sample $s$ of $n$ units from one register is linked to the other via a unique common identifier, thus defining $n$ matched ( $y_{i}, x_{i}$ ) pairs. Our aim is to use these linked sample data, plus auxiliary information corresponding to the population averages of $Y$ and $X$ from each register, to estimate the parameters $\alpha, \beta$ and $\sigma^{2}$ that characterise the population linear regression model $y_{i}=\alpha+\beta x_{i}+\sigma e_{i}$ where the errors $e_{i}$ are distributed as independent and identically distributed (iid) Gaussian random variables with zero mean and unit variance.

The classical approach to fitting a population regression model like the one above given sample data is to use a pseudo-likelihood approach. See Kish and Frankel (1974), Binder (1983), Godambe and Thompson (1986) and Pfeffermann (1993). This is usually motivated as follows. Let $\mathbf{Y}_{U}$ and $\mathbf{X}_{U}$ denote the vectors of population values of $Y$ and $X$, with $f\left(\mathbf{Y}_{U} \mid \mathbf{X}_{U} ; \theta\right)$ denoting the conditional probability density of these population values. Then, if the pair $\left(\mathbf{Y}_{U}, \mathbf{X}_{U}\right)$ were to be observed, $\theta$ would be estimated as a solution to
$\mathrm{sc}_{U}=\partial_{\theta} \log f\left(\mathbf{Y}_{U} \mid \mathbf{X}_{U} ; \theta\right)=0$. But for any specified value of $\theta, \mathrm{sc}_{U}$ defines a finite population parameter (the "census score") that we can estimate using the sample data and the sample weights, say by $\mathrm{sc}_{w}$. The maximum pseudo-likelihood estimator (MPLE) of $\theta$ is then the solution to the estimating equation $\mathrm{sc}_{w}=0$.

Note that this approach does not specify how the sample weights should be constructed, only that $\mathrm{sc}_{w}=0$ defines a "design-consistent" estimator of $\mathrm{sc}_{U}$ for any permissible value of $\theta$. In particular, calibration weights can be used. For the case described above it is clear that there are three calibration constraints, defined by the population size $N$, the population mean of $X$ and the population mean of $Y$. Substituting $\mathbf{Z}_{U}=\left[\mathbf{1}_{U} \mathbf{Y}_{U} \mathbf{X}_{U}\right]$ and setting $\Omega$ equal to the identity matrix of order $N$, calibration weights that satisfy these three constraints are given by

$$
w_{s}^{\text {cal }}=\frac{N}{n} \mathbf{1}_{s}+N\left[\mathbf{1}_{s} \mathbf{y}_{s} \mathbf{x}_{s}\right]\left[\begin{array}{lll}
\mathbf{1}_{s}^{\prime} \mathbf{1}_{s} & \mathbf{y}_{s}^{\prime} \mathbf{1}_{s} & \mathbf{x}_{s}^{\prime} \mathbf{1}_{s} \\
\mathbf{1}_{s}^{\prime} \mathbf{y}_{s} & \mathbf{y}_{s}^{\prime} \mathbf{y}_{s} & \mathbf{x}_{s}^{\prime} \mathbf{y}_{s} \\
\mathbf{1}_{s}^{\prime} \mathbf{x}_{s} & \mathbf{y}_{s}^{\prime} \mathbf{x}_{s} & \mathbf{x}_{s}^{\prime} \mathbf{x}_{s}
\end{array}\right]^{-1}\left(\begin{array}{c}
0 \\
\bar{y}_{U}-\bar{y}_{s} \\
\bar{x}_{U}-\bar{x}_{s}
\end{array}\right) .
$$

Put $\bar{x}_{w s}^{\text {cal }}=N^{-1} \sum_{s} w_{i}^{\text {cal }} x_{i}$ and $\bar{y}_{w s}^{\text {cal }}=N^{-1} \sum_{s} w_{i}^{\text {cal }} y_{i}$. The corresponding calibrated MPLEs are then

$$
\begin{aligned}
& \hat{\beta}_{\text {CALmple }}=\left(\sum_{s} w_{i}^{\text {cal }} x_{i}\left(x_{i}-\bar{x}_{w s}^{\text {cal }}\right)\right)^{-1} \sum_{s} w_{i}^{\text {cal }} x_{i}\left(y_{i}-\bar{y}_{w s}^{\text {cal }}\right) \\
& \hat{\alpha}_{\text {CALmple }}=\bar{y}_{w s}^{\text {cal }}-\hat{\beta}_{\text {CALmple }} \bar{x}_{w s}^{\text {cal }} \\
& \hat{\sigma}_{\text {CALmple }}^{2}=N^{-1} \sum_{s} w_{i}^{\text {cal }}\left(y_{i}-\hat{\alpha}_{\text {CALmple }}-\hat{\beta}_{\text {CALmple }} x_{i}\right)^{2} .
\end{aligned}
$$

The alternative to this hybrid calibration-MPLE approach is to use the MIP. To start, note that the face value MLEs (i.e., MLEs based on an assumption of simple random sampling and no auxiliary information) for $\alpha, \beta$ and $\sigma^{2}$ are

$$
\begin{aligned}
\hat{\beta}_{\mathrm{FVmle}} & =\frac{\sum_{s}\left(x_{i}-\bar{x}_{s}\right)\left(y_{i}-\bar{y}_{s}\right)}{\sum_{s}\left(x_{i}-\bar{x}_{s}\right)^{2}} \\
\hat{\alpha}_{\mathrm{FVmle}} & =\bar{y}_{s}-\hat{\beta}_{\mathrm{FVmle}} \bar{x}_{s} \\
\hat{\sigma}_{\mathrm{FVmle}}^{2} & =n^{-1} \sum_{s}\left(y_{i}-\hat{\alpha}_{\mathrm{FV} m \mathrm{le}}-\hat{\beta}_{\mathrm{FVmle}} x_{i}\right)^{2} .
\end{aligned}
$$

However, there is extra information. In particular, we know the population means $\bar{y}_{U}$ and $\bar{x}_{U}$ of $Y$ and $X$. So the face value MLEs are no longer full information MLEs. The latter can be computed using the MIP. Given the Gaussian assumption, the components of the ideal (i.e., population) data score function are

$$
\begin{aligned}
& \mathrm{sc}_{1 U}=\frac{1}{\sigma^{2}} \sum_{U}\left(y_{i}-\alpha-\beta x_{i}\right) \\
& \mathrm{sc}_{2 U}=\frac{1}{\sigma^{2}} \sum_{U} x_{i}\left(y_{i}-\alpha-\beta x_{i}\right) \\
& \mathrm{sc}_{3 U}=\frac{N}{2 \sigma^{2}}+\frac{1}{2 \sigma^{4}} \sum_{U}\left(y_{i}-\alpha-\beta x_{i}\right)^{2} .
\end{aligned}
$$

Let a subscript of $U-s$ denote the non-sampled part of the population. The corresponding components of the available data score function are then obtained by replacing the components of the ideal data score function by their conditional expectations given the sample values of $Y$ and $X$ and the values $\bar{y}_{U-s}$ and $\bar{x}_{U-s}$ of the non-sample means of $Y$ and $X$. In order to do this, note that for non-sampled unit $i$ our Gaussian assumption for the error term in the population regression model, plus the fact of random sampling, allows us to write

$$
\binom{y_{i}}{\bar{y}_{U-s}} \left\lvert\, \mathbf{X}_{U-s} \sim N\left\{\binom{\alpha+\beta x_{i}}{\alpha+\beta \bar{x}_{U-s}},\left[\begin{array}{cc}
\sigma^{2} & (N-n)^{-1} \sigma^{2} \\
(N-n)^{-1} \sigma^{2} & (N-n)^{-1} \sigma^{2}
\end{array}\right]\right\}\right.
$$

where $\mathbf{X}_{U-s}$ denotes the non-sampled values of $X$. It is straightforward to see that the conditional distribution of $Y$ given $X, \bar{y}_{U-s}$ and $\bar{x}_{U-s}$ is then

$$
y_{i} \mid x_{i}, \bar{x}_{U-s}, \bar{y}_{U-s} \sim N\left\{\bar{y}_{U-s}+\beta\left(x_{i}-\bar{x}_{U-s}\right), \sigma^{2}\left(1-\frac{1}{N-n}\right)\right\}
$$

and so the MIP-based available data score function has components

$$
\begin{aligned}
& \mathrm{sc}_{1 s}=\frac{1}{\sigma^{2}}\left\{\sum_{s}\left(y_{i}-\alpha-\beta x_{i}\right)+(N-n)\left(\bar{y}_{U-s}-\alpha-\beta \bar{x}_{U-s}\right)\right\} \\
& \mathrm{sc}_{2 s}=\frac{1}{\sigma^{2}}\left\{\sum_{s} x_{i}\left(y_{i}-\alpha-\beta x_{i}\right)+(N-n) \bar{x}_{U-s}\left(\bar{y}_{U-s}-\alpha-\beta \bar{x}_{U-s}\right)\right\} \\
& \mathrm{sc}_{3 s}=-\frac{(n+1)}{2 \sigma^{2}}+\frac{1}{2 \sigma^{4}}\left\{\sum_{s}\left(y_{i}-\alpha-\beta x_{i}\right)^{2}+(N-n)\left(\bar{y}_{U-s}-\alpha-\beta \bar{x}_{U-s}\right)^{2}\right\} .
\end{aligned}
$$

The MIP-based MLEs are obtained by setting these score components to zero and solving for $\alpha, \beta$ and $\sigma^{2}$. The solutions are

$$
\begin{gathered}
\hat{\beta}_{\text {MIPmle }}=\frac{\sum_{s}\left(x_{i}-\bar{x}_{s}\right)\left(y_{i}-\bar{y}_{s}\right)+n \bar{x}_{s}\left(\bar{y}_{s}-\bar{y}_{U}\right)+(N-n) \bar{x}_{U-s}\left(\bar{y}_{U-s}-\bar{y}_{U}\right)}{\sum_{s}\left(x_{i}-\bar{x}_{s}\right)^{2}+n \bar{x}_{s}\left(\bar{x}_{s}-\bar{x}_{U}\right)+(N-n) \bar{x}_{U-s}\left(\bar{x}_{U-s}-\bar{x}_{U}\right)} \\
\hat{\alpha}_{\text {MIPmle }}=\bar{y}_{U}-\hat{\beta}_{\text {MIPmle }} \bar{x}_{U}
\end{gathered}
$$

and

$$
\hat{\sigma}_{\text {MIPmle }}^{2}=\frac{1}{n+1} \sum_{s}\left(y_{i}-\hat{\alpha}_{\text {MIPmle }}-\hat{\beta}_{\text {MIPmle }} x_{i}\right)^{2}+(N-n)\left(\bar{y}_{U-s}-\hat{\alpha}_{\text {MIPmle }}-\hat{\beta}_{\text {MIPmle }} \bar{x}_{U-s}\right)^{2} .
$$

These are just the weighted least squares (WLS) estimators of these parameters based on an extended sample consisting of the data values in $s$ (each with weight equal to one) plus an additional data value (with weight equal to $N-n$ ) defined by the known non-sample means $\bar{y}_{U-s}$ and $\bar{x}_{U-s}$. Standard WLS variance estimation methods can therefore be applied. Furthermore, these MIP-based MLEs depend only on the sample values of $Y$ and $X$ and on the population means of $Y$ and $X$ and not also on the individual values in the population vector $\mathbf{X}_{U}$ so they are also the available data MLEs when all one has is auxiliary summary information corresponding to the population means of $Y$ and $X$.

Related results are reported in Handcock, Rendall and Cheadle (2005) who tackle the problem by maximising the face value likelihood generated by the sample values of $Y$ and $X$ subject to the constraint $\bar{y}_{U}=\hat{\alpha}+\hat{\beta} \bar{x}_{U}$. This leads to the estimators

$$
\begin{gathered}
\hat{\beta}_{\mathrm{con}}=\frac{\sum_{s}\left(x_{i}-\bar{x}_{s}\right)\left(y_{i}-\bar{y}_{s}\right)+n\left(\bar{x}_{s}-\bar{x}_{U}\right)\left(\bar{y}_{s}-\bar{y}_{U}\right)}{\sum_{s}\left(x_{i}-\bar{x}_{r}\right)^{2}+n\left(\bar{x}_{s}-\bar{x}_{U}\right)^{2}} \\
\hat{\alpha}_{\mathrm{con}}=\bar{y}_{U}-\hat{\beta}_{\mathrm{con}} \bar{x}_{U}
\end{gathered}
$$

and

$$
\hat{\sigma}_{\mathrm{con}}^{2}=n^{-1} \sum_{s}\left(y_{i}-\hat{\alpha}_{\mathrm{con}}-\hat{\beta}_{\mathrm{con}} x_{i}\right)^{2} .
$$

In general, the differences between these constraint-based estimators and the MIP-based MLEs defined earlier will be small.

### 3.3 Imprecise benchmarks

So far, the population benchmarks $\bar{y}_{U}$ and $\bar{x}_{U}$ have been assumed to be precise. However, this is not always true. For example, they could be estimated from survey data themselves, albeit from surveys with much larger samples, and so may in fact have errors. This can arise, for example, if census coverage is incomplete, and so census outputs are adjusted for coverage error. It can also be the case that we have access to estimates derived from another larger survey rather than census values for these benchmarks. As long as the error or imprecision of such estimation is small, our MIP-based MLEs above are still valid. Asymptotically, if the benchmark estimates ( $\tilde{y}_{U}, \tilde{x}_{U}$ ) for ( $\bar{y}_{U}, \bar{x}_{U}$ ) satisfy $\tilde{y}_{U}=\bar{y}_{U}+o_{p}\left(n^{-1 / 2}\right)$ and $\tilde{x}_{U}=$ $\bar{x}_{U}+o_{p}\left(n^{-1 / 2}\right)$, and if they are used in place of $\left(\bar{y}_{U}, \bar{x}_{U}\right)$, then it is easily seen that the resulting estimators $\left(\tilde{\alpha}_{\text {MIPme }}, \tilde{\beta}_{\text {MIPmele }}\right)$ are asymptotically equivalent to ( $\hat{\alpha}_{\text {MIPmele }}, \hat{\beta}_{\text {MIPmle }}$ ) apart from a negligible error of $o_{p}\left(n^{-1 / 2}\right)$. However, this conclusion is not valid for $\tilde{\sigma}_{\text {miPme }}^{2}$ unless a generally higher order accuracy of $o_{p}\left(n^{1 / 2} / N\right)$ for the benchmark estimates is assumed.

Intuitively, one expects the extra information from knowing $\left(\bar{y}_{U}, \bar{x}_{U}\right)$ to contribute mainly to estimation of the intercept term $\alpha$ in the population regression model. To see that this is the case we write down the variances of $\hat{\beta}_{\text {MIPmle }}$ and $\hat{\alpha}_{\text {MIPmle }}$. This can be done by differentiating the available data score function components, changing signs and evaluating at these MLEs to get the observed information matrix for the regression model parameters. This matrix can then be inverted to get the (asymptotic) variances and covariances of these MLEs. Alternatively, exploiting their equivalence to a WLS fit, we can obtain the variances of $\hat{\alpha}_{\text {MIPmle }}$ and $\hat{\beta}_{\text {MIPmle }}$ directly. These are

$$
\begin{gathered}
\operatorname{Var}\left(\hat{\alpha}_{\text {MIPmle }}\right)=n^{-1} \sigma^{2}\left(\frac{\bar{x}_{s}^{(2)}-\left(1-n N^{-1}\right)\left(\bar{x}_{s}^{(2)}-\bar{x}_{U-s}^{2}\right)}{\bar{x}_{s}^{(2)}-\bar{x}_{U-s}^{2}+n N^{-1}\left(\bar{x}_{U-s}^{2}-\bar{x}_{U}^{2}\right)}\right) \\
\operatorname{Var}\left(\hat{\beta}_{\text {MIPme }}\right)=\frac{n^{-1} \sigma^{2}}{\bar{x}_{s}^{(2)}-\bar{x}_{U-s}^{2}+N n^{-1}\left(\bar{x}_{U-s}^{2}-\bar{x}_{U}^{2}\right)}=\frac{n^{-1} \sigma^{2}}{\bar{x}_{s}^{(2)}-\bar{x}_{s}^{2}+N^{-1}(N-n)\left(\bar{x}_{s}-\bar{x}_{U-s}\right)^{2}} .
\end{gathered}
$$

Here $\bar{x}_{s}^{(2)}$ is the mean of the squares of the sample $X$-values. It can be shown that $\operatorname{Var}\left(\hat{\beta}_{\text {MIPmle }}\right) \leq$ $\operatorname{Var}\left(\hat{\beta}_{\mathrm{FV} m \mathrm{e}}\right)$, with equality only if $\bar{x}_{s}=\bar{x}_{U-s}$. Similarly $\operatorname{Var}\left(\hat{\alpha}_{\text {MIPmle }}\right) \leq \operatorname{Var}\left(\hat{\alpha}_{\mathrm{FVmle}}\right)$, with equality only if $\bar{x}_{s}^{(2)}=\bar{x}_{s} \bar{x}_{U-s}$, which is extremely unlikely in practice. This confirms our intuition above.

### 3.4 Comparing the efficiency of MIP with that of calibration for data integration

How much more efficient is using a MIP-based approach to data integration compared with a calibrationbased approach? Some insight can be obtained from the results of a small model-based simulation study set out in Tables 3.1 and 3.2. Here population values were generated as $y_{i}=5+x_{i}+e_{i}$ with $x_{i}=e^{z_{i}}$ and $z_{i}$ and $e_{i}$ generated independently of one another as standard Gaussian. A total of 1,000 simulations were carried out for each scenario, corresponding to choice of $N, n$ and the degree of imprecision in the benchmarks. Sampling in Table 3.1 was carried out using simple random sampling without replacement (SRSWOR) and three levels of imprecision in the population benchmarks were examined - no error in the benchmarks, benchmarks subject to census-level error (benchmark equal to true value plus a random error with standard deviation equal to the actual marginal standard deviation multiplied by $N^{-1 / 2}$ ) and benchmarks subject to larger survey error (benchmark equal to true value plus a random error with standard deviation equal to the actual marginal standard deviation multiplied by $\left.(N / 5)^{-1 / 2}\right)$.

The values shown in Table 3.1 are relative efficiencies, defined as the ratio of the $5 \%$ trimmed RMSE of a reference estimator to the corresponding $5 \%$ trimmed RMSE of the estimator of interest, expressed as a percentage. Values over 100 therefore indicate superior relative efficiency for the alternative estimator. A trimmed RMSE was used to measure efficiency in order to avoid distortions caused by a small number of outlying error values generated in the simulations. The reference estimator in Table 3.1 is the face value MLE under SRSWOR. It is clear that the MIP-based MLEs perform well. In contrast, the MPLEs based on calibrated weights are consistently less efficient for all three parameters of interest, even when the benchmarks contain errors. It is only when the benchmark errors are relatively large that the efficiency of the MIP-based MLEs falls below that of the face value MLEs.

Table 3.2 shows the relative performances of the same estimators as in Table 3.1, but this time where an informative sampling method is used. In particular, the sample data here are selected with inclusion probabilities that are approximately proportional to their $Y$-values, and the reference estimation method is MPLE, with weights defined by inverse sample inclusion probabilities under probability proportional to $Y$ (PPY) sampling. In contrast the calibrated MPLE is based on calibrated versions of these sample weights while the MIP-based MLE is the same as in Table 3.1, i.e., it based on an assumption of SRSWOR. This allows one to investigate the degree to which incorporation of auxiliary population information helps protect against bias induced by misspecification of the sampling method. The gains from using the MIP-based MLE, even under a misspecified sampling method, are very clear. In contrast, the MPLE based on calibration weights is much less efficient, even though it is based on essentially unbiased sampling weights.

Table 3.1
Linear population model under SRSWOR and population benchmarks of varying quality.

|  |  | $\boldsymbol{N}=\mathbf{5 0 0}$ <br> $\boldsymbol{n}=\mathbf{2 0}$ | $\boldsymbol{N}=\mathbf{1 , 0 0 0}$ <br> $\boldsymbol{n}=\mathbf{5 0}$ | $\boldsymbol{N}=\mathbf{5 , 0 0 0}$ <br> $\boldsymbol{n}=\mathbf{2 0 0}$ |
| :--- | :--- | ---: | ---: | ---: |
|  |  |  | Benchmarks Known Precisely |  |
| $\alpha$ | CALmple | 103 | 127 | 143 |
|  | MIPmle | 134 | 145 | 150 |
| $\beta$ | CALmple | 81 | 90 | 96 |
|  | MIPmle | 106 | 102 | 101 |
| $\sigma^{2}$ | CALmple | 84 | 94 | 99 |
|  | MIPmle | 102 | 100 | 100 |
|  |  | Benchmarks Subject to Census-level Error |  |  |
| $\alpha$ | CALmple | 84 | 101 | 112 |
|  | MIPmle | 116 | 111 | 116 |
| $\beta$ | CALmple | 73 | 89 | 96 |
|  | MIPmle | 104 | 100 | 100 |
| $\sigma^{2}$ | CALmple | 78 | 88 | 97 |
|  | MIPmle | 103 | 101 | 100 |
|  |  | Benchmarks Subject to Larger Survey Error |  |  |
| $\alpha$ | CALmple | 64 | 71 | 75 |
|  | MIPmle | 86 | 80 | 78 |
| $\beta$ | CALmple | 71 | 84 | 93 |
|  | MIPmle | 100 | 95 | 100 |
| $\sigma^{2}$ | CALmple | 63 | 77 | 94 |
|  | MIPmle | 99 | 94 | 99 |

Notes: Values shown are per cent relative efficiencies with respect to $5 \%$ trimmed RMSE of the face value MLE under SRSWOR. CALmple denotes the MPLE based on calibrated weights, while MIPmle denotes the MIP-based MLE.
MIP $=$ Missing information principle; MLE $=$ Maximum likelihood estimator; MPLE $=$ Maximum pseudo-likelihood estimator; RMSE $=$ Root mean square error; SRSWOR $=$ Simple random sampling without replacement.

Table 3.2
Linear population model under PPY sampling and population benchmarks of varying quality.

|  |  | $\begin{array}{r} N=500 \\ n=20 \end{array}$ | $\begin{array}{r} N=\mathbf{1 , 0 0 0} \\ n=50 \end{array}$ | $\begin{array}{r} N=5,000 \\ n=200 \end{array}$ |
| :---: | :---: | :---: | :---: | :---: |
|  |  | Benchmarks Known Precisely |  |  |
| $\alpha$ | CALmple | 118 | 143 | 159 |
|  | MIPmle | 201 | 210 | 222 |
| $\beta$ | CALmple | 63 | 73 | 81 |
|  | MIPmle | 109 | 110 | 117 |
| $\sigma^{2}$ | CALmple | 78 | 89 | 91 |
|  | MIPmle | 106 | 106 | 111 |
|  |  | Benchmarks Subject to Census-level Error |  |  |
| $\alpha$ | CALmple | 98 | 120 | 135 |
|  | MIPmle | 136 | 139 | 152 |
| $\beta$ | CALmple | 65 | 70 | 77 |
|  | MIPmle | 107 | 112 | 121 |
| $\sigma^{2}$ | CALmple | 77 | 82 | 90 |
|  | MIPmle | 108 | 107 | 109 |
|  |  | Benchmarks Subject to Larger Survey Error |  |  |
| $\alpha$ | CALmple | 69 | 74 | 89 |
|  | MIPmle | 84 | 76 | 82 |
| $\beta$ | CALmple | 54 | 57 | 66 |
|  | MIPmle | 103 | 107 | 117 |
| $\sigma^{2}$ | CALmple | 62 | 71 | 87 |
|  | MIPmle | 99 | 101 | 102 |

Notes: Values shown are per cent relative efficiencies with respect to 5\% trimmed RMSE of expansion-weighted MPLE under PPY sampling. CALmple denotes the MPLE based on calibrated weights, while MIPmle denotes the MIP-based MLE.
MIP = Missing information principle; MLE = Maximum likelihood estimator; MPLE = Maximum pseudo-likelihood estimator; PPY = probability proportional to $Y ;$ RMSE $=$ Root mean square error .

The results set out in Tables 3.1 and 3.2 provide some evidence for claiming that parameter estimation based on application of the MIP is more efficient, and sometimes considerably more efficient, than parameter estimation based on maximum pseudo-likelihood, particularly when this approach is based on calibrated weights. As one might expect, the MIP-based estimate of $\alpha$ benefits most from the auxiliary information. However there are non-negligible gains for MIP-based estimates of $\beta$ and $\sigma^{2}$ as well.

Why is the use of calibration weights so inefficient here? One answer follows from taking a model-based perspective on calibration. Recollect that calibrated weighting implicitly assumes a linear model linking $Y$ and the variables defining the calibration constraints. But one of those constraints involves $Y$, implying an over-parameterised model. It is known (Bardsley and Chambers, 1984) that such models lead to highly variable weights and inefficient inference.

### 3.5 Another example of the use of the MIP for analysis based on integrated data sources

Integration of information from external sources when analysing survey data can arise in many different ways, and using the MIP as a general-purpose tool for these situations can be beneficial. For example, Merkouris (2004) describes a situation where independent generalized regression (GREG) estimators of the population total of a variable $Y$ based on data from multiple surveys need to be efficiently combined. The solution that is put forward in this paper is to form an efficiently weighted average of these different GREG estimators, where the efficient weights are based on a common auxiliary variable, say $C$, measured in the different surveys. But a MIP-based approach is also possible. To illustrate, suppose that there are just two surveys, say A and B, with survey A using calibrated weights based on constraints defined by auxiliary variables $X$ and $C$ and survey B using calibrated weights based on constraints defined by auxiliary variables $Z$ and $C$. From a model-based perspective, the ideal data set would be where all three auxiliaries are measured in both surveys, in which case the data from both surveys could be stacked and values of $Y$ fitted to a linear model with three covariates ( $X, C$ and $Z$ ). Parameters of this model can therefore be estimated using the MIP, with unknown values of $Z$ in survey A replaced by their conditional expectations and unknown values of $X$ in survey B replaced by their conditional expectations. The model-based regression estimator of the population total of $Y$ using the combined data from both surveys is then just $N$ times the fitted value of the three parameter model at the population means of $X, C$ and $Z$. This is very similar to data fusion (Raessler, 2004).

## 4. Using the MIP under informative sampling

### 4.1 What do we mean by saying that a method of sampling is informative?

In Section 3 above I assumed that the method of sampling was simple random, so that $\mathbf{S}_{U}$ and $\mathbf{Y}_{U}$ are conditionally independent given $\mathbf{Z}_{U}$. This allowed the sample label set $s$ to be treated as fixed since $\mathbf{S}_{U}$ is
then ancillary for the parameters of the ideal data model. If $\mathbf{S}_{U}$ is not ancillary, application of the MIP requires one to model the joint distribution of the ideal data and the outcome of the sampling process. This is specific to the method used to select the sample, and so it is impossible to provide general results. Instead, in this section I provide some insight into the use of the MIP under informative sampling by showing how two special cases of informative sampling impact on inference in the case of a very simple single parameter population distribution. These simple examples illustrate how using the MIP to integrate the information in $\mathbf{S}_{U}$ in these situations can substantially improve inference. Before doing this, however, it is useful to be a little clearer about what we mean when we say a method of sampling is informative.

Broadly speaking, sampling is informative if distributions of population and sample values of $Y$ are different (Pfeffermann, 1993). However, after conditioning on a population auxiliary, the two distributions can be the same. Sampling is non-informative (informative) for inference about the distribution of $Y$ given some information if the associated conditional probability of observing a particular value of $Y$ given a random population draw is equal (not equal) to the same conditional probability given the value of a random sample draw. That is, informative/non-informative status depends on what is being conditioned on. In particular, suppose that we have complete response, so $\mathbf{r}_{s}$ contains no information and our sample values of $Y$ are $\mathbf{y}_{s}$. This allows us to concentrate on the impact of conditioning on $\mathbf{S}_{U}$ and $\mathbf{Z}_{U}$. In the same way that the concepts of Missing Completely At Random, Missing At Random and Non-Ignorable Missingness are defined in the missing data literature (Rubin, 1976; Little and Rubin, 1987; Little, 2003), we can define Completely Non-Informative Sampling: The distribution of $\mathbf{Y}_{U}$ is independent of $\mathbf{S}_{U}$ and $\mathbf{Z}_{U}$, so the marginal distribution of $\mathbf{y}_{s}$ contains all relevant information for $\theta$.

Non-Informative Sampling Given $\mathbf{Z}_{U}$ : The distribution of $\mathbf{Y}_{U} \mid \mathbf{Z}_{U}$ is independent of that of $\mathbf{S}_{U} \mid \mathbf{Z}_{U}$ (i.e., $\mathbf{S}_{U}$ is ancillary for $\theta$ given $\mathbf{Z}_{U}$ ), so we have the same parameters for distributions of $\mathbf{y}_{s} \mid \mathbf{Z}_{U}$ and $\mathbf{Y}_{U} \mid \mathbf{Z}_{U}$ and the parameter of interest $\theta$ depends on the parameters of joint distribution of $\mathbf{y}_{s}$ and $\mathbf{Z}_{U}$ (i.e., the parameters of distribution of $\mathbf{y}_{s} \mid \mathbf{Z}_{U}$ and the parameters of the marginal distribution of $\mathbf{Z}_{U}$ ). Here we can ignore the sampling process in likelihood-based inference but cannot throw away $\mathbf{Z}_{U}$ information.

Informative Sampling: Here $\mathbf{Y}_{U}, \mathbf{S}_{U}$ and $\mathbf{Z}_{U}$ are jointly dependent and the parameter of interest $\theta$ can depend on all the parameters of the joint distribution of these quantities. An immediate consequence is that the conditional distributions of $\mathbf{Y}_{U}$ and $\mathbf{y}_{s}$ given the auxiliary information $\mathbf{Z}_{U}$ can be very different, and so inference about the parameters of $\mathbf{Y}_{U} \mid \mathbf{Z}_{U}$ cannot just focus on the likelihood generated by the conditional distribution of $\mathbf{y}_{s} \mid \mathbf{Z}_{U}$.

It should be clear from the above that informativeness of the sampling method depends on the auxiliary information available to the survey data analyst, and how much this information "explains" the outcome of the sampling process. For example, cluster and multi-stage sampling can be modelled when the auxiliary information includes indicators for the population groupings corresponding to sampling units at the different stages of sampling. A sampling method that is informative in one situation may not be informative in another. For example, even if the sampling mechanism is entirely determined by the auxiliary information,
this mechanism can be informative if we do not (or cannot) include it in our survey data, as often happens in secondary data analysis. Furthermore, even if $\mathbf{Z}_{U}$ is included in the available data, the sampling method can still be informative if it also depends on variables not included in $\mathbf{Z}_{U}$ that are correlated with those in $\mathbf{D}_{U}$. In this case, including the outcome $\mathbf{S}_{U}$ of the sampling mechanism as part of $\mathbf{D}_{U}$ requires us to specify the joint density of $\mathbf{Y}_{U}$ and $\mathbf{S}_{U}$ given $\mathbf{Z}_{U}$. From this we can determine the distribution of $\mathbf{y}_{s}$ and hence write down the likelihood for $\theta$. The traditional approach to likelihood inference under informative sampling achieves this by directly specifying the distribution of $\mathbf{Y}_{S}$ given $\mathbf{Z}_{U}$, where $S$ is a random subset of $U$ with distribution determined by the outcome $\mathbf{S}_{U}$. An alternative is to use the MIP to specify the score and information functions directly.

### 4.2 Applying the MIP to size-biased and cut-off sampling

In order to illustrate the use of the MIP in this context, first note that a commonly used model for informative sampling is where sample inclusion depends directly on $\mathbf{Y}_{U}$. Two ways this can happen are when inclusion probabilities are functions of $Y$ and where there is cut-off sampling on $Y$. The key ideas for dealing with these two situations are straightforwardly developed by assuming that the $N$ population values of $Y$ are independent and identically distributed draws from a single parameter exponential distribution with marginal density $f(y ; \theta)=\theta \exp (-\theta y)$, allowing one to obtain explicit results for both cases. In what follows I therefore make this assumption, with the target of inference then being $\mu=E(Y)=\theta^{-1}$.

First, suppose that the sample of $n$ units is selected using size-biased sampling with known inclusion probabilities,

$$
\pi_{i}=\frac{n\left(y_{i}+\alpha z_{i}\right)}{N\left(\bar{y}_{U}+\alpha \bar{z}_{U}\right)}
$$

but where $\alpha$ is unknown. Here $z_{i}$ is a auxiliary "size" value associated with population unit $i$ and there is complete response. It is easy to see that $\left(\pi_{i} N\right) \bar{y}_{U}+\left(\pi_{i} N \bar{z}_{U}-n z_{i}\right) \alpha=n y_{i}$, and so provided $n \geq 2$, values of $\bar{y}_{U}$ and $\alpha$ are deducible from the sample values of $Y$ and their known inclusion probabilities. Consequently the available data are the sample $Y$ values $\left\{y_{i} ; i \in s\right\}$ and $\bar{y}_{U}$. Applying the MIP, it immediately follows that the available data score for $\theta$ is

$$
\mathbf{s c}_{s}=E\left\{\sum_{U}\left(\theta^{-1}-y_{i}\right) \mid\left\{y_{i} ; i \in s\right\}, \bar{y}_{U}\right\}=N\left\{\theta^{-1}-E_{s}\left(\bar{y}_{U}\right)\right\}=N\left(\theta^{-1}-\bar{y}_{U}\right) .
$$

The MIP-based MLE for $\mu$ is then $\hat{\mu}_{\text {MIPmle }}=\bar{y}_{U}$, i.e., the ideal data MLE. Similarly, the available data information for $\theta$ is the population information for this parameter, $N \theta^{-2}$, and, since $\mu=\theta^{-1}$, the estimated variance of $\hat{\mu}_{\text {MIPmle }}$ is $N^{-1} \bar{y}_{U}^{2}$.

Next consider what happens under cut-off sampling. Again assume complete response and population values distributed as one parameter exponential, with mean $\mu$ the target of inference. But now suppose that
the vector of sample inclusion indicators is random, corresponding to all population units with $y_{i}>K$, for known $K$. Then

$$
E\left(\bar{y}_{U} \mid \mathbf{y}_{s}, K\right)=N^{-1}\left\{n \bar{y}_{s}+(N-n) E(Y \mid Y \leq K)\right\}=N^{-1}\left\{n \bar{y}_{s}+(N-n)\left(\theta^{-1}-K e^{-\theta K}\left(1-e^{-\theta K}\right)^{-1}\right)\right\}
$$

and so the available data score for $\theta$ is

$$
\mathrm{sc}_{s}=n\left(\theta^{-1}-\bar{y}_{s}\right)+(N-n)\left(K e^{-\theta K}\right)\left(1-e^{-\theta K}\right)^{-1} .
$$

There is no analytic solution to setting this score function to zero, but it is easy to obtain numerically. If we let $\hat{\theta}_{\text {MIPmle }}$ denote this solution then the MIP-based MLE of $\mu$ is

$$
\hat{\mu}_{\text {MIPmle }}=\bar{y}_{s}-(N-n) n^{-1}\left(K e^{-K \hat{\theta}_{\text {MIPme }}}\right)\left(1-e^{-K \hat{\theta}_{\text {MIPml }}}\right)^{-1} .
$$

Here it is easiest to obtain the available data information for $\theta$ by direct differentiation of the available data score for this parameter, i.e., as $\operatorname{info}_{s}=n \theta^{-2}+(N-n) K^{2} e^{-\theta K}\left(1-e^{-\theta K}\right)^{-2}$, and so a large sample estimate of the variance of $\hat{\mu}_{\text {MIPmle }}$ is $\hat{V}\left(\hat{\mu}_{\text {MIPmle }}\right)=\left(\left.\theta^{4} \mathrm{info}_{s}\right|_{\theta=\hat{\theta}_{\text {MPPme }}}\right)^{-1}$.

### 4.3 Maximum pseudo-likelihood under size-biased and cut-off sampling

Alternatively, we could adopt a maximum pseudo-likelihood approach for both sampling methods above. For the case of size-biased sampling the maximum pseudo-likelihood estimator (MPLE) of $\theta$, obtained as the zero of the sample weighted estimate of the ideal data score, is the inverse of the Hajek estimator for the population mean of $Y$. It immediately follows that the MPLE $\hat{\mu}_{\text {MPLE }}$ of $\mu$ is this Hajek estimator. Clearly $\hat{\mu}_{\text {MPLE }}$ is suboptimal - we know the inclusion probabilities $\pi_{i}$, so we know $\bar{y}_{U}$, which is the ideal data MLE. However, $\hat{\mu}_{\text {MpLE }}$ is approximately unbiased in large populations since

$$
E\left(\hat{\mu}_{\text {MPLE }}\right)=E\left\{E\left(\left.\frac{\sum_{U} I_{i} y_{i}\left(y_{i}+\alpha z_{i}\right)^{-1}}{\sum_{U} I_{i}\left(y_{i}+\alpha z_{i}\right)^{-1}} \right\rvert\, \mathbf{Y}_{U}\right)\right\} \approx E\left(\frac{n\left(\bar{y}_{U}+\alpha \bar{z}_{U}\right)^{-1} \bar{y}_{U}}{n\left(\bar{y}_{U}+\alpha \bar{z}_{U}\right)^{-1}}\right)=E\left(\bar{y}_{U}\right) .
$$

What about the case of cut-off sampling? Since pseudo-likelihood depends essentially on design consistency for its validity, and since this is turn requires that all population units have a non-zero chance of sample inclusion, it is clear that there is no MPLE for $\mu$ under cut-off sampling.

### 4.4 Maximum sample likelihood under size-biased and cut-off sampling

The other well-known approach to inference under informative sampling is to maximize the sample likelihood. This is a model-based methodology (Krieger and Pfeffermann, 1992; Pfeffermann, Krieger and Rinott, 1998; Pfeffermann and Sverchkov, 2003) motivated by inferential methods for size-biased sampling that approximate the probability density $f_{s}$ of the sample values making up $\mathbf{y}_{s}$ as a function of the probability density $f_{U}$ of the population values making up $\mathbf{Y}_{U}$ and the sampling weights. In particular, Bayes Theorem is used to obtain the probability density of a randomly chosen sample value $y_{i}$ as

$$
f_{s}\left(y_{i} ; \omega, \theta\right)=f_{U}\left(y_{i} \mid i \in s\right)=\frac{\operatorname{Pr}\left(i \in s \mid y_{i} ; \omega\right) f_{U}\left(y_{i} ; \theta\right)}{\operatorname{Pr}(i \in s ; \omega, \theta)}
$$

where $\omega$ is a nuisance parameter that characterizes the sample selection method. The estimator for the parameter $\theta$ of interest is then defined by maximising the "sample likelihood" for $\theta$,

$$
\operatorname{SL}\left(\theta, \omega ; \mathbf{y}_{s}\right)=\prod_{i \in s} \frac{\operatorname{Pr}\left(i \in s \mid y_{i} ; \omega\right) f_{U}\left(y_{i} ; \theta\right)}{\operatorname{Pr}(i \in s ; \omega, \theta)}
$$

as a function of $\theta$. Note that under this approach one needs to independently estimate the nuisance parameter $\omega$.

Applying the sample likelihood approach to estimation of the mean of an exponential distribution under the size-biased sampling scheme above, we first note the large sample approximation

$$
\log \left\{\operatorname{SL}\left(\theta ; \mathbf{y}_{s}\right)\right\} \propto \log \left\{\prod_{i \in s} \frac{\left(y_{i}+\alpha z_{i}\right)}{\left(\theta^{-1}+\alpha \bar{z}_{U}\right)} \theta \exp \left(-\theta y_{i}\right)\right\}=n \log (\theta)-n \log \left(\theta^{-1}+\alpha \bar{z}_{U}\right)-\theta n \bar{y}_{s}+C
$$

where $C$ does not depend on $\theta$. Differentiating with respect to $\theta$ leads to the sample likelihood score

$$
\mathrm{sc}_{\mathrm{SL}}=n \theta^{-1}\left(1+\left(1+\alpha \bar{z}_{U} \theta\right)^{-1}\right)-n \bar{y}_{s} .
$$

When $\alpha=0$ we have an explicit solution to $\mathrm{sc}_{\mathrm{SL}}=0$ given by $\theta=2 / \bar{y}_{s}$, implying a maximum sample likelihood estimator (MSLE) for $\mu$ of the form $\hat{\mu}_{\text {MSLE }}=\bar{y}_{s} / 2$. No explicit solution exists for $\mathrm{sc}_{\mathrm{sL}}=0$ when $\alpha>0$, so numerical methods need to be used. Also, when $\alpha=0$, it is straightforward to show that in large samples $E\left(\bar{y}_{s}\right) \approx 2 \mu$, so $\hat{\mu}_{\text {MSLE }}$ is approximately unbiased.

A simple MSLE for $\mu$ also exists under cut-off sampling. Here $\operatorname{Pr}(i \in s)=\operatorname{Pr}\left(y_{i}>K\right)=\exp (-\theta K)$, so, up to an additive constant, the $\log$ of the sample likelihood for $\theta$ under cut-off sampling is $n \log (\theta)-$ $\theta n\left(\bar{y}_{s}-K\right)$. It is easy to see that then the MSLE of $\mu$ is $\hat{\mu}_{\text {MSLE }}=\bar{y}_{s}-K$. This estimator is unbiased.

### 4.5 Comparison of MIPmle, MPLE and MSLE

Once again, small scale simulation results help put some perspective on how much efficiency is lost by using pseudo-likelihood or sample likelihood instead of full information likelihood based on application of the MIP with data from a single parameter exponential population. Tables 4.1 and 4.2 show bias and RMSE for size-biased sampling with $\alpha>0$ while Tables 4.3 and 4.4 show similar results for cut-off sampling. In all cases these results are based on 1,000 independent simulations. They show that a MIP-based MLE (MIPmle) is consistently preferable to estimation using maximum sample likelihood (MSLE) or maximum pseudo-likelihood (MPLE). Tables 4.1 and 4.2 also show that, as expected, maximum sample likelihood outperforms maximum pseudo-likelihood. These results are in line with what has been observed in other studies where using maximum pseudo-likelihood (by far the most prevalent method of parametric estimation with survey data) is inefficient (Dorfman, Chambers and Wang, 2002). Its only advantage would appear to be its simplicity and the widespread availability of software.

Table 4.1
Size-biased sampling from a single parameter exponential population of size $N=5,000$ with $\theta=1$ and with $n=100$.

| $\boldsymbol{\rho}$ | Bias |  |  |  | RMSE |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
|  | MIPmle | MPLE | MSLE | MIPmle | MPLE | MSLE |
| 0.05 | -0.0006 | 0.0040 | -0.0018 | 0.0145 | 0.1424 | 0.0950 |
| 0.25 | -0.0002 | 0.0070 | 0.0009 | 0.0138 | 0.1147 | 0.0827 |
| 0.50 | 0.0004 | 0.0094 | 0.0017 | 0.0139 | 0.1186 | 0.0813 |
| 0.75 | 0.0004 | 0.0028 | -0.0084 | 0.0140 | 0.1091 | 0.0763 |
| 0.95 | -0.0002 | 0.0074 | -0.0047 | 0.0145 | 0.1134 | 0.0713 |

Notes: Values of the auxiliary variable $Z$ were generated as a single parameter exponential with $\theta=1$ and with $\rho=\operatorname{Cor}(Y, Y+\alpha Z)$ where $\alpha=\sqrt{\rho^{-2}-1}$, so increasing (decreasing) correlation implied decreasing (increasing) $\alpha$.
MIPmle $=$ MIP-based MLE; MLE $=$ Maximum likelihood estimator; MPLE $=$ Maximum pseudo-likelihood estimator; MSLE $=$ Maximum sample likelihood estimator; RMSE = Root mean square error.

Table 4.2
Same scenario as in Table 4.1 except that $\rho=0.5$ and the impact of increasing $n$ is shown.

| $\boldsymbol{n}$ | Bias |  |  |  |  | RMSE |
| :--- | ---: | :---: | :---: | :---: | :---: | :---: |
|  | MIPmle | MPLE | MSLE | MIPmle | MPLE | MSLE |
| 10 | -0.0007 | 0.0611 | -0.0139 | 0.0138 | 0.3457 | 0.2539 |
| 30 | 0.0002 | 0.0255 | -0.0061 | 0.0147 | 0.1986 |  |
| 100 | 0.0004 | 0.0094 | 0.0017 | 0.0139 | 0.1186 |  |
| 300 | -0.0005 | -0.0090 | -0.0090 | 0.0142 | 0.0650 | 0.0511 |
| 900 | -0.0001 | -0.0344 | -0.0267 | 0.0144 | 0.0460 |  |
| Notes: | MIPmle = MIP-based MLE; MLE = Maximum likelihood estimator; MPLE $=$ Maximum pseudo-likelihood estimator; MSLE $=$ |  |  |  |  |  |
|  | Maximum sample likelihood estimator; RMSE = Root mean square error. |  |  |  |  |  |
|  |  |  |  |  |  |  |

Table 4.3
Cut-off sampling from a single parameter exponential distribution of $\operatorname{size} N=5,000$ with $\theta=1$ and with cutoff $K=2$.

| $\boldsymbol{\mu}$ | $\boldsymbol{E}(\boldsymbol{n})$ | Bias |  | RMSE |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  |  | MIPmle | MSLE | MIPmle | MSLE |
| 0.4343 | 50 | 0.0108 | -0.0016 | 0.0400 | 0.0646 |
| 0.5112 | 100 | 0.0016 | -0.0031 | 0.0197 | 0.0515 |
| 0.7109 | 300 | -0.0003 | 0.0020 | 0.0133 | 0.0393 |
| 1.0172 | 700 | -0.0008 | 0.000 | 0.0162 | 0.0387 |
| 1.6612 | 1,500 | -0.0002 | 0.0010 | 0.0240 | 0.0424 |
| Notes: The impact of increasing expected sample size is shown. |  |  |  |  |  |
|  | MIPmle = MIP-based MLE; MLE $=$ Maximum likelihood estimator; MSLE $=$ Maximum sample likelihood estimator; RMSE = Root |  |  |  |  |
|  | mean square error. |  |  |  |  |

Table 4.4
Same scenario as in Table 4.3 except that the cut-off $K$ changes, with $\theta$ modified to ensure expected sample sizes are as shown.

| $\boldsymbol{K}$ | $\boldsymbol{E}(\boldsymbol{n})$ | Bias |  | RMSE |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  |  | MIPmle | MSLE | MIPmle | MSLE |
| 5 | 50 | 0.0254 | -0.0020 | 0.0925 | 0.1522 |
| 4 | 100 | 0.0038 | -0.0025 | 0.0429 | 0.1045 |
| 3 | 300 | -0.0003 | -0.0041 | 0.0202 | 0.0607 |
| 2 | 700 | 0.0002 | -0.0005 | 0.0160 | 0.0381 |
| 1 | 1,500 | 0.0002 | -0.0004 | 0.0122 | 0.0210 |

Notes: The average value of $\theta$ is 1.0048 .
MIPmle $=$ MIP-based MLE; MLE $=$ Maximum likelihood estimator; MSLE $=$ Maximum sample likelihood estimator; RMSE $=$ Root mean square error.

### 4.6 Other examples of the use of the MIP under informative sampling

There are other situations where sampling is informative because the sample design itself is informative. For example, size stratification (i.e., stratification based on a size variable $Z$ correlated with $Y$ ) is informative when the size stratum boundaries are known, but the analyst does not have access to nonsampled population values of $Z$. This would often be the case in secondary analysis. For example, Dorfman, Chambers and Wang (2002) describe how the MIP can be used to approximate maximum likelihood estimates when $Z$ and $Y$ coincide. In a small-scale simulation study they show that using a MIP-based approach in this case leads to significant gains in efficiency compared with a maximum pseudo-likelihood approach using stratification weights. And even if $Z$ and $Y$ differ, it will usually be the case that they are highly correlated, in which case knowing that the non-sampled units within a stratum have values of $Z$ that lie between known bounds provides the analyst with information that can be used to modify inference about the stratum mean of $Y$ and hence its overall population mean.

It has already been noted that using a maximum pseudo-likelihood approach can be inefficient. However, a powerful argument for its use in the past has been that it is design consistent, and so robust to misspecification of the population distribution of $Y$. But this is usually relative to the use of a face value maximum likelihood approach, which ignores the information in the sample design and implicitly assumes simple random sampling. To illustrate, consider the following scenario, based on Examples 2 and 3 in Binder and Roberts (2003). Suppose that our assumed or working model for the population values of $Y$ is that they are independently and identically distributed as Gaussian with mean $\mu$ and with variance $\sigma^{2}$. The sample design is stratified sampling based on an auxiliary size variable $Z$. In particular, there are two strata, with stratum 1 (low values of $Z$ ) sampled disproportionately less than stratum 2 (high values of $Z$ ). In this case the face value MLE of $\mu$ is the unweighted sample mean $\bar{y}_{s}$, ignoring the disproportionate stratification. However, the MPLE is the stratified sample mean $\bar{y}_{s t}=N^{-1}\left(N_{1} \bar{y}_{s 1}+N_{2} \bar{y}_{s 2}\right)$, where $\bar{y}_{s j}$ is the sample mean in stratum $j$.

Now suppose that the working Gaussian model of a common mean and variance is misspecified, and in reality it is the conditional distribution of $Y$ given $Z$ that is Gaussian, with $E(Y \mid Z)=\beta Z$ and $\operatorname{Var}(Y \mid Z)=$ $\gamma^{2} Z$. We refer to this as the "true" model below. The target parameter $\mu$ (the marginal mean of $Y$ across the population) under this true model is then $\mu=E(E(Y \mid Z))=\beta E(Z)$. When $N$ is large it is reasonable to approximate it by $\beta E\left(\bar{z}_{U}\right)=\beta N^{-1}\left(N_{1} \eta_{1}+N_{2} \eta_{2}\right)$ where $\eta_{1}, \eta_{2}$ are the means of $Z$ in strata 1 and 2 respectively, with corresponding variances $\omega_{1}^{2}, \omega_{2}^{2}$. Consequently, under the true model, $E\left(\bar{y}_{s}\right)=\beta n^{-1}\left(n_{1} \eta_{1}+\right.$ $\left.n_{2} \eta_{2}\right)$ and $E\left(\bar{y}_{s t}\right)=\beta N^{-1}\left(N_{1} \eta_{1}+N_{2} \eta_{2}\right)$, while

$$
\operatorname{Var}\left(\bar{y}_{s}\right)=n^{-2}\left(n_{1}\left(\gamma^{2}+\omega_{1}^{2} \beta^{2}\right)+n_{2}\left(\gamma^{2}+\omega_{2}^{2} \beta^{2}\right)\right)
$$

and

$$
\operatorname{Var}\left(\bar{y}_{s t}\right)=N^{-2}\left(N_{1}^{2} n_{1}^{-1}\left(\gamma^{2}+\omega_{1}^{2} \beta^{2}\right)+N_{2}^{2} n_{2}^{-1}\left(\gamma^{2}+\omega_{2}^{2} \beta^{2}\right)\right)
$$

Typically $\omega_{1}^{2}<\omega_{2}^{2}$ so since $n_{2} / n<N_{2} / N$ we see that that $\bar{y}_{s t}$ is unbiased and has smaller variance than $\bar{y}_{s}$ under the true model.

However, the face value MLE $\bar{y}_{s}$ makes no use of the available data on the size variable $Z$. At a minimum this corresponds to the sample values of $Z$, which then allows the true regression model to be identified, and $\beta$ estimated by $\hat{\beta}=\bar{y}_{s} \bar{z}_{s}^{-1}$. Applying the MIP to this situation, and still using the working model to define the ideal data score, the MIP-based MLE is the solution of the available data score equation (here $r$ denotes the set of non-sampled population units)

$$
\sum_{s}\left(y_{i}-\mu\right)+\sum_{r}\left\{E\left(y_{i} \mid z_{i}\right)-\mu\right\}=0 .
$$

In order to proceed further, one needs to be more specific about what one knows about the non-sampled $Z$ values. If this is just the stratum population sizes then a reasonable assumption is that the expected value and variance of $Y$ vary between the strata, and $\bar{y}_{s t}$ is the MIP-based MLE. If in addition one knows the size stratum boundaries then the approach discussed following Table 4.4 can be adopted. However, suppose that one also knows the population average $\bar{z}_{U}$ of $Z$. It is easy to see that the MIP-based MLE is then $\hat{\mu}_{\text {MIPmle }}=\hat{\beta} \bar{z}_{U}$. Furthermore, this MIP-based MLE is unbiased under the true model since $E\left(\hat{\mu}_{\text {MIPme }}\right)=$ $\beta N^{-1}\left(N_{1} \eta_{1}+N_{2} \eta_{2}\right)$, with

$$
\operatorname{Var}\left(\hat{\mu}_{\text {MIIPmle }}\right)=n^{-1} \gamma^{2} E\left(\bar{z}_{U}^{2} / \bar{z}_{s}^{2}\right)+N^{-2} \beta^{2}\left(N_{1} \omega_{1}^{2}+N_{2} \omega_{2}^{2}\right) .
$$

Also, since $E\left(\bar{z}_{U}^{2} / \bar{z}_{s}^{2}\right)<1$ under the specified stratified sample design,

$$
\operatorname{Var}\left(\bar{y}_{s t}\right)-\operatorname{Var}\left(\hat{\mu}_{\mathrm{MIPmle}}\right)>\gamma^{2} \sum_{j=1}^{2} n_{j}^{-1}\left(\frac{N_{j}^{2}}{N^{2}}-\frac{n_{j}^{2}}{n^{2}}\right)+\beta^{2} \sum_{j=1}^{2} n_{j}^{-1} \omega_{j}^{2}\left(\frac{N_{j}^{2}}{N^{2}}\right)\left(1-\frac{n_{j}}{N_{j}}\right) .
$$

The expression on right hand side above will typically be positive. This is supported by the small-scale simulation results shown in Table 4.5. That is, although the face value MLE is biased and inefficient under the true model, the MIP-based MLE that takes the information on $Z$ into account is unbiased under this model and usually more efficient than the MPLE.

Table 4.5
Simulation results for $\mathbf{1 , 0 0 0}$ independent repetitions of $N=1,000, n=100, Z$ distributed as single parameter exponential with mean $4, \beta=1, \gamma=0.1$ and a sample design with two strata defined by values below/above the population mean of $Z$.


## 5. Modelling using non-deterministically linked data

### 5.1 Using the MIP when there is data linkage error

Data linkage is the joining of two or more administrative or survey datasets using statistical matching (ADRN, 2012). A key feature of many linkage applications is a clear separation between the linkage process and subsequent analysis of the linked data. Typically, this separation is for reasons of confidentiality, in the sense that the linking agencies often use confidential data in their record matching. See Harron, Goldstein and Dibben (2016). The linked data set that eventuates is then made available to analysts but the information used in the linking is not. This is a secondary analysis situation.

My focus here is on the bias in this analysis due to incorrect links, i.e., the bias that may arise if the linkage is not accurate enough (Lahiri and Larsen, 2005; Chambers, 2009; Kim and Chambers, 2012). Many of the linked data sets that are created are based on non-deterministic linking, where there is uncertainty about whether the data values in the linked record are actually for the same population unit. I also focus on the simplest scenario, where two population registers, denoted $\mathscr{\mathscr { y }}$ and $\mathscr{X}$, are linked, and the analyst, who has full access to the $\mathscr{X}$-register, is provided with a sample of the records from the linked register. In particular, the $\mathscr{\mathscr { y }}$-register contains values of a scalar random variable $Y$ and the $\mathscr{X}$-register contains values of vector random variable $X$.

Suppose that we are interested in modelling the conditional distribution of $Y$ given $X$. This is straightforward given a random sample of correctly linked $(Y, X)$ values. But, we do not have such a sample. Instead we have a sample of linked values $\left(Y^{*}, X\right)$ where $Y^{*}=Y$ if the linkage is correct, but possibly not if it is incorrect. I say "possibly" here because depending on the scale of measurement of $Y$, we can have $Y^{*}=Y$ even if the linkage is incorrect.

To proceed further I introduce a set of sandbox assumptions that simplify further analysis. They are unlikely to hold in practice, but serve to define a useful "working model" for inference. These are

- Both registers contain $N$ records, with no duplications, and linkage is 1-1 and complete. That is, all records in both registers are linkable, and no record in one register can be (eventually) linked to more than one record in the other register;
- There is a categorical "blocking" variable $B$ recorded on both registers, measured without error on both, and taking $Q$ distinct values $q=1,2, \ldots, Q$ such that all matching takes place within a block, i.e., records in both registers with the same value of $B$;
- The records on the linked data set are indexed in exactly the same way as they are indexed on the $\mathscr{X}$-register.

Suppose there are $M_{q}$ records in each register with $B=q$ (so $N=\sum_{q} M_{q}$ ). Our second assumption above then constrains linkage errors to only occur within "blocks". Let $\mathbf{y}_{q}$ and $\mathbf{y}_{q}^{*}$ denote the original and linked values of $Y$ in block $q$. Under 1-1 and complete linkage it immediately follows that $\mathbf{y}_{q}^{*}=\mathbf{A}_{q} \mathbf{y}_{q}$ where $\mathbf{A}_{q}$ is an unknown random permutation matrix of order $M_{q}$, i.e., entries of $\mathbf{A}_{q}$ are either zero or one,
with a value of one occurring just once in each row and column. Let $\mathbf{X}$ denote the matrix of $X$ values in the $\mathscr{X}$ register. Put $E\left(\mathbf{A}_{q} \mid \mathbf{X}\right)=\mathbf{T}_{q}$ and suppose that we have non-informative linkage given $\mathbf{X}$. That is, $\mathbf{A}_{q}$ is independent of $\mathbf{y}_{q}$ given $\mathbf{X}$ and so

$$
E\left(\mathbf{y}_{q}^{*} \mid \mathbf{X}\right)=\mathbf{T}_{q} E\left(\mathbf{y}_{q} \mid \mathbf{X} ; \theta\right)
$$

where $\theta$ denotes the vector of parameters of the conditional distribution of $Y$ given $\mathbf{X}$. These parameters are our primary target of inference.

An efficient linkage process should ensure that correct linkages within a block are more likely than incorrect linkages. We therefore impose the restrictive but practically useful assumption that linkage errors are exchangeable within a block, i.e., the probability of a record being correctly linked in block $q$ is $\lambda_{q}$ while the probability that it is incorrectly linked is $\eta_{q}$. We refer to this as an Exchangeable Linkage Errors or ELE model. Under 1-1 and complete linkage the ELE model implies $\lambda_{q}+\left(M_{q}-1\right) \eta_{q}=1$. It immediately follows that $\mathbf{T}_{q}=\left(\lambda_{q}-\eta_{q}\right) \mathbf{I}_{q}+\eta_{q} \mathbf{1}_{q} \mathbf{1}_{q}^{\prime}$, where $\mathbf{I}_{q}$ is the identity matrix of order $M_{q}$ and $\mathbf{1}_{q}$ is the unitary vector of length $M_{q}$.

However, we do not have access to the full linked register. Instead we have a random sample $s$ of $n$ records from this linked register. We extend the idea of non-informative linking by assuming that the random processes underpinning sample selection and linking are mutually independent and that these processes are both non-informative for the parameters of the distribution of $Y$ given $X$. This ensures that linkage of a sample to a register is stochastically equivalent to sampling from a completely linked register. This register can be partitioned into $Q$ blocks with block $q$ itself partitioned into $m_{q}$ sampled values followed by $M_{q}-m_{q}$ non-sampled (and hence unobserved) linked values. Following standard practice, we use subscripts of $s$ and $r$ to denote a partition into sampled and non-sampled values. Consequently $\mathbf{A}_{s q}$ is the matrix defined by those rows of $\mathbf{A}_{q}$ that correspond to sampled units, with $\mathbf{A}_{s s q}$ denoting those columns of $\mathbf{A}_{s q}$ that correspond to sampled units, and so on. We can then write the vector of sampled linked values in block $q$ as $\mathbf{y}_{s q}^{*}=\mathbf{A}_{s q} \mathbf{y}_{q}$, with $\mathbf{T}_{s q}=E\left(\mathbf{A}_{s q} \mid \mathbf{X}_{q}\right)=\left[\mathbf{T}_{s s q} \mathbf{T}_{s r q}\right]$ where $\mathbf{T}_{s s q}=\left(\lambda_{q}-\eta_{q}\right) \mathbf{I}_{s q}+\eta_{q} \mathbf{1}_{s q} \mathbf{1}_{s q}^{\prime}$ and $\mathbf{T}_{s r q}=$ $\eta_{q} \mathbf{1}_{s q} \mathbf{1}_{r q}^{\prime}$.

In order to carry out a maximum likelihood analysis in this situation we need to specify a model for the conditional distribution of $Y$ given $X$. We assume a simple linear model with Gaussian errors for this purpose, i.e., we put $\mathbf{y} \mid \mathbf{X} \sim N\left(\mathbf{X} \beta, \sigma^{2} \mathbf{I}\right)$ where $\mathbf{I}$ is the identity matrix of order $N$. However, our data consist of a sample of values of $Y$ from linked records. In order to apply the MIP in this case, we need to specify the conditional distribution of the correctly linked population values of $Y$ given these linked values. Now $E\left(\mathbf{y}_{q} \mid \mathbf{X}\right)=\mathbf{X}_{q} \beta$ and $\operatorname{Var}\left(\mathbf{y}_{q} \mid \mathbf{X}\right)=\sigma^{2} \mathbf{I}_{q}$, while under ELE $E\left(\mathbf{y}_{s q}^{*} \mid \mathbf{X}\right)=\mathbf{T}_{s q} \mathbf{X}_{q} \beta$ and $\operatorname{Var}\left(\mathbf{y}_{s q}^{*} \mid \mathbf{X}\right)=$ $\Sigma_{s q}^{*}=\sigma^{2} \mathbf{I}_{s q}+\mathbf{V}_{s q}$. Here $\mathbf{V}_{s q}$ is the sample component of $\mathbf{V}_{q}=\operatorname{Var}\left(\mathbf{A}_{q} \mathbf{X}_{q} \beta \mid \mathbf{X}\right)$ and represents the increased heterogeneity in $\mathbf{y}_{s q}^{*}$ caused by incorrect linkage. Chambers (2009) shows that $\mathbf{V}_{q}$ is well approximated by the diagonal matrix with $i^{\text {th }}$ diagonal term $\left(1-\lambda_{q}\right)\left(\lambda_{q}\left(f_{i}-\bar{f}_{q}\right)^{2}+\bar{f}_{q}^{(2)}-\left(\bar{f}_{q}\right)^{2}\right)$ where $f_{i}=\mathbf{x}_{i}^{\prime} \beta$ and $\bar{f}_{q}^{(2)}$ is the average of the $f_{i}^{2}$ in block $q$. Finally, we note that

$$
\operatorname{Cov}\left(\mathbf{y}_{q}, \mathbf{y}_{s q}^{*} \mid \mathbf{X}\right)=\operatorname{Cov}\left(\mathbf{y}_{q}, \mathbf{A}_{s q} \mathbf{y}_{q} \mid \mathbf{X}\right)=\operatorname{Cov}\left(\mathbf{y}_{q}, \mathbf{y}_{q} \mid \mathbf{X}\right) \mathbf{T}_{s q}^{\prime}=\sigma^{2} \mathbf{T}_{s q}^{\prime} .
$$

It is tempting to conclude from this (as I have done in the past) that the joint distribution of $\mathbf{y}_{q}$ and $\mathbf{y}_{s q}^{*}$ given $\mathbf{X}$ is then multivariate Gaussian with these moments. However, as pointed out by Zhang and Tuoto (2021), since the support of $\mathbf{y}_{s q}^{*}$ is just $\mathbf{y}_{q}$ this clearly cannot be true. However, if we are prepared to approximate this joint distribution by a Gaussian copula with the same first and second moments, then the MIP can be used to construct a corresponding approximation to the MLEs for the parameters $\beta$ and $\sigma^{2}$ of the conditional distribution of $\mathbf{y}_{q}$ given $\mathbf{y}_{s q}^{*}$. This argument turns out to be surprisingly fruitful.

Put $\mathbf{D}_{s q}=\mathbf{T}_{s q}^{\prime}\left(\sigma^{2} \mathbf{I}_{s q}+\mathbf{V}_{s q}\right)^{-1}$. Then

$$
E\left(\mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X}\right)=\mathbf{X}_{q} \beta+\sigma^{2} \mathbf{D}_{s q}\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right)=\mathbf{a}_{s q}
$$

and

$$
\operatorname{Var}\left(\mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X}\right)=\sigma^{2} \mathbf{I}_{q}-\sigma^{4} \mathbf{D}_{s q} \mathbf{T}_{s q}=\mathbf{B}_{s q} .
$$

That is, we can write $\mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X} \sim \mathbf{a}_{s q}+\mathbf{B}_{s q}^{1 / 2} \mathbf{g}_{q}$ where $\mathbf{g}_{q} \sim N\left(\mathbf{0}_{q}, \mathbf{I}_{q}\right)$. Next since $\mathbf{y}_{s q}^{*}=\mathbf{A}_{s q} \mathbf{y}_{q}$, the ideal data in block $q$ is the set $\left\{\mathbf{y}_{q}, \mathbf{X}_{q}\right\}$ while the available data is the set $\left\{\mathbf{y}_{s q}^{*}, \mathbf{X}_{q}\right\}$, when we treat $\mathbf{A}_{q}$, and hence $\mathbf{A}_{s q}$, as ancillary. In order to use the MIP we therefore first note that since $\mathbf{y}_{q} \sim N\left(\mathbf{X}_{q} \beta, \sigma^{2} \mathbf{I}_{q}\right)$, the score functions for $\beta$ and $\sigma^{2}$ based on the ideal data are

$$
\begin{gathered}
\mathrm{sc}_{U}(\beta)=\sigma^{-2} \sum_{q=1}^{Q} \mathbf{X}_{q}^{\prime}\left(\mathbf{y}_{q}-\mathbf{X}_{q} \beta\right) \\
\operatorname{sc}_{U}\left(\sigma_{e}^{2}\right)=\sigma^{-4} \sum_{q=1}^{Q}\left(\mathbf{y}_{q}-\mathbf{X}_{q} \beta\right)^{\prime}\left(\mathbf{y}_{q}-\mathbf{X}_{q} \beta\right)-N \sigma^{-2}=\sigma^{-4} \sum_{q=1}^{Q}\left\{\mathbf{y}_{q}^{\prime} \mathbf{y}_{q}-2 \beta^{\prime} \mathbf{X}_{q}^{\prime} \mathbf{y}_{q}+\beta^{\prime} \mathbf{X}_{q}^{\prime} \mathbf{X}_{q} \beta\right\}-N \sigma^{2}
\end{gathered}
$$

so the MIP-based score function for $\beta$ using the available data is

$$
\mathbf{s c}_{s}(\beta)=\sigma^{-2} \sum_{q=1}^{Q} \mathbf{X}_{q}^{\prime}\left(E\left(\mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X}\right)-\mathbf{X}_{q} \beta\right)=\sum_{q=1}^{Q} \mathbf{X}_{q}^{\prime} \mathbf{D}_{s q}\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right) .
$$

In order to define the corresponding MIP-based score for $\sigma^{2}$ we first note that

$$
E\left(\mathbf{y}_{q}^{\prime} \mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X}\right)=E\left(\left(\mathbf{a}_{s q}+\mathbf{B}_{s q}^{1 / 2} \mathbf{g}_{q}\right)^{\prime}\left(\mathbf{a}_{s q}+\mathbf{B}_{s q}^{1 / 2} \mathbf{g}_{q}\right)\right)=\mathbf{a}_{s q}^{\prime} \mathbf{a}_{s q}+\operatorname{tr}\left(\mathbf{B}_{s q}\right)
$$

and

$$
E\left(\beta^{\prime} \mathbf{X}_{q}^{\prime} \mathbf{y}_{q} \mid \mathbf{y}_{s q}^{*}, \mathbf{X}\right)=\beta^{\prime} \mathbf{X}_{q}^{\prime} \mathbf{a}_{s q} .
$$

This leads to a MIP-based available data score for $\sigma^{2}$ of the form

$$
\begin{aligned}
\mathrm{sc}_{s}\left(\sigma^{2}\right) & =\sigma^{-4} \sum_{q=1}^{Q}\left\{\left(\mathbf{a}_{s q}-\mathbf{X}_{q} \beta\right)^{\prime}\left(\mathbf{a}_{s q}-\mathbf{X}_{q} \beta\right)+\operatorname{tr}\left(\mathbf{B}_{s q}\right)\right\}-N \sigma^{2} \\
& =\sum_{q=1}^{Q}\left\{\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right)^{\prime} \mathbf{D}_{s q}^{\prime} \mathbf{D}_{s q}\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right)+\operatorname{tr}\left(\mathbf{B}_{s q}\right)\right\}-N \sigma^{2} .
\end{aligned}
$$

Formal representations for the resulting estimators of $\beta$ and $\sigma^{2}$ are obtained by setting these available data scores to zero and solving for these parameters. This leads to

$$
\hat{\beta}_{\mathrm{MIPmle}}=\left[\sum_{q=1}^{Q} \mathbf{X}_{q}^{\prime} \mathbf{D}_{s q} \mathbf{T}_{s q} \mathbf{X}_{q}\right]^{-1}\left[\sum_{q=1}^{Q} \mathbf{X}_{q}^{\prime} \mathbf{D}_{s q} \mathbf{y}_{s q}^{*}\right]
$$

and

$$
\hat{\sigma}_{\text {MIPmle }}^{2}=N^{-1} \sum_{q=1}^{Q}\left\{\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right)^{\prime} \mathbf{D}_{s q}^{\prime} \mathbf{D}_{s q}\left(\mathbf{y}_{s q}^{*}-\mathbf{T}_{s q} \mathbf{X}_{q} \beta\right)+\operatorname{tr}\left(\mathbf{B}_{s q}\right)\right\}
$$

Since $\mathbf{D}_{s q}$ (and hence $\mathbf{B}_{s q}$ ) is a function of $\beta$ and $\sigma^{2}$, the above estimators are computed iteratively. They also require one to know (or at least have a good estimate of) the correct linkage probabilities in each block. This issue is discussed in more detail in Chambers and Diniz da Silva (2019) and highlights the importance of the simultaneous release of paradata about the linking process when linked data are released for secondary analysis. An important practical point that also needs to be made here is that the block size $M_{q}$ will usually be very large, making computation of block-dimensioned quantities like $\mathbf{D}_{s q}$ and $\mathbf{B}_{s q}$ time consuming. So in the development below I introduce a further approximation, replacing $\mathbf{y}_{q}$ by $\mathbf{y}_{s q}$ in the ideal data.

### 5.2 Application to small area estimation using non-deterministically linked data

Probably the most common application of model-based ideas in survey sampling is small area estimation or SAE. That is, where the sampled population is partitioned into $D$ non-overlapping domains such that each domain is represented in the sample, but where the domain sample sizes are small, and sometimes even zero. It is standard to refer to these domains then as "small areas", where "small" is actually a reference to the domain sample size. See Rao and Molina (2015) for a comprehensive discussion of methods that have been proposed for estimation of domain-specific quantities in this situation, with the most common target being the domain average of a variable $Y$ measured on the sampled population units.

Here I focus on the special case where $Y$ is not measured directly on the sample but is obtained by linking the sample frame to another population register and then integrating the data from this register with the data directly obtained from the sampled units. This type of sample data acquisition is now increasingly common. Analysis variables in this integrated data set can exhibit increased heteroskedasticity (compared with an ideal data set where linkage is perfect) when records are incorrectly linked. This has the potential to lead to biased small area inference. See Briscolini, Di Consiglio, Liseo, Tancredi and Tuoto (2018).

In order to show how MIP-based ideas coupled with an ELE linkage errors specification can be used in SAE I assume that the population distribution of $Y$ given a set of covariates $X$ is adequately modelled at the unit level by a linear mixed model with Gaussian random effects of the form

$$
y_{j}=\mathbf{x}_{j}^{\prime} \beta+\mathbf{z}_{j}^{\prime} \mathbf{u}+e_{j}
$$

where $j$ indexes individual population units, $e_{j}$ denotes individual model error, $\mathbf{u}$ denotes a set of random area effects and $\mathbf{z}_{j}$ is a covariate characterising the impact of these area random effects on an individual population unit. The most common specification for this model in SAE is a random intercepts specification, where we associate a random effect $u_{i}$ with each area $i$ and $\mathbf{u}$ denotes the vector of these effects. In this case $\mathbf{z}_{j}$ is the vector that "picks out" the area in which unit $j$ is located. It is standard to assume noninformative sampling within each area, in which case the sample data on $Y$ can be written in matrix form as

$$
\mathbf{y}_{s}=\mathbf{X}_{s} \beta+\mathbf{Z}_{s} \mathbf{u}+\mathbf{e}_{s}
$$

where $\mathbf{e}_{s}$ is a $n$-vector of uncorrelated zero mean Gaussian random variables with common variance $\sigma_{e}^{2}$, $\mathbf{u}$ is a $D$-vector of uncorrelated zero mean Gaussian random variables with variance $\sigma_{u}^{2}$, and $\mathbf{e}_{s}$ and $\mathbf{u}$ are distributed independently.

With linked data spread across $Q$ blocks, however, we see

$$
\mathbf{y}_{s}^{*}=\left(\begin{array}{c}
\mathbf{y}_{s 1}^{*}=\mathbf{A}_{s 1} \mathbf{y}_{1} \\
\mathbf{y}_{s 2}^{*}=\mathbf{A}_{s 2} \mathbf{y}_{2} \\
\vdots \\
\mathbf{y}_{s Q}^{*}=\mathbf{A}_{s Q} \mathbf{y}_{Q}
\end{array}\right)=\mathbf{A}_{s} \mathbf{y}
$$

where $\mathbf{A}_{s}=\operatorname{diag}\left(\mathbf{A}_{s q}\right)$ and $\mathbf{y}$ denotes the vector of actual (but unknown) $Y$-values in the population. Put $\mathbf{T}_{s}=\operatorname{diag}\left(\mathbf{T}_{s q}\right)$ and $\mathbf{J}_{s}=\operatorname{diag}\left(\mathbf{J}_{s q}\right)$, where $\mathbf{T}_{s q}=\left(\lambda_{q}-\gamma_{q}\right) \mathbf{J}_{s q}+\gamma_{q} \mathbf{1}_{m_{q}} \mathbf{1}_{M_{q}}^{\prime}$ and $\mathbf{J}_{s q}=\left[\mathbf{I}_{m_{q}} \mid \mathbf{0}_{m_{q} \times\left(M_{q}-m_{q}\right)}\right]$. Then

$$
\begin{gathered}
E\left(\mathbf{y}_{s}^{*} \mid \mathbf{X}, \mathbf{Z}\right)=\left(\begin{array}{c}
\mathbf{T}_{s 1} \mathbf{X}_{1} \\
\mathbf{T}_{s 2} \mathbf{X}_{2} \\
\vdots \\
\mathbf{T}_{s Q} \mathbf{X}_{Q}
\end{array}\right) \beta=\mathbf{T}_{s} \mathbf{X} \beta \\
\operatorname{Var}\left(\mathbf{y}_{s}^{*} \mid \mathbf{X}, \mathbf{Z}\right)=\left[\left(\sigma_{e}^{2} \mathbf{I}_{s q}+\mathbf{V}_{s q}+\sigma_{u}^{2} \mathbf{K}_{s q}\right) 1(p=q)+\left(\sigma_{u}^{2} \mathbf{T}_{s p} \mathbf{Z}_{p} \mathbf{Z}_{q}^{\prime} \mathbf{T}_{s q}^{\prime}\right) 1(p \neq q)\right]=\Sigma_{s}^{*}
\end{gathered}
$$

and

$$
\operatorname{Cov}\left(\mathbf{y}_{s}, \mathbf{y}_{s}^{*} \mid \mathbf{X}, \mathbf{Z}\right)=\operatorname{Cov}\left(\mathbf{y}_{s}, \mathbf{A}_{s} \mathbf{y} \mid \mathbf{X}, \mathbf{Z}\right)=\left(\sigma_{e}^{2} \mathbf{J}_{s}+\sigma_{u}^{2} \mathbf{Z}_{s} \mathbf{Z}^{\prime}\right) \mathbf{T}_{s}^{\prime}=\mathbf{C}_{s} \mathbf{T}_{s}^{\prime}
$$

where $1(w)$ equals one if statement $w$ is true and is zero otherwise, and $\mathbf{K}_{s q}$ (see Samart and Chambers, 2014) represents the extra heterogeneity in $\mathbf{y}_{s}^{*}$ due to incorrect linkage of units in the same block but in different areas. Put $\operatorname{Var}\left(\mathbf{y}_{s} \mid \mathbf{X}, \mathbf{Z}\right)=\Sigma_{s}=\sigma_{e}^{2} \mathbf{I}_{s}+\sigma_{u}^{2} \mathbf{Z}_{s} \mathbf{Z}_{s}^{\prime}$. Making the same Gaussian copula assumption as before, we can then write $\mathbf{y}_{s} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z} \sim \mathbf{a}_{s}+\mathbf{B}_{s}^{1 / 2} \mathbf{g}_{s}$, with

$$
\mathbf{a}_{s}=E\left(\mathbf{y}_{s} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z}\right)=\mathbf{X}_{s} \beta+\mathbf{C}_{s} \mathbf{T}_{s}^{\prime}\left(\Sigma_{s}^{*}\right)^{-1}\left(\mathbf{y}_{s}^{*}-\mathbf{T} \mathbf{X} \beta\right)
$$

and

$$
\mathbf{B}_{s}=\operatorname{Var}\left(\mathbf{y}_{s} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z}\right)=\Sigma_{s}-\mathbf{C}_{s} \mathbf{T}_{s}^{\prime}\left(\Sigma_{s}^{*}\right)^{-1} \mathbf{T}_{s} \mathbf{C}_{s}^{\prime}
$$

Since $\mathbf{y}_{s}^{*}=\mathbf{A}_{s} \mathbf{y}$ the ideal data set underpinning the use of the MIP in this situation would normally include the population vector $\mathbf{y}$ and, as in the previous development, application of the MIP would then proceed using the properties of the conditional distribution $\mathbf{y} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z}$. However, this involves manipulating $N$-dimensioned quantities, which is usually impractical. I therefore introduce a further approximation that replaces $\mathbf{y}$ by $\mathbf{y}_{s}$ in the ideal data set. This has the immediate effect of replacing $N$-dimensioned quantities by $n$-dimensioned quantities in the score identity, which now depends on the first and second moments of the conditional distribution $\mathbf{y}_{s} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z}$ derived above. These can now be used to approximate the score functions for $\beta, \sigma_{u}^{2}$ and $\sigma_{e}^{2}$ based on the linked sample data, replacing the score functions for these parameters based on the ideal data set by their conditional expectations given the actual (i.e., linked) sample data. This process is the same as that already outlined for the simple regression case earlier so no details are provided here. Instead, I note that the popular maximum likelihood version of the Empirical Best Linear Predictor (EBLUP) of the mean $\bar{y}_{i}$ of $Y$ in area $i$ is of the form $\bar{y}_{i}^{\mathrm{EBL}}=\overline{\mathbf{x}}_{i}^{\prime} \hat{\beta}+\overline{\mathbf{z}}_{i}^{\prime} \hat{\mathbf{u}}$ where $\hat{\beta}$ is the MLE for $\beta$ and $\hat{\mathbf{u}}$ is the minimum MSE linear predictor for the vector of area effects $\mathbf{u}$ when $\beta, \sigma_{e}^{2}$ and $\sigma_{u}^{2}$ are replaced by their MLEs. However, given linked sample data, the minimum MSE predictor of $\mathbf{u}$ is its conditional expectation given these data. Under the Gaussian copula assumption this is

$$
E\left(\mathbf{u} \mid \mathbf{y}_{s}^{*}, \mathbf{X}, \mathbf{Z}\right)=\sigma_{u}^{2} \mathbf{Z}^{\prime} \mathbf{T}_{s}^{\prime}\left(\sigma_{s}^{*}\right)^{-1}\left(\mathbf{y}_{s}^{*}-\mathbf{T}_{s} \mathbf{X} \beta\right) .
$$

When MIP-based MLE approximations for $\beta, \sigma_{e}^{2}$ and $\sigma_{u}^{2}$ are substituted in this expression we obtain a linkage error corrected predictor of the random effects vector $\mathbf{u}$, which we denote by $\hat{\mathbf{u}}_{\text {MIP }}$. Combining this with the MIP-based MLE approximation $\hat{\beta}_{\text {MIP }}$ for $\beta$ one can then compute a MIP-based predictor for the mean of $Y$ in area $i$ as $\bar{y}_{i}^{\mathrm{MIP}}=\overline{\mathbf{x}}_{i}^{\prime} \hat{\beta}_{\mathrm{MIP}}+\overline{\mathbf{z}}_{i}^{\prime} \hat{\mathbf{u}}_{\mathrm{MIP}}$.

I can illustrate the gains from using this MIP-based approach to SAE based on linked data via a small simulation. This assumes linkage errors follow an ELE model with known parameters (see Chambers (2009) for how one deals with an ELE model with estimated parameters). The target population consisted of $D=$ 40 areas, with an average area population size of 500 , so $N=20,000$. A random intercepts model was used to generate the ideal data values of $Y$ for unit $j$ in area $i$ according to $y_{j}=100+5 x_{j}+u_{i}+e_{j}$, where the values of $x_{j}$ were generated as independent and identically distributed lognormal with a log scale mean of $\log (4.5)-0.5$ and a $\log$ scale variance of 0.5 . The area random effects $u_{i}$ were independently generated as Gaussian with mean zero and variance $\sigma_{u}^{2}=2$ while the individual random effects $e_{j}$ were independently generated as Gaussian with mean zero and variance $\sigma_{e}^{2}=7$. The actual linked values of $Y$ were then generated by independent repetitions of an ELE model within $Q=40$ blocks covering the population of interest. The blocks were defined independently of the small areas of interest, with $\lambda_{q}=1$ in blocks 1-10, $\lambda_{q}=0.95$ in blocks 11-20, $\lambda_{q}=0.9$ in blocks 21-30 and $\lambda_{q}=0.85$ in blocks 31-40. Blocks contained units from multiple small areas, with a block including units from an average of 5 small areas. As a consequence there was across area linkage error. Independent simple random samples were taken from each area, with area sample sizes ranging from 5 to 40 with an average of 25 , so $n=1,000$ and the linked sample values of $Y$ as well as the population values of $X$ were then used to fit the random intercepts model.

The above scenario was independently simulated 100 times. In each simulation estimates of the model parameters were calculated for the ideal case (no linkage error for sample values) and for the naive case (linkage error ignored), in both cases via REML using the function lmer in R (R Core Team, 2019). Estimates were also calculated using the MIP-based approach described above, using the naive estimates as starting values. Table 5.1 shows the average values and RMSEs over the 100 simulations, while the boxplots in Figure 5.1 show the distributions of these parameter estimates over the same simulations. Observe that the measurement error due to linkage error causes naive estimates of the fixed effects to be biased, reflecting the fact that linkage error shrinks slope parameters towards zero, with naive estimates of between area variation reduced and corresponding estimates of within area variation greatly increased. This is exactly what one expects. The MIP-based estimates do not suffer from these problems.

In addition, EBLUP-type estimates of the population average of $Y$ in each small area were calculated in each simulation, using the same parameter estimation methods (Ideal, Naive and MIP). For each small area and each simulation, the squared error and the absolute error of these EBLUP-type estimates were also computed. Figure 5.2 shows the boxplots of their corresponding mean squared error and mean absolute error values over the simulations for each area and for each parameter estimation method. These are denoted Area-MSE and Area-MAE respectively.

These results show that a method for fitting a mixed model that allows for linkage error can lead to significant improvement over a naive approach that ignores linkage error. This is consistent with results presented in Samart and Chambers (2014), Briscolini et al. (2018) and Salvati, Fabrizi, Ranalli and Chambers (2021). Of these, it is only the first paper where linkage errors are allowed between distinct small areas. Note that the approach described in that paper is not based on use of the MIP but on direct development of the likelihood function generated by the linked data followed by approximation of the relevant score functions. It also assumes balanced data (all block by area cells have sample) in order to obtain a formula for $\mathbf{K}_{s q}$. The same formula was used in the simulation reported here. In related research not presented here, Nicola Salvati and Enrico Fabrizi have empirically demonstrated that the small area estimates generated by the Samart-Chambers approach are less efficient than those generated by the MIP approach.

Table 5.1
Simulation results for 100 independent repetitions of the ELE linkage error scenario with $N=\mathbf{2 0 , 0 0 0}, n=1,000$ and $Q=40$ blocks.

| Parameter <br> (True value) | Average |  |  | RMSE |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| $\left(\begin{array}{ll}\text { Ideal } & \text { Naive }\end{array}\right.$ | MIP | Ideal | Naive | MIP |  |  |
| $\beta_{1}(5)$ | 99.993 | 101.115 | 99.981 | 0.296 | 1.206 | 0.358 |
| $\sigma_{u}^{2}(2)$ | 5.004 | 4.637 | 5.006 | 0.051 | 0.384 | 0.078 |
| $\sigma_{u}^{2}(7)$ | 2.078 | 1.889 | 2.050 | 0.560 | 0.656 | 0.649 |

[^1]Figure 5.1 Boxplots showing the distributions of parameter estimate values in the simulations of the ELE linkage error scenario.


Notes: Dotted horizontal line shows the true value of each parameter. ELE = Exchangeable linkage errors; MIP = Missing information principle.

Figure 5.2 Boxplots showing the distributions of area specific mean squared error (Area-MSE) and mean absolute error (Area-MAE) for the 40 areas in the ELE linkage error simulations.


Finally, it should be reiterated that application of the MIP approach with linked data is numerically intensive. This is because use of the ELE model to characterise linkage errors means that computations are effectively performed over all records in each of the blocks making up the $\mathscr{X}$-register. A practical implementation of the MIP algorithm that can handle large-scale linked population registers (which can contain millions of records) is an ongoing research project.

## 6. Discussion

I never knew Joseph Waksberg, but I certainly knew of him. Lohr (2021) describes a research career at the US Census Bureau and at Westat that made important contributions to many areas of survey methodology, including two that I subsequently became actively involved with - census coverage adjustment and calibration of survey weights. However, it was Waksberg's work on design and estimation using multiple frames that aligns most closely with the aims of this paper, since at their core these are about making maximum use of the information in combined data structures. In particular, his work shows us how to design sampling strategies that take advantage of this complexity to produce efficient estimates that relate to the population underpinning the combined data.

My aim in this paper has not been design but estimation, and in particular the use of the Missing Information Principle as a guide for defining parametric estimators when modelling messy data. In the context of a multiple frames scenario with random sampling of each frame, the ideal data are the values associated with the union of the distinct population frames, and the estimating equations for model parameters given multiple frame sample data are defined by replacing the sufficient statistics in the ideal data score function by their conditional expectations given these available data. When these sufficient statistics are linear in functions of the ideal data values this usually corresponds to replacing function values for individual units by their conditional expectations given the information derived from their (potentially multiple) frame memberships. This can be a complex specification process, requiring different models for different amounts of frame overlap.

There are many other messy data situations where application of the MIP leads to useful insights. Thus, Steel, Beh and Chambers (2004) report on how it can be used in likelihood-based inference with ecological data, i.e., where parameters of a joint distribution are of interest, but where the available data only provide information about marginal distributions. Here also having access to a very small sample taken from the joint distribution can have a very large impact in terms of improving the quality of inference. Another important application area where sampling is clearly informative is case-control sampling, see Prentice and Pyke (1979). Following the approach of Scott and Wild (1997) in this situation, Chambers and Wang (2008) use the MIP to develop MLEs for the parameters of a logistic regression model given case-control data. The simulation results they report show substantial improvement in efficiency over the standard approach for this model, which assumes simple random sampling in the fitting process and then discards the intercept
estimate. The MIP has been employed for efficient design as well, with Chipperfield, Barr and Steel (2018) using it in the context of efficient split questionnaire design.

The book Chambers et al. (2012) contains many more examples of application of the MIP as well as much more detailed developments of the results sketched out in this paper. In particular these show the information functions based on the available data. However, as I stated at the start of this paper, I believe that it is the score identity component of the MIP that is most useful since it shows how estimation should proceed. Uncertainty estimation given these estimates can be derived from the information function, but they can also be derived via more direct Taylor Series approximations or via numerically intensive methods such as bootstrapping.

When one views the score identity in the MIP from a more abstract perspective, it is clear that it is a special case of estimation based on the conditional expectation of a convenient estimating function. Consequently, if one generalises from the standard frequentist likelihood focus of this paper, it is interesting to note that an equivalent formulation of the score identity has been developed for estimating functions based on quasi-likelihood (Lin, Steel and Chambers, 2004). That is, there is scope for extension of the use of a MIP-based approach to estimation in messy data situations, for example those based on nonparametric likelihood approaches like empirical likelihood. Whether this leads to further insights remains to be seen, however. In any case such nonparametric extensions will also require methods for calculating the nonparametric equivalent of the conditional expectation operator reflecting the available information, perhaps via constrained parametric simulation. It will be interesting to see whether the development of these generalisations of the MIP will then allow it to accommodate the types of "large" machine learning models that are becoming more common.

I am very grateful to the Waksberg Award Committee for giving me this opportunity to prepare this paper for Survey Methodology. Hopefully it will encourage other statisticians working with messy data to investigate whether the MIP (and its potential generalisations) can be a useful tool for making inferences based on these data.

Finally, I would like to dedicate the preparation of this paper to my three friends and former colleagues from the University of Southampton, Fred Smith, Chris Skinner and Tim Holt, who have all now sadly passed away. Without the impetus of their groundbreaking book, Skinner, Holt and Smith (1989), and their insight and support during my years at Southampton, many of my personal contributions reported in this paper would not have been possible.

## References

ADRN (2012). The UK Administrative Data Research Network. Improving Access for Research and Policy. Report from the Administrative Data Taskforce.

Bardsley, P., and Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. Applied Statistics, 33, 290-299.

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. Journal of Official Statistics, 3, 141-153.

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

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. Analysis of Survey Data, (Eds., R.L. Chambers and C.J. Skinner). Chichester: Wiley.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from survey data. International Statistical Review, 62, 349-363.

Briscolini, D., Di Consiglio, L., Liseo, B., Tancredi, A. and Tuoto, T. (2018). New methods for small area estimation with linkage uncertainty. International Journal of Approximate Reasoning, 94, 30-42.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. Journal of Official Statistics, 12, 3-32.

Chambers, R.L. (2009). Regression analysis of probability-linked data. Statisphere, 4. Available from https://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=FL1356824.

Chambers, R.L., and Clark, R.G. (2012). An Introduction to Model-Based Survey Sampling with Applications. Oxford: Oxford University Press.

Chambers, R., and Diniz da Silva, A. (2019). Improved secondary analysis for linked data: A framework and an illustration. Journal of the Royal Statistical Society, Series A, 183, 37-59.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. Journal of the Royal Statistical Society, Series B, 60, 397-412.

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

Chambers, R.L., Steel, D.G., Wang, S. and Welsh, A.H. (2012). Maximum Likelihood Estimation for Sample Surveys. CRC Press: Boca Raton, Florida.

Chambers, R., and Wang, S. (2008). Maximum likelihood logistic regression with auxiliary information. Working Paper 12-08, Centre for Statistical and Survey Methodology, University of Wollongong.

Chipperfield, J.O., Barr, M.L. and Steel, D.G. (2018). Split questionnaire designs: Collecting only the data you need through MCAR and MAR designs. Journal of Applied Statistics, 45, 1465-1475. DOI 10.1080/02664763.2017.1375085.

Clark, R.G., and Chambers, R.L. (2008). Adaptive calibration for prediction of finite population totals. Survey Methodology, 34, 2, 163-172. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10757-eng.pdf.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-37.

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

Dorfman, A., Chambers, R. and Wang, S. (2002). Are survey weights necessary? The maximum likelihood approach to sample survey inference. Proceedings of the $162^{\text {nd }}$ Annual Meeting of the American Statistical Association, New York, August 11-15.

Elliott, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. Applied Statistics, 54, 595-609.

Galloway, A. (2005). Non-probability sampling. Encyclopedia of Social Measurement, 2, (Ed., K. KempfLeonard), Elsevier, 859-864.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. Statistical Science, 22, 153164.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of super populations and survey population: Their relationship and estimation. International Statistical Review, 54, 37-59.

Handcock, M., Rendall, M. and Cheadle, J. (2005). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. Sociological Methodology, 35, 291-334.

Harron, K., Goldstein, H. and Dibben, C. (Editors) (2016). Methodological Developments in Data Linkage. Chichester: Wiley.

Imbens, G.W., and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. Review of Economic Studies, 61, 655-680.

Kim, G., and Chambers, R. (2012). Regression analysis under incomplete linkage. Computational Statistics and Data Analysis, 56, 2756-2770.

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

Krieger, A.M., and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. Survey Methodology, 18, 2, 225-239. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14484-eng.pdf.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. Journal of the American Statistical Association, 100, 222-230.

Lin, Y.-X., Steel, D. and Chambers, R.L. (2004). Restricted quasi-score estimating functions for sample survey data. Stochastic Methods and Their Applications (Journal of Applied Probability, 41A, (Eds., J. Gani and E. Seneta). Sheffield: Applied Probability Trust.

Little, R.J.A. (2003). The Bayesian approach to sample survey inference. Chapter 4, Analysis of Survey Data, (Eds., R.L. Chambers and C.J. Skinner). Chichester: Wiley.

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

Little, R.J.A. (2022). Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference. Survey Methodology, 48, 2, 257-281. Paper available at https://www150.statcan.gc.ca/ n1/en/pub/12-001-x/2022002/article/00001-eng.pdf.

Little, R.J.A., and Rubin, D.B. (1987). Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics, $1^{\text {st }}$ Edition. New York: John Wiley \& Sons, Inc.

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

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

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. Proc. $6^{\text {th }}$ Berkeley Symp. Math. Statist., 1, 697-715.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society Series, Series B, 12, 241-255.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. International Statistical Review, 61, 317-337.

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

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. Statistica Sinica, 8, 1087-1114.

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

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Analysis of Survey Data, (Eds., C. Skinner and R. Chambers), New York: John Wiley \& Sons, Inc., 175195.

Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika, 66, 403-411.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Raessler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. Austrian Journal of Statistics, 33, 153-171.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. Survey Methodology, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K., and Molina, I. (2015). Small Area Estimation. Hoboken, NJ: Wiley.
Royall, R.M. (1970). On finite population sampling under certain linear regression models. Biometrika, 57, 377-387.

Royall, R.M. (1976). The least squares linear approach to two-stage sampling. Journal of the American Statistical Association, 71, 657-664.

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

Salvati, N., Fabrizi, E., Ranalli, M.G. and Chambers, R. (2021). Small area estimation with linked data. Journal of the Royal Statistical Society, Series B, 83, 78-107.

Samart, K., and Chambers, R. (2014). Linear regression with nested errors using probability-linked data. Australian and New Zealand Journal of Statistics, 56, 27-46.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. Survey Methodology, 33, 2, 99-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488eng.pdf.

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

Scott, A. (2006). Population-based case control studies. Survey Methodology, 32, 2, 123-132. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9546-eng.pdf.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. Biometrika, 84, 57-71.

Silva, P.L.D.N., and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. Survey Methodology, 23, 1, 23-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3102-eng.pdf.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Editors) (1989). Analysis of Complex Surveys. Chichester: Wiley.

Steel, D.G., Beh, E.J. and Chambers, R.L. (2004). The information in aggregate data. Ecological Inference: New Methodological Strategies, (Eds., G. King, O. Rosen and M. Tanner). Cambridge: Cambridge University Press.

Strauss, W.J., Carroll, R.J., Bortnick, S.M., Menkedick, J.R. and Schultz, B.D. (2001). Combining datasets to predict the effects of regulation of environmental lead exposure in housing stock. Biometrics, 57, 203210.

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

Zhang, L.-C., and Tuoto, T. (2021). Linkage-data linear regression. Journal of the Royal Statistical Society: Series A, 184, 522-547.

# Jean-Claude Deville's contributions to survey theory and official statistics 

Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé ${ }^{1}$


#### Abstract

Jean-Claude Deville, who passed away in October 2021, was one of the most influential researchers in the field of survey statistics over the past 40 years. This article traces some of his contributions that have had a profound impact on both survey theory and practice. This article will cover the topics of balanced sampling using the cube method, calibration, the weight-sharing method, the development of variance expressions of complex estimators using influence function and quota sampling.


Key Words: Calibration; Balanced sampling; Quota sampling; Variance estimation; Cube method; Weight-share method.

## 1. Introduction

Jean-Claude Deville, who passed away in October 2021, will undoubtedly leave an important legacy in survey statistics. For more than 40 years, as part of the National Institute of Statistics and Economic Studies (INSEE) and then École nationale de la statistique et de l'analyse de l'information (ENSAI) in France, he developed major innovations including calibration techniques; balanced sampling; indirect sampling and weight sharre methods; variance calculation, particularly for complex estimators; processing of nonignorable non-response; and quota surveys. That being said, he has worked in all survey fields and beyond. His exceptional productivity is mainly attributable to a very fruitful imagination combined with a remarkable mastery of mathematical tools. It was also fed by the concrete cases encountered at INSEE, which, like all national statistical institutes, was constantly confronted with various constraints and obstacles that had to be overcome, generally quickly and at a low cost. As head of the statistical methodology unit, he had to meet the technical challenges presented to him as they arose.

The following is an overview of Jean-Claude Deville's developments, all of which have moved on to posterity and can be found in depth in the many articles he published throughout his career, some shared with colleagues with whom he had privileged relationships. Clearly, some of his developments have found considerable international applications since their publication. There has even been an "industrial" implementation for calibration, the development of which was designed with another prestigious statistician, CarlErik Särndal.

## 2. Unequal and balanced probability sampling

### 2.1 Innovations in sampling algorithms

A sample is said to be balanced on a variable if the Horvitz-Thompson estimators of the totals calculated from a sample are equal to or nearly equal to the population total $U=\{1, \ldots, k \ldots, N\}$. Formally, suppose
that a vector of auxiliary variables $\mathbf{z}_{k}=\left(z_{k 1}, \ldots, z_{k Q}\right)^{\top}$ is known for all population units. Sample $S$ is balanced on $\mathbf{z}_{k}$ if

$$
\sum_{k \in S} \frac{\mathbf{z}_{k}}{\pi_{k}}=\sum_{k \in U} \mathbf{z}_{k}
$$

where $\pi_{k}$ is the inclusion probability, i.e., the probability that unit $k$ is selected in the random sample $S$.
The idea of selecting a balanced sample dates back to the very beginning of survey theory. Kiær (1896, $1899,1903,1905)$ was the first to propose what he called "representative counts". It is actually a selection of samples by quota. However, it was Gini and Galvani (1929) who first selected a balanced sample in official statistics. They selected 29 Italian districts (circondari) out of 214 to match several population averages as well as possible (Langel and Tillé, 2011; Tillé, 2016; Brewer, 2013). This method was harshly criticized by Jerzy Neyman because the sample was not randomly selected (see Bellhouse, 1988). Yates (1949) and Thionet (1953) proposed methods for which a sample is selected and then improved by successively replacing units to approach a balancing setting. Hájek $(1964,1981)$ proposed using rejective sampling, which consists of selecting a series of samples until a sufficiently balanced sample is obtained. However, this method has the drawback of changing the inclusion probabilities of units without being able to calculate them accurately afterwards (Choudhry and Singh, 1979; Dupačová, 1979; Fuller, 2009; Legg and Yu, 2010; Boistard, Lopuhaä and Ruiz-Gazen, 2012; Fuller, Legg and Li, 2017).

Jean-Claude Deville quickly became interested in sampling methods. In 1987, he published a book chapter with Jean-Marie Grosbras in which sampling methods were described and compared (Deville and Grosbras, 1987). The following year, along with Nicole Roth, they proposed a first balanced sampling method (Deville, Grosbras and Roth, 1988). The method applies only to equal probability of selection. The idea is to divide the variable space into quadrants and select one unit in the quadrant at each step that will contribute the most to achieving balancing. In the proceedings of the Örebro Conference held at Statistics Sweden in 1992, Jean-Claude Deville expressed his views on the three facets of the use of auxiliary information, namely constrained samples (i.e., balanced), conditional inference and weighting (Deville, 1992).

In parallel with this work, Jean-Claude Deville conducted very specific research on sampling issues. He proposed a formalization of sampling in a continuous population (Deville, 1989) long before the publication by Cordy (1993), who is often cited as the first reference in this field. He also proposed a selection method with unequal probabilities (Deville, 1998c), which is a variant of systematic sampling.

Then, with Yves Tillé, he proposed the splitting method (Deville and Tillé, 1998) to select samples with unequal probabilities of selection. This class of methods consists of starting with the vector of inclusion probabilities, $\boldsymbol{\pi}(0)=\boldsymbol{\pi}$, the sum of which is equal to an integer $n$. Then, at each step $t, 0,1,2, \ldots$, this vector $\boldsymbol{\pi}(t)$ is randomly modified until a vector containing only values equal to 0 or 1 is obtained, which corresponds to the selection of a sample. For the method to be correct, three conditions must be met:

1. All components of the $\pi(t)$ remain in the interval $[0,1]$.
2. The sum of the components of $\boldsymbol{\pi}(t)$ remains equal to $n$.
3. The martingale property must be satisfied:

$$
\begin{equation*}
\mathrm{E}_{p}\{\boldsymbol{\pi}(t) \mid \boldsymbol{\pi}(t-1)\}=\boldsymbol{\pi}(t-1), \text { for all } t, \tag{2.1}
\end{equation*}
$$

where $\mathrm{E}_{p}(\cdot)$ is the expectation with respect to the sampling design that takes the sampling randomization into account.

The martingale property is sufficient to show that the inclusion probabilities are respected at each stage. In fact, through the law of total expectation, we readily obtain $\mathrm{E}_{p}\{\boldsymbol{\pi}(t)\}=\boldsymbol{\pi}(0)$.

The splitting method is a very general way of representing a sampling method. Almost all selection algorithms can be viewed as a splitting procedure. This allows one to focus on basic steps. Checking the three conditions allows one to quickly check whether or not the method is correct.

One of the methods proposed as a special case of the splitting method is the pivot method for which only two components of the vectors $\boldsymbol{\pi}(t)$ are changed at each step. This method was generalized to select multiple non-overlapping samples in the same population with equal or unequal probabilities (Deville and Tillé, 2000b).

The transition from the splitting method to balanced sampling was relatively simple when Jean-Claude Deville and Yves Tillé realized that samples $\left(I_{1}, \ldots, I_{N}\right)^{\top}$ coded as vectors containing only 0 and 1 are the vertices of a $N$-cube of $\mathbb{R}^{N}$. In addition, conditions 1 and 2 of the splitting method can be interpreted geometrically. The vectors $\boldsymbol{\pi}(t)$ must remain in the simplex $\mathcal{P}=\left\{c_{k} \in[0,1] \mid \sum_{k=1}^{N} c_{k}=n\right\}$. Figure 2.1 shows a representation of this simplex for a sample of size $n=2$ in a population of size $N=3$. The splitting method is therefore a random walk in a simplex that must statisfy the martingale property.

Figure 2.1 Simplex bringing together samples of size $n=2$ in a population of $\operatorname{size} N=3$ within a cube where the samples are the vertices. Here, the simplex is an equilateral triangle.


### 2.2 Balanced sampling using the cube method

The shift to balanced sampling then became self-evident. It was simply a matter of replacing condition 2 in the splitting method to obtain the general principles of a balanced sampling method. The following three conditions are therefore obtained:

1. All components of the $\boldsymbol{\pi}(t)$ remain in the interval [0.1].
2. At each step $t=0,1,2, \ldots$ the vectors $\pi(t)=\left(\pi_{1}(t), \ldots, \pi_{N}(t)\right)^{\top}$ must meet the balancing equations:

$$
\sum_{k \in U} \frac{\mathbf{z}_{k}}{\pi_{k}} \pi_{k}(t)=\sum_{k \in U} \mathbf{z}_{k} .
$$

3. The martingale property must be satisfied

$$
\begin{equation*}
\mathrm{E}_{p}\{\boldsymbol{\pi}(t) \mid \boldsymbol{\pi}(t-1)\}=\boldsymbol{\pi}(t-1), \text { for all } t \tag{2.2}
\end{equation*}
$$

Conditions 1 and 2 now define a polytope

$$
\mathcal{P}=\left\{c_{k} \in[0,1] \left\lvert\, \sum_{k \in U} \frac{\mathbf{z}_{k}}{\pi_{k}} c_{k}=\sum_{k \in U} \mathbf{z}_{k}\right.\right\},
$$

in which the vectors $\boldsymbol{\pi}(t)$ must remain to statisfy the balancing constraints at each step. An example of a polytope is shown in Figure 2.2. However, when the constraints are complex, the vertices of the polytope $\mathcal{P}$ are not necessarily the vertices of the cube, meaning that there may not be exactly balanced samples. This will result in a roughly balanced sample. That is why the cube method consists of two phases: the flight phase and the landing phase.

The flight phase is a random walk in the polytope $\mathcal{P}$ that ends on one of the polytope vertices. The landing phase consists of selecting an approximately balanced sample close to the vertex of the polytope obtained at the end of the flight phase while satisfying the inclusion probabilities.

Figure 2.2 Polytope $\mathcal{P}$ in cases where the polytope vertices are not the vertices of the cube.


The cube method is a family of methods that allows such a random walk to be generated. For the flight phase, to go from $\boldsymbol{\pi}(t)$ to $\boldsymbol{\pi}(t+1)$, the cube method is conducted in the following manner.

1. A vector $\mathbf{u}(t)=\left(u_{1}(t), \ldots, u_{N}(t)\right)^{\top}$ is generated so that

$$
\sum_{k \in U} \frac{\mathbf{z}_{k}}{\pi_{k}} u_{1}(t)=\mathbf{0},
$$

and $u_{k}(t)=0$, if $\pi_{k}(t)$ is an integer ( 0 or 1 ). If such a vector does not exist, the flight phase stops.
2. We look for the largest positive values, $\lambda_{1}$ and $\lambda_{2}$, that satisfy

$$
0 \leq \pi_{k}(t)+\lambda_{1} u_{k}(t) \leq 1 \text { and } 0 \leq \pi_{k}(t)-\lambda_{2} u_{k}(t) \leq 1, \text { for all } k \in U .
$$

3. We update

$$
\boldsymbol{\pi}(t+1)=\left\{\begin{array}{l}
\boldsymbol{\pi}(t)+\lambda_{1} \mathbf{u}(t) \text { with probability } q \\
\boldsymbol{\pi}(t)-\lambda_{2} \mathbf{u}(t) \text { with probability } 1-q,
\end{array}\right.
$$

where $q=\lambda_{2} /\left(\lambda_{1}+\lambda_{2}\right)$.

There are several ways to generate the vector $\mathbf{u}(t)$, which allows you to define several variants of the method. After a maximum of $N$ steps, the flight phase ends on the vertex of the polytope $\mathcal{P}$. This vertex is a vector containing at most $Q$ values that are neither 0 nor 1 , where $Q$ is the number of auxiliary variables. To obtain a sample, one must apply the landing phase. Two variations are proposed in Deville and Tillé (2004).

The cube method was first published in the proceedings of the Journées de méthodologie statistique (Deville and Tillé, 2000a) and then as a chapter of a book (Deville and Tillé, 2001). The English publication was much more difficult but was eventually accepted in Biometrika (Deville and Tillé, 2004, 2005). A referee could not accept that the samples could be balanced and random at the same time. Another could not admit that the method worked without listing all possible samples. Another criticism was that the method did not provide exactly balanced samples. However, the existence of an exact solution does not depend on the method but on the geometry of the problem.

### 2.3 Implementation, method applications and research extensions

A first prototype of a SAS-IML function was written by three students from the ENSAI (Bousabaa, Lieber and Sirolli, 1999) under the supervision of Frédéric Tardieu and Yves Tillé. The first version was very slow, to the point its applicability was doubtful, but progress was quickly made. Chauvet and Tillé (2006a) proposed an implementation that considers only a small portion of the population at each stage, significantly reducing the computational time. An SAS function was written using this procedure (Chauvet and Tillé, 2006b). Several R packages also allow the selection of a balanced sample (Tillé and Matei, 2021; Grafström and Lisic, 2019; Jauslin, Eustache, Panahbehagh and Tillé, 2021). Their method is especially
simple because the functions depend on only two arguments: the matrix of balancing variables and the vector of inclusion probabilities.

Jean-Claude Deville was instrumental in changing the census procedure in France to a continuous system (Deville and Jacod, 1996). The cube method has been a valuable tool for constructing rotation groups (Durr and Dumais, 2002). The primary sampling units of the master sample were also selected using the cube method. The method was very quickly used in many applications (Tillé, 2011). As with calibration, balancing has become a standard procedure in survey statistics.

The cube method has also generated a lot of academic work. The accuracy of balancing is discussed in Chauvet, Haziza and Lesage (2015). Leuenberger, Eustache, Jauslin and Tillé (2022) suggest sorting observations in ascending depth order in the scatter plot, reducing the rounding problem. The issue of optimal inclusion probabilities is addressed in Nedyalkova and Tillé $(2008,2012)$ and Chauvet, Bonnéry and Deville (2011). These results generalized the optimal stratification of Neyman (1934). Several articles deal with balancing for stratified populations (Chauvet, 2009; Hasler and Tillé, 2014; Jauslin, Eustache and Tillé, 2021).

Several studies have been dedicated to spatial sampling. Grafström, Lundström and Schelin (2012) use the repulsive aspect of the pivot method to obtain samples properly spread out in space, increasing accuracy when data are autocorrelated. Grafström and Tillé (2013) then propose a variation of the cube method to obtain samples that are properly spread out and balanced on totals. Lastly, Jauslin and Tillé (2020a, b) balance on micro-strata containing the neighborhood of each unit to obtain particularly well-spread samples.

Jean-Claude Deville did a lot of work on maximum entropy plans, which he left several handwritten notes on (Deville, 2000b; Deville, nda, ndb, ndc). These results finally enabled a relatively quick implementation of this plan. Deville and Qualité (2005) then proposed an extension to the multidimensional case. As a result of a referee's remark during the submission of the article on the cube method, Jean-Claude Deville focused on determining a necessary and sufficient condition for balancing to have no rounding problems (Deville, 2015, 2014). Unfortunately, the condition obtained is very restrictive. In cases where the condition is met, it develops maximum entropy balanced designs (Deville, 2014).

Jean-Claude Deville quickly understood the value of the cube method for applications other than sampling. Several balanced imputation methods were proposed by Chauvet, Deville and Haziza (2011); Hasler and Tillé (2016); Eustache, Vallée and Tillé (2022). These methods have the advantage of properly restoring the distribution of the imputed variable while reducing the variance caused by random imputation. The cube method is also used in fields of application far from sampling such as in the MCMC (Monte Carlo Markov Chain) methods (Chopin and Ducrocq, 2021).

## 3. Calibration

The papers by Deville and Särndal (1992) and Deville, Särndal and Sautory (1993) on calibration methods (also called recovery methods) and published in the prestigious Journal of the American Statistical

Association are considered to be two of the most important and influential articles in the past 30 years in the field of sampling and official statistics. These two articles propose a unified theory of estimation in the presence of auxiliary information, the premises of which are discussed in Lemel (1976) and Huang and Fuller (1978). The two papers co-authored by Jean-Claude Deville have generated numerous research articles over the past three decades. The reader is referred to Särndal (2007), Haziza and Beaumont (2017), Devaud and Tillé (2019), and Zhang, Han and Wu (2022) for reviews on calibration methods. Poststratification (e.g., Holt and Smith 1979), raking methods (Deming and Stephan, 1940; Stephan, 1942), generalized regression estimation (see, for example, Särndal, Swensson and Wretman, 1992) can be obtained as special cases of calibration methods.

Calibration methods use auxiliary information available at the estimation stage to ensure consistency between the survey estimates produced and known or estimated external totals. In practice, calibration methods are also used to reduce non-response and coverage errors.

### 3.1 Calibration in the absence of non-sampling errors

In this section, we consider an ideal theoretical framework for which non-response and coverage errors are assumed to be negligible. Calibration is based on the availability of a vector of auxiliary variables, $\mathbf{x}_{k}=\left(x_{1 k}, \ldots, x_{J k}\right)^{\top}$, and the corresponding vector of population totals, $\mathbf{t}_{\mathbf{x}}=\left(t_{x_{1}}, \ldots, t_{x_{j}}\right)^{\top}$, where $t_{x_{j}}=$ $\sum_{k \in U} x_{j k}, j=1, \ldots, J$. The vector $\mathbf{t}_{\mathbf{x}}$ is obtained from an external source such as the census, an administrative file or another survey.

When selecting a sample $S$ of a population $U$, it is almost certain that the sample will suffer from a random distortion in terms of the vector of auxiliary variables $\mathbf{x}$, in the sense $\hat{\mathbf{t}}_{\mathbf{x}, \pi} \neq \mathbf{t}_{\mathbf{x}}$, with $\hat{\mathbf{t}}_{\mathbf{x}, \pi}=$ $\sum_{k \in S} \mathbf{x}_{k} / \pi_{k}$. Unlike a systematic distortion (as is usually encountered in a non-response context), we face a random distortion since $\mathrm{E}_{p}\left(\hat{\mathbf{t}}_{\mathbf{x}, \pi}\right)-\mathbf{t}_{\mathbf{x}}=\mathbf{0}$. The purpose of calibration is therefore to correct this distortion.

More formally, we are seeking a set of calibration weights $\left\{w_{k} ; k \in S\right\}$ such that

$$
\begin{equation*}
\sum_{k \in S} \frac{d_{k} G\left(w_{k} / d_{k}\right)}{q_{k}} \tag{3.1}
\end{equation*}
$$

is minimized subject to the $J$ calibration constraints

$$
\begin{equation*}
\sum_{k \in S} w_{k} \mathbf{x}_{k}=\mathbf{t}_{\mathbf{x}}, \tag{3.2}
\end{equation*}
$$

where $d_{k}=1 / \pi_{k}$ and $q_{k}$ is a scaling factor selected by the user (see Deville and Särndal, 1992; Deville et al., 1993). In the majority of cases encountered in practice, we set $q_{k}=1$ for all $k$. The function $G(\cdot)$ is a pseudo-distance function to measure the closeness between the weights before calibration $d_{k}$ and the weights after calibration $w_{k}$.

Calibration weights $w_{k}$ are given by

$$
\begin{equation*}
w_{k}=d_{k} F\left(q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right) \tag{3.3}
\end{equation*}
$$

where $\hat{\lambda}$ is a vector of size $J$ of estimated coefficients ensuring that the constraints (3.2) are satisfied, and $F(\cdot)=g^{-1}(\cdot)$ is the calibration function, defined as the inverse function of $g(.) \equiv \partial G(t) / \partial t$. The calibrated weight (3.3) can be viewed as the product of the weight before calibration, $d_{k}$, and an adjustment factor, $F\left(q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right)$. In addition to the vector $\hat{\lambda}$, the latter is dependent on the calibration function $F(\cdot)$ (and therefore pseudo-distance function $G(\cdot))$ as well as the characteristics of the unit $k, q_{k}$ and $\mathbf{x}_{k}$. In some situations, the term $q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}$ does not depend on $k$, in which case all calibration functions will lead to the same set of weights $w_{k}$. This occurs in the cases of post-stratification or ratio estimation (Haziza and Beaumont, 2017).

The calibration estimator of $t_{y}$ is given by

$$
\begin{equation*}
\hat{t}_{y, C}=\sum_{k \in S} w_{k} y_{k} \tag{3.4}
\end{equation*}
$$

Deville and Särndal (1992) considered a range of functions $G(\cdot)$ some of which are presented in Table 3.1. For the generalized chi-squared distance, Deville and Särndal (1992) showed that calibration weights are given by

$$
w_{k}=d_{k}\left(1+q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right)
$$

where

$$
\hat{\lambda}=\left(\sum_{k \in S} d_{k} \mathbf{x}_{k} q_{k} \mathbf{x}_{k}^{\top}\right)^{-1}\left(\mathbf{t}_{\mathbf{x}}-\hat{\mathbf{t}}_{\mathbf{x}, \pi}\right)
$$

It follows that the calibration estimator coincides with the well-known generalized linear regression estimator (see, for example, Särndal et al., 1992)

$$
\begin{equation*}
\hat{t}_{y, C}=\hat{t}_{y, \pi}+\left(\mathbf{t}_{\mathbf{x}}-\hat{\mathbf{t}}_{\mathbf{x}, \pi}\right)^{\top} \hat{\mathbf{B}} \tag{3.5}
\end{equation*}
$$

where

$$
\hat{\mathbf{B}}=\left(\sum_{k \in S} d_{k} \mathbf{x}_{k} q_{k} \mathbf{x}_{k}^{\top}\right)^{-1} \sum_{k \in S} d_{k} \mathbf{x}_{k} q_{k} y_{k} .
$$

This result is one of the important breakthrough in the field of estimation in the presence of auxiliary information: it is possible to construct the generalized regression estimator using calibration. Deville et al. (1993) have established that the use of Kullback-Leibler information (see Table 3.1) leads to the standard raking estimator, which is another major contribution. The truncated generalized chi-squared distance and the logit distance (see Table 3.1) allow for bounds to be placed on the calibration adjustment factors, $F\left(q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right)$, to limit the dispersion of the weights.

Although the calibration estimators are biased with respect to the sampling design, they are consistent, which is a desirable property (Deville and Särndal, 1992). When the sample size $n$ is large enough, the
square bias of calibration estimators becomes negligible in front of their variance. Therefore, the mean square error of calibration estimators is approximately equal to their variance, provided that $n$ is large enough.

Deville and Särndal (1992) showed that the variance of a calibration estimator can be approximated by

$$
\begin{equation*}
V_{p}\left(\hat{t}_{y, C}\right) \approx \sum_{k \in U} \sum_{\ell \in U} \Delta_{k \ell} \frac{E_{k}}{\pi_{k}} \frac{E_{\ell}}{\pi_{\ell}}, \tag{3.6}
\end{equation*}
$$

where $E_{k}=y_{k}-\mathbf{x}_{k}^{\top} \mathbf{B}$ is the "census residual" associated with unit $k$ with

$$
\mathbf{B}=\left(\sum_{k \in U} \mathbf{x}_{k} q_{k} \mathbf{x}_{k}^{\top}\right)^{-1} \sum_{k \in U} \mathbf{x}_{k} q_{k} y_{k} .
$$

This is a remarkable property: all calibration estimators have the same asymptotic variance regardless of the calibration function $F(\cdot)$. Expression (3.6) suggests that calibration estimators are efficient when residuals $E_{k}$ are small, which will occur when the relationship between the variable of interest $y$ and the calibration variables $\mathbf{x}$ is linear and strong. What if the relationship is not linear? In this case, the model may not fit the data well, leading to large residuals and a large variance. This has led Wu and Sitter (2001) to propose a model calibration procedure that allows for non-linear relationships through, for example, generalized linear models. However, unlike the classic calibration of Deville and Särndal (1992), model calibration requires the availability of the vector $\mathbf{x}$ for all population units. This requirement is generally not met in practice, especially in household surveys.

Table 3.1
A few distance functions introduced in Deville and Särndal (1992).

|  | Distance function $\boldsymbol{G}\left(\boldsymbol{w}_{\boldsymbol{k}} / \boldsymbol{d}_{\boldsymbol{k}}\right)$ | Calibration adjustment factor $\boldsymbol{F}\left(q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right)$ |
| :---: | :---: | :---: |
| Generalized chi-square distance | $\frac{1}{2}\left(\frac{w_{k}}{d_{k}}-1\right)^{2}$ | $1+q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}$ |
| Kullback-Leibler Information | $\frac{w_{k}}{d_{k}} \log \frac{w_{k}}{d_{k}}-\frac{w_{k}}{d_{k}}+1$ | $\exp \left(q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}\right)$ |
| Inverse Kullback-Leibler information | $\log \frac{d_{k}}{w_{k}}+\frac{w_{k}}{d_{k}}-1$ | $\frac{1}{1-q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}}$ |
| Hellinger distance | $2\left\{\sqrt{\frac{w_{k}}{d_{k}}}-1\right\}^{2}$ | $\frac{1}{\sqrt{1-2 q_{k} \hat{\lambda}^{\top} \mathbf{x}_{k}}}$ |
| Truncated generalized chi-square distance | $\begin{cases}\frac{1}{2}\left(\frac{w_{k}}{d_{k}}-1\right)^{2} & L<\frac{w_{k}}{d_{k}}<M \\ \infty & \text { otherwise }\end{cases}$ | $\begin{cases}1+q_{k} \lambda^{\top} \mathbf{x}_{k} & (L-1) \leq q_{k} \lambda^{\top} \mathbf{x}_{k} \leq(M-1) \\ M & q_{k} \lambda^{\top} \mathbf{x}_{k}>(M-1) \\ L & q_{k} \lambda^{\top} \mathbf{x}_{k}<(L-1)\end{cases}$ |
| Logit distance | $\begin{cases}\left(a_{k} \log \frac{a_{k}}{1-L}+b_{k} \log \frac{b_{k}}{M-1}\right) \frac{d_{k}}{A} & L<\frac{w_{k}}{d_{k}}<M \\ \infty & \text { otherwise }\end{cases}$ | $\frac{L(M-1)+M(1-L) \exp \left(A q_{k} \lambda^{\top} \mathbf{x}_{k}\right)}{M-1+(1-L) \exp \left(A q_{k} \lambda^{\top} \mathbf{x}_{k}\right)}$ |

In multi-stage or multi-phase surveys, we face several layers of auxiliary information. For example, in two-stage sampling, we may have auxiliary information at the household level (number of individuals in the household, number of individuals in each age group, owner or tenant status, etc.) and information at the individual level (gender, age group, etc.). The reader is referred to Sautory and Le Guennec (2003) and Estevao and Särndal $(2002,2006)$ for a discussion of calibration methods in a multi-stage or multi-phase sampling.

### 3.2 Adjustment of non-response by calibration

Post-stratification and raking methods have long been used to treat unit non-response; see, for example, Thomsen (1978), Bethlehem and Keller (1987), and Bethlehem (1988). The first work on a unified nonresponse calibration approach is presented in Deville and Dupont (1993) and Dupont (1993). The approach was further investigated by Lundström and Särndal (1999) and Särndal and Lundström (2005). The idea is to obtain final weights $w_{k}$ from the initial weights in order $d_{k}$ to achieve the following two objectives: (i) reduce non-response bias and (ii) ensure consistency between survey estimates and known population totals.

We consider a population $U$ in which a sample $S$ is selected and only a subset $S_{r}$ of units have responded. We therefore have $S_{r} \subset S \subset U$. We have two levels of auxiliary information: (1) The vector $\mathbf{x}_{U k}$ that is observed for $k \in S_{r}$ and for which the vector of population totals, $\sum_{k \in U} \mathbf{x}_{U k}$, is known. (2) The vector $\mathbf{x}_{S k}$ that is observed for $k \in S$ and for which the vector of Horvitz-Thompson estimates, $\sum_{k \in S} d_{k} \mathbf{x}_{S k}$, is available. Variables $\mathbf{x}_{U k}$ are those that will ensure consistency between survey estimates and known population totals. Ideally, the variables $\mathbf{x}_{S k}$ are those that explain both the response status $R_{k}$ and the variables of interest. For each $k \in S_{r}$, we create the stacked vector $\mathbf{x}_{k}=\binom{\mathbf{x}_{U k}}{\mathbf{x}_{S k}}$. We are seeking a final weighting system, $\left\{w_{k} ; k \in S_{r}\right\}$, such that

$$
\sum_{k \in S_{r}} \frac{d_{k} G\left(w_{k} / d_{k}\right)}{q_{k}}
$$

is minimized subject to the calibration constraints

$$
\sum_{k \in S_{r}} w_{k} \mathbf{x}_{k}=\binom{\sum_{k \in U} \mathbf{x}_{U k}}{\sum_{k \in S} d_{k} \mathbf{x}_{S k}} .
$$

The final weights are given by

$$
w_{k}=d_{k} \times F\left(q_{k} \hat{\lambda}_{r}^{\top} \mathbf{x}_{k}\right) .
$$

The estimator for $t_{y}$ is given by

$$
\begin{equation*}
\hat{t}_{y, C}=\sum_{k \in S_{r}}\left\{d_{k} \times F\left(q_{k} \hat{\lambda}_{r}^{\top} \mathbf{x}_{k}\right)\right\} y_{k} . \tag{3.7}
\end{equation*}
$$

Although any pseudo-distance function $G(\cdot)$ can be used, caution should be exercised. Indeed, selecting a pseudo-distance function in the context of non-response comes down to imposing a parametric model describing the relationship between the inverse of the response probabilities and the vector $\mathbf{x}$ (Haziza and Lesage, 2016). In general, an erroneous choice of the function $G(\cdot)$ will generally lead to a biased calibration estimator. An exception to this rule will occur when the variable of interest $y$ is linearly related to the vector $\mathbf{x}$ and non-response is of the Missing At Random type (Rubin, 1976).

Another important contribution by Jean-Claude Deville is calibration on instrumental variables, also known as generalized calibration (Deville, 1998a, 2000a, 2002). This approach has also been studied and discussed by, among others, Sautory and Le Guennec (2003), Kott (2006), Chang and Kott (2008), Kott and Chang (2010), Haziza and Beaumont (2017), and Lesage, Haziza and D’Haultfoeuille (2019). This approach is especially useful in the context of nonignorable non-response (Rubin, 1976). In this case, the response probability depends on fully observed variables but also variables available for the respondents only. As a result, estimating the response probabilities is not easy. Generalized calibration leads to a consistent estimator of a total if the exclusion restriction conditions are met.

## 4. The weight-sharing method

Indirect sampling involves selecting a sample from a target population using a different sampling frame, but somewhat related to that target population. Many developments related to the indirect survey can be found in Lavallée's books $(2002,2007)$ to which we add more recent contributions such as Deville and Maumy-Bertrand (2006); Falorsi, Piersante and Bako (2016); Kiesl (2010); Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine and Puech (2023). We will see that Jean-Claude Deville played a leading role in the development of indirect sampling.

### 4.1 The very beginning: Longitudinal surveys

The genesis of the indirect sampling relates to a weighting problem in the context of longitudinal surveys. This involved weighting individuals interviewed in a social longitudinal survey that tracks individuals belonging to an household over time.

After a selection of households (and thus individuals) in the first wave, changes in household composition throughout the waves, partly due to marriages and deaths, made the weighting process difficult. The solution is achieved by using the weight-share method (see Lavallée, 1995).

The problem of weighting longitudinal household surveys has attracted interest from several authors, including Huang (1984); Judkins, Hubble, Dorsch, McMillen and Ernst (1984); Ernst, Hubble and Judkins (1984); Ernst (1989); and Kalton and Brick (1995). The article by Ernst (1989) clearly described the basis of the problem and proposed a solution related to the weight share method.

Consider a longitudinal survey of individuals drawn from households. Two waves of data are available: the wave $A$ (or first wave) and the wave $B$ (a subsequent wave). A sample $S^{A}$ containing $m^{A}$ individuals was drawn from the population in the wave $A$ containing $M^{A}$ individuals. Let $\pi_{k}^{A}>0$, the probability of selection of individual $k$. In wave $B$, the population then contains $M^{B}$ individuals distributed among $N^{B}$ households $U_{i}^{B}$, where the household $i$ contains $M_{i}^{B}$ individuals.

The longitudinal survey process is as follows. For each individual $k$ from $S^{A}$, a list is established of $M_{i}^{B}$ individuals from the household $i$ in wave $B$ containing this individual, i.e., $S^{B}$, all of the $n^{B}$ households identified by the individuals $k \in S^{A}$. Once households from $S^{B}$ have been identified, all individuals $k$ in households $i \in S^{B}$ are surveyed to measure the variable of interest $y$. The weight-share method assigns an estimation weight $w_{i k}$ to each individual $k$ in a surveyed household $U_{i}^{B}$. The method steps are as follows:

- Step 1 For each individual $k$ in the households $i$ of $S^{B}$, we calculate the initial weight $w_{i k}^{\prime}=\gamma_{k} / \pi_{k}^{A}$, where $\gamma_{k}=1$ if $k \in S^{A}$, and 0 otherwise.
- Step 2 For each household $i$ of $S^{B}$, we obtain the total number of individuals $M_{i}^{A B}$ in the household $i$ present in wave $A$ (but not necessarily contained in $S^{A}$ ).
- Step 3 The final weight $w_{i}=\sum_{k \in U_{i}^{B}} w_{i k}^{\prime} / M_{i}^{A B}$ is calculated.
- Step 4 Lastly, we set $w_{i k}=w_{i}$ for all $k \in U_{i}^{B}$.

We could consider calculating the selection probability $\pi_{i k}^{B}$ of the individual $k$ in household $i$ of $S^{B}$. This probability corresponds to the probability of selecting any of the $M_{i}^{B}$ individuals in the household $i$, and therefore we must know each of the $M_{i}^{B}$ probabilities $\pi_{k}^{A}$ in the household $i$ of $S^{B}$. Unfortunately, especially in the case of multistage surveys, the probabilities $\pi_{k}^{A}$ are often unknown. In addition, apart from relatively simple cases (for example, when individuals $k$ are independently selected in $S^{A}$ ), the calculation of the weights $\pi_{i k}^{A}$ can be very complex. The weight-share method thus offers a simple solution to a weighting problem that is difficult, if not impossible, to carry out in practice.

### 4.2 A generalization of the problem

Imagine links (or correspondence) between individuals in both waves of the survey. Since it involves tracking individuals over time, these links can be seen as "one-to-one" (Figure 4.1). During discussions with Jean-Claude Deville, he came up with the following idea: "Why not generalize the links?" So instead of having "one-to-one" links, why not consider "many-to-many" links (Figure 4.2)? Figures 4.1 and 4.2 provide a graphical representation of the methods. The sample $S^{A}$ is the yellow subset of the wave $A$. The green subsets of the wave $B$ are clusters $U_{i}^{B}$ (the households) encountered in the second wave.

Figure 4.1 Longitudinal links ("one-to-one").


Figure 4.2 Arbitrary links ("many-to-many").


With this new way of looking at links, the question then became how to associate a weight (unbiased) to the surveyed units of $U^{B}$ (the target population) following the selection of units in $U^{A}$ (survey frame). In fact, the problem was much broader than that of longitudinal surveys.

The new problem studied was the following. Let $U^{A}$ and $U^{B}$ be two populations related to each other. An estimate is required for $U^{B}$ (target population), but a survey frame is available only for $U^{A}$. The
proposed solution is to then draw a sample from $U^{A}$ to produce estimates for $U^{B}$ using the existing correspondence (links) between the two populations. For this new approach, Jean-Claude Deville then coined the term indirect sampling.

Indirect sampling proceeds as follows. First, for each unit $j$ of $S^{A}$, we identify the units $k$ from the clusters $i$ of $U^{B}$ that have a link with $j$. Let $U_{i}^{B}$ be the set of all the units $k$ in cluster $i$. For each unit $k$ identified, we list the $M_{i}^{B}$ units from cluster $i$ containing this unit. Lastly, we survey all the units $k$ from the clusters $i \in S^{B}$ to measure the variable of interest $y$.

To illustrate indirect sampling, Jean-Claude Deville suggested an example where the goal is to survey people (units) living in dwellings (clusters). A sampling frame of dwellings is available, but unfortunately not up to date. This sampling frame does not include, among other things, the renovations impacting the divisions of dwelling in buildings. An example of this type of renovation is shown to the left of Figure 4.3. It can be seen that dwellings $a, b, c, d$ and $e$ were transformed to obtain dwellings $a^{\prime}, b^{\prime}, c^{\prime}$ and $d^{\prime}$. Drawing a sample of dwellings from the sampling frame, we get to the new dwellings using the correspondence between the old and new dwellings. This correspondence is illustrated to the right of Figure 4.3.

Figure 4.3 Pre- and post-renovation dwellings (left) and indirect sampling of renovated dwellings (right).


Estimating the total $t_{y}^{B}$ of the target population $U^{B}$ can be done using $S^{A}$ drawn from $U^{A}$. However, note that this can be a major challenge if the links between the units of $U^{A}$ and $U^{B}$ are not one-to-one. In fact, in this case, it is difficult, if not impossible, to associate a selection probability, or an estimation weight, to the units surveyed in $U^{B}$. The solution then is to use the Generalized Weight-Share Method (GWSM), which yields an estimation weight for each surveyed unit of the target population $U^{B}$.

As with longitudinal surveys, we consider a sample $S^{A}$ containing $m^{A}$ units drawn from $U^{A}$ containing $M^{A}$ units. The target population $U^{B}$ contains $M^{B}$ units and it is divided into $N^{B}$ clusters, where the cluster $i$ contains $M_{i}^{B}$ units. The links (or correspondence) between the units $j$ of $U^{A}$ and the units $k$ of the clusters $i$ of $U^{B}$ are identified by the variable $l_{i, j k}$, where $l_{j, i k}=1$ if there is a link between the unit $j \in U^{A}$ and the unit $k$ of the cluster $i$ of $U^{B}$, and 0 otherwise.

To apply the GWSM, simply follow these steps (reminiscent of the weight-share method, but more general):

- Step 1 For each unit $k$ of the clusters $i$ of $S^{B}$, we calculate the initial weight $w_{i k}^{\prime}=$ $\sum_{j \in U^{A}} l_{j, k} \gamma_{j} / \pi_{j}^{A}$, where $\gamma_{j}=1$ if $j \in S^{A}$, and 0 otherwise.
- Step 2 For each unit $k$ of the clusters $i$ of $S^{B}$, we obtain the total number of links $L_{i k}^{B}=$ $\sum_{j \in U^{4}} l_{j, i k}$.
- Step 3 The final weight $w_{i}=\sum_{k \in U_{i}^{B}} w_{i k}^{\prime} / \sum_{k \in U_{i}^{B}} L_{i k}^{B}$ is calculated.
- Step 4 Lastly, we apply $w_{i k}=w_{i}$ for all $k \in U_{i}^{B}$.

Lavallée $(2002,2007)$ mentioned that indirect sampling and the GWSM are useful because they offer a simple solution to complex survey and weighting problems. In addition, the GWSM generally yields the same results as classical results in the context of simple problems. In fact, the GWSM is an interesting solution, although it is not always the most accurate (minimum variance) compared with another more complex estimation method.

### 4.3 Properties of the Generalized Weight-Share Method

The development of the properties of the GWSM took place during discussions with Jean-Claude Deville, which began in 1995. These led to the following theorem and its two corollaries:

Theorem 1 Duality of the form of $\hat{t}_{y}^{B}$ with respect to $U^{A}$ and $U^{B}$. The estimator $\hat{t}_{y}^{B}$ can be written in both forms.

$$
\hat{t}_{y}^{B}=\sum_{i \in S^{B}} \sum_{k \in U_{i}^{B}} w_{i k} y_{i k}
$$

(with GWSM weights) and

$$
\hat{t}_{y}^{B}=\sum_{j \in U^{A}} \gamma_{j} Z_{j} / \pi_{j}^{A},
$$

where $Z_{j}=\sum_{i \in U^{B}} \sum_{k \in U_{i}^{l}} l_{j, i k} Y_{i} / L_{i}^{B}$.

Theorem 1 shows that we are, ultimately, in the presence of a simple Horvitz-Thompson estimator. From this finding, we obtain the following two corollaries:

Corollary 1 Bias of $\hat{t}_{y}^{B}$. The estimator $\hat{t}_{y}^{B}$ is unbiased for the estimation of $Y^{B}$, with respect to the sampling design.

Corollary 2 Variance of $\hat{t}_{y}^{B}$. The variance formula for the estimator $\hat{t}_{y}^{B}$, with respect to the sampling design, is given by

$$
V_{p}\left(\hat{t}_{y}^{B}\right)=\sum_{j \in U^{A}} \sum_{j^{\prime} \in U^{A}}\left(\pi_{i j^{\prime}}^{A}-\pi_{j}^{A} \pi_{j^{\prime}}^{A}\right) \frac{Z_{j} Z_{j^{\prime}}}{\pi_{j}^{A} \pi_{j^{\prime}}^{A}},
$$

where $\pi_{j j^{\prime}}^{A}$ is the joint probability of selection of units $j$ and $j^{\prime}$.

### 4.4 Calibration

Let's assume that we want to correct the GWSM weights so that the estimates produced correspond to known totals (auxiliary information). The most commonly used technique is the calibration developed by Deville and Särndal (1992).

In the context of indirect sampling, there are two possible sources of auxiliary information:
(i) From the survey frame $U^{A}$, we have a column vector $\mathbf{x}_{j}^{A}$ and its total $\mathbf{t}_{x}^{A}=\sum_{j \in U^{A}} \mathbf{x}_{j}^{A}$ (assumed to be known).
(ii) From the target population $U^{B}$, we have a column vector $\mathbf{x}_{i k}^{B}$ and its total $\mathbf{t}_{x}^{B}=\sum_{i \in U^{B}} \sum_{k \in U_{i}^{B}} \mathbf{x}_{i k}^{B}$ (assumed to be known).

The calibration constraints associated with the GWSM are:
(i) $\hat{\mathbf{t}}_{x}^{\mathrm{CAL}, A}=\sum_{j \in S^{A}} w_{j}^{\mathrm{CAL}, A} \mathbf{x}_{j}^{A}=\mathbf{t}_{x}^{A}$ and
(ii) $\hat{\mathbf{t}}_{x}^{\mathrm{CAL}, B}=\sum_{i \in S^{B}} \sum_{k \in U_{i}^{B}} w_{i k}^{\mathrm{CAL}, B} \mathbf{x}_{i k}^{B}=\mathbf{t}_{x}^{B}$, where $w_{j}^{\mathrm{CAL}, A}$ is the calibration weight obtained from $d_{j}^{A}=$ $1 / \pi_{j}^{A}$, and $w_{i k}^{\mathrm{CAL}, B}$ is the calibration weight of the unit $k$ from the surveyed cluster $i$ where the GWSM was applied.

Based on Theorem 1, the latter constraint can be rewritten as: $\hat{\mathbf{t}}_{x}^{\mathrm{CAL}, B}=\sum_{j \in S^{A}} w_{j}^{\mathrm{CAL}, A} \boldsymbol{\Gamma}_{j}=t_{x}^{B}$, where $\Gamma_{j}=\sum_{i \in U^{B}} \sum_{k \in U_{i}^{B}} l_{j, i k} X_{i}^{B} / L_{i}^{B}$. This constraint is now expressed in terms of units $j \in s^{A}$.

By defining the vectors

$$
\mathbf{x}_{j}^{A B}=\binom{\mathbf{x}_{j}^{A}}{\Gamma_{j}} \quad \text { and } \quad \mathbf{t}_{x}^{A B}=\binom{\mathbf{t}_{x}^{A}}{\mathbf{t}_{x}^{B}}
$$

we get the single constraint encompassing $U^{A}$ and $U^{B}$ :

$$
\hat{\mathbf{t}}_{x}^{\mathrm{CAL}, A B}=\sum_{j \in S^{A}} w_{j}^{\mathrm{CAL}, A} \mathbf{x}_{j}^{A B}=\mathbf{t}_{x}^{A B}
$$

The formulation of the problem for determining the estimator $\hat{t}_{y}^{\mathrm{CAL}, B}=\sum_{j \in S^{4}} w_{j}^{\mathrm{CAL}, A} Z_{j}$ associated with the GWSM is: Determine $w_{j}^{\mathrm{CAL}, A}$, for $j \in S^{A}$, to minimize the total distance

$$
\sum_{j \in S^{A}} G_{j}\left(w_{j}^{\mathrm{CLL}, A}, d_{j}^{A}\right)
$$

subject to the single constraint

$$
\hat{\mathbf{t}}_{x}^{\mathrm{CAL}, A B}=\sum_{j \in S^{4}} w_{j}^{\mathrm{CAL}, A} \mathbf{x}_{j}^{A B}=\mathbf{t}_{x}^{A B} .
$$

This formulation is consistent with that of Deville and Särndal (1992). Calibration can therefore be readily applied to indirect sampling and GWSM.

It is important to note that this calibration work was done without direct collaboration with Jean-Claude Deville. However, after all the work presented by Lavallée (2001) at the Colloque francophone sur les sondages in Brussels, he discovered the article of Deville (1998b), which provides the same solution to the calibration problem associated with GWSM.

### 4.5 Optimization of the links

The indicator variable $l_{j, j k}$ indicates whether or not there is a link between the units $j$ of the sampling frame $U^{A}$ and the units $k$ of the clusters $i$ of the target population $U^{B}$. However, it does not indicate the relative importance that some links might have over others. It is possible to replace $l_{j, i k}$ with a quantitative variable $\theta_{j, i k}$ representing the importance we want to give to the link $l_{j, i k}$. This variable $\theta_{j, i k}$ is defined on $[0,+\infty)$, where $\theta_{j, i k}=0$ is equivalent to $l_{j, i k}=0$. It should be noted that if the process for assigning values of $\theta_{j, i k}$ is independent of the selection of $S^{A}$, the GWSM remains unbiased.

By replacing the links $l_{j, i k}$ with $\theta_{j, j k}$, we obtain a new estimator (unbiased) $\hat{t}_{y \theta}^{B}$. The problem is then to determine optimal values $\theta_{j, k}$ so as to minimize the variance of $\hat{t}_{y \theta}^{B}$. Indeed, since the estimator $\hat{t}_{y \theta}^{B}$ remains unbiased regardless of the values of $\theta_{j, i k}$, it must be possible to determine the values of the latter to maximize the precision of $\hat{t}_{y \theta}^{B}$. The problem is therefore to determine $\left[\theta_{j, j k}\right]_{M^{1} \times M^{B}}$ to minimize $V_{p}\left(\hat{t}_{y \theta}^{B}\right)=$ $f\left(y_{i k} ; i=1, \ldots, N^{B} ; k=1, \ldots, M_{i}^{B}\right)$.

Deville and Lavallée (2006) determined the values of $\theta_{j, j k}$ such that the variance of the estimator $\hat{t}_{y \theta}^{B}$ is (almost) minimal. The optimal solution is relatively complex, and often depends on the variable of interest $y$. However, Jean-Claude Deville came up with the idea of defining the concept of weak optimality as well as that of strong optimality independent of the $y$-variables.

Weak optimality consists of determining values of $\theta_{j, i k}$ to minimize the variance of $\hat{t}_{y \theta}^{B}$ for a very specific choice of a variable of interest: $y_{i k}=1$ for a given unit $k$ of a cluster $i$ of $U^{B}$ and $y_{i k^{\prime}}=0$ for all other units of $U^{B}$. The optimization problem reduces to determining $\left[\theta_{j, j, k}\right]_{M^{A} \times M^{B}}$ as to minimize $V_{p}\left(\hat{t}_{y \theta}^{B}\right)=$ $f\left(y_{i k}=1 ; y_{i k^{\prime}}=0 ; \forall i \neq i^{\prime}\right.$ and $\left.k \neq k^{\prime}\right)$. Deville and Lavallée (2006) mention that weak optimality consists of minimizing the variance $\mathrm{V}\left(w_{i k}^{\theta} \mid i k \in S^{B}\right)$ of the weight $w_{i k}^{\theta}$ obtained by the GWSM (with $\theta_{j, i k}$ instead of $l_{j, j k}$ ) for all possible values of $\left[\theta_{j, j k}\right]_{M^{4} \times M^{B}}$. We note that the resulting weakly optimal weighted links do
not involve the variable $y$ itself (since the values of $y$ have been replaced by 1 and 0 ). In addition, the weakly optimal values of $\theta_{j, i k}$ are generally relatively easy to calculate.

Strong optimality independent of $y$ involves an additional step to weak optimality. It consists of checking that the weakly optimal values of $\theta_{j, i k}$ do not generally depend (i.e., for any value of $y$ other than 1 and 0 ) on the variable of interest $y$. With this in mind, Deville and Lavallée (2006) proposed a criterion to check whether the weak optimality corresponds to the strong optimality (minimum variance of $\hat{t}_{y \theta}^{B}$ ). If this criterion is met, strong optimality does not depend on the variable of interest $y$.

## 5. The development of variance expression and its estimation for complex estimators

The development of a variance formula and its estimation for a sample estimator is an essential step in producing confidence intervals that will inform users of statistics on their reliability. Conventional theory uses either an analytical approach, which is, by nature, based on mathematical formulas, or a sample replication approach (bootstrap, jackknife, random groups). Roughly speaking, the analytical approach may be considered more applicable when sampling is complicated and the expression estimator is rather simple, whereas the replication approach is used more readily in the opposite configuration, that is, in the presence of simple sampling and a complex estimator. This was certainly a common strategy before the development of the linearization theory for complex statistics. Jean-Claude Deville has contributed a great deal to the theory of linearization. Of course, the technique of linearization of estimators defined as functions of linear components, typically differentiable functions of estimators of totals, such as a ratio or coefficient of regression or estimators that are the solution of estimating equations (Woodruff, 1971; Binder, 1983; Wolter, 1985; Binder, 1991; Francisco and Fuller, 1991; Binder and Kovačević, 1995; Binder, 1996), has been known for many years. In the late 1990s, Jean-Claude Deville proposed a formal framework based on the influence function in the journal Survey Methodology (Deville, 1999) to deal with highly non-linear statistics, such as fractiles or parameters defined as solutions to certain equations (implicit parameters), in an asymptotic and general setting. When the sample size is large, linearization eventually allows a very complex estimator to be approximated by a classical linear estimator of the Horvitz-Thompson type, and then the variance of the former is approximated by the variance of the latter, thus producing the desired result.

More specifically, the historical approach considers the parameter $\theta$ and estimator $\hat{\theta}$ as differentiable functions of the individual variables of interest. Linearization is then based on a Taylor expansion procedure of $\hat{\theta}$ around its expectation $\theta$. Survey weights are present in the expression of $\hat{\theta}$ but are not treated as variables. Jean-Claude Deville reverses the approach in some ways by considering $\hat{\theta}$ as a function of survey weights and uses a derivative with respect to the weights; it is the influence function, introduced in the next part.

### 5.1 The theoretical framework

The proposed methodology is explained in three stages: first, an asymptotic framework, second, a formalization using the concept of measure on a probabilistic space, and third, the use of the concept of influence function, which is appropriately adapted to the context.

The asymptotic framework is the one defined in Isaki and Fuller (1982), and considers a series of nested populations, with respective sizes $N$ going to infinity, within which samples $s$ are selected, whose size $n$ also goes to infinity. For any individual variable $x_{k}$, if $t_{x}$ is the true total of $x_{k}$ and

$$
\hat{t}_{x}=\sum_{k \in S} w_{k} x_{k}
$$

its estimator, we assume that $N^{-1} t_{x}$ has a limit, and that $N^{-1}\left(\hat{t}_{x}-t_{x}\right)$ converges towards 0 in probability and that $\sqrt{n} N^{-1}\left(\hat{t}_{x}-t_{x}\right)$ converges in distribution towards a Gauss distribution. Any complex statistic $S$ constructed from true or estimated totals is based on similar assumptions; depending to its expression, it is consistent when it is multiplied by $N^{-\alpha}$, where $\alpha$ is a positive integer. The integer $\alpha$ is called the degree of homogeneity. A ratio is a homogeneous statistic of degree 0 and a variance is a homogeneous statistic of degree 2 . The first axiom is therefore extended by assuming that $N^{-\alpha} S$ has a limit.

Then comes the formalization of the estimators by using the concept of measure. In the expression of any parameter, individuals in the finite population are "naturally" weighted by a weight equal to 1 , interpreted as a mass associated with a measure $M$ with finite support. In the same population, sampling results in weighting any individual $k$ in the selected sample $s$ by the survey weight $w_{k}$ and any individual $k$ outside $s$ by 0 . This leads to the measure $\hat{M}$. A parameter, however complex, can be expressed as a function of $M$, noted $T(M)$ and called "functional of $M$ ". As an example, consider a total

$$
T(M)=\sum_{k=1}^{N} y_{k}=\sum_{k=1}^{N} y_{k} M(k) .
$$

Adopting a general notation familiar in the context of measure theory, we write $T(M)=\int y d M$, where we integrate over all the individuals in the population. Turning to estimators, again in the case of a total, we consider

$$
T(\hat{M})=\sum_{k \in s} w_{k} y_{k}=\sum_{k=1}^{N} y_{k} \hat{M}(k),
$$

in which case $T(\hat{M})=\int y d \hat{M}$. This parallelism applies to any complex parameter, initially expressed as $T(M)$ and estimated by $T(\hat{M})$, an estimator obtained by substituting $M$ with $\hat{M}$.

The third component of the theory uses the notion of influence function, which is used in the theory of robust statistics (Hampel, Ronchetti, Rousseeuw and Stahel, 1985). We consider a specific measure $\delta_{k}$ obtained by assigning a mass equal to 1 to the individual $k$, and the measure $M+t \delta_{k}$ leading to the mass
$1+t$ to the individual $k$ and to the mass 1 for all other individuals. The influence function is defined when the limit exists - by

$$
I T(M, k)=\lim _{t \rightarrow 0} \frac{T\left(M+t \delta_{k}\right)-T(M)}{t}
$$

It can be shown, under certain technical conditions most often satisfied in practice, that when the measure space is equipped with a distance, if a measure $M_{2}$ converges to a measure $M_{1}$, then

$$
T\left(M_{2}\right)=T\left(M_{1}\right)+\int I T\left(M_{1}, k\right) d M_{2}-\int I T\left(M_{1}, k\right) d M_{1}+R_{\epsilon}
$$

where $R_{\epsilon}$ is a random residual that converges to 0 in probability. This equality must be adapted to the postulated initial asymptotic conditions: to deal generally with the homogeneous statistics of degree $\alpha$, it is the functionals $N^{-\alpha} T$ that have the required asymptotic properties and must therefore be considered here. Also, noting that the total mass associated with the measurement $M$ is $N$, by setting a distance

$$
d\left(\frac{M_{1}}{N}, \frac{M_{2}}{N}\right)=\left|\int y d\left(\frac{M_{1}}{N}\right)-\int y d\left(\frac{M_{2}}{N}\right)\right|
$$

and by setting $M_{1}=M$ and $M_{2}=\hat{M}$, it follows that $\hat{M} / N$ converges toward $M / N$, which ultimately leads to:

$$
N^{-\alpha} T(\hat{M})=N^{-\alpha} T(M)+\int I T(M, k) d\left(\frac{\hat{M}}{N}\right)-\int I T(M, k) d\left(\frac{M}{N}\right)+R_{\epsilon}
$$

where the residual $R_{\epsilon}$ is negligeable in probability in front of $1 / \sqrt{n}$, under these conditions, that is,

$$
\text { for all } \epsilon>0, P\left(\left|\sqrt{n R_{\epsilon}}>\epsilon\right|\right) \rightarrow 0
$$

By noting $\operatorname{IT}(M, k)=z_{k}$, it follows that

$$
\begin{equation*}
N^{-\alpha}(T(\hat{M})-T(M))=\frac{1}{N} \sum_{k \in s} w_{k} z_{k}-\frac{1}{N} \sum_{k=1}^{N} z_{k}+R_{\epsilon} \tag{5.1}
\end{equation*}
$$

By noting $\hat{t}_{z}=\sum_{k \in s} w_{k} z_{k}$ the natural linear estimator of the total $t_{z}=\sum_{k=1}^{N} z_{k}$,

$$
\sqrt{n} N^{-\alpha}(T(\hat{M})-T(M))=\sqrt{n} \frac{1}{N}\left(\hat{t}_{z}-t_{z}\right)+R_{\epsilon}
$$

According to the third asymptotic assumption, the term on the right has a Gaussian limit, and therefore the term on the left hand-side has an asymptotic variance, which is equal to that on the right hand-side. It is customary to use $N^{2(\alpha-1)} V\left(\hat{t}_{z}\right)$ as an approximate variance of $T(\hat{M})$ when $n$ is considered to be "sufficiently large". The variable $z_{k}$ is called a linearized variable associated with the functional $T$. To carry out the variance estimation in practice, when $z_{k}$ depends on a finite number of parameters that can be
estimated by using the sample data, $z_{k}$ will be replaced by its natural estimator $\hat{z}_{k}$. The estimated variance then differs from the true variance by a term whose order of magnitude is $n^{-1 / 2}$.

Deville's paper was once again innovative since it was followed by several other works on variance estimation. While Deville proposed to linearize the parameter of interest at the population level, Demnati and Rao $(2004,2010)$ derive the estimator directly from survey weights. This method is a simple way to calculate the influence function on the estimator. $\operatorname{Graf}$ (2011); Antal, Langel and Tillé (2011); Graf and Tillé (2014); and Vallée and Tillé (2019) derived the estimator by the sample selection indicators, allowing for both the non-linearity of the estimator and calibration. The results given by the different methods are not always identical because the weights may depend on the selection indicators, especially when the estimator is calibrated.

### 5.2 The tools

From the previous theory, we obtain rules for calculating linearized variables that allow the treatment of complex estimators in a simple fashion, by breaking them down. The "total" functional $T(M)=\sum_{i=1}^{N} y_{i}$ is the simplest. Since

$$
T\left(M+t \delta_{k}\right)=\sum_{i=1}^{N} y_{i} M(i)+\sum_{i=1}^{N} y_{i} t \delta_{k}(i)=\sum_{i=1}^{N} y_{i}+t y_{k},
$$

it follows that $z_{k}=I T(M, k)=y_{k}$. The expression (5.1) is here tautological, with $R_{\epsilon}=0$.
Various useful properties are cited in the founding article.
(i) Let $T(M)$ and $S(M)$ be two functionals:

$$
I(T+S)(M, k)=I T(M, k)+I S(M, k)
$$

and

$$
I(T \cdot S)(M, k)=I T(M, k) \cdot S(M, k)+I S(M, k) \cdot T(M, k) .
$$

(ii) Let $T(M)$ be a vector of totals in $\mathbb{R}^{p}$ and $f$ be a differentiable function of $\mathbb{R}^{p}$ in $\mathbb{R}$ with the $(p, p)$ matrix of partial derivatives evaluated at $T(M)$ is noted $D_{f, T(M)}$. Then, $\operatorname{If}(T)(M, k)=$ $D_{f, T(M)} \cdot I T(M, k)$. This rule is useful for finding well known linearized, of smooth functions of totals, such as ratios or linear correlation coefficients.

Now consider functionals parametrized by a vector $\alpha$ of $\mathbb{R}^{p}$, noted $T(M, \alpha)$. For example, if $p=1$ and $\alpha \in[0,1]$, noting $F$ the distribution function associated with the distribution of $y_{k}$, $T(M, \alpha)=F^{-1}(\alpha)$ is the quantile of order $\alpha$ of the distribution.
(iii) If $T(M, \lambda)$ as a function of $\lambda$ has sufficient regularity, a reciprocal functional $\Lambda(M, \alpha)$ can therefore be defined satisfying $T(M, \Lambda(M, \alpha))=\alpha$. For example, to treat the variance of a quantile, we will consider $T(M, \lambda)=F(\alpha)$ and we obtain $\Lambda(M, \alpha)=F^{-1}(\alpha)$. It can be shown that

$$
\begin{equation*}
I \Lambda(M, \alpha, k)=-\left\{\frac{\partial T}{\partial \alpha}(M, \Lambda(M, \alpha))\right\}^{-1} \cdot I T(M, \alpha, k) \tag{5.2}
\end{equation*}
$$

This theorem is useful for obtaining, for example, the linearization of a quantile or implicit parameter (see 5.3).
(iv) Suppose that the parameter is written as a function of the value $y$, that is, let's consider a functional of the type $T(M, \phi(y))$ where $\phi$ is a function with the right technical properties, then the functional $S(M)=\int T(M, \phi(y)) d M$. So,

$$
\begin{equation*}
I S(M, k)=T\left(M, \phi\left(y_{k}\right)\right)+\int I T(M, \phi(y), k) d M \tag{5.3}
\end{equation*}
$$

This theorem is used, for example, to determine the linearized Gini coefficient (see 5.3).
(v) If the parameter is itself a functional $S(M)$, we can obtain the influence function of $T(M, S(M))$ - that is, the linearization of a compound of functionals - as

$$
I T(M, S(M), k)=I T(M, \alpha, k)+\frac{\partial T}{\partial \alpha}(M, \alpha) \cdot I S(M, k),
$$

where $\alpha$ takes the value $S(M)$ in the final expression.

We also mention an interesting property: for any functional of degree $\alpha$, we have

$$
\sum_{k=1}^{N} I T(M, k)=\alpha \cdot T(M) .
$$

In particular if $\alpha=0$ (a ratio for example), the population total of the linearized variables $z_{k}$ is equal to zero.

### 5.3 A few applications

The above theory makes it possible to linearize virtually all the estimators that are encountered in survey statistics. It is therefore general in scope, and ultimately permits all analytical variance calculations, for (about) every conceivable parameter, under some asymptotic conditions, that is, when the sample size $n$ is considered to be "sufficiently large".

The original article by Jean-Claude Deville presents several examples of applications for complex parameters. It contains, along with the technical developments justifying them, the cases of implicit parameters, quantiles, the Gini coefficient, the poverty line (defined as the proportion of individuals in a population whose income is less than half its median), and the Kendall rank coefficient of correlation. The variance estimation for a principal component and that of the projection of a point representing any subpopulation on a factorial axis as part of a multiple correspondence analysis are also discussed. Below, we provide some results regarding implicit parameters, quantiles, and the Gini inequality coefficient.

An implicit parameter is a vector of $\mathbb{R}^{p}$ solution of an equation of the form $H(M, \mathbf{B})=0$, where $H(M, \mathbf{B})=\sum_{k \in U} l_{k}(\mathbf{B})$, the functions $l_{k}$ being regular functions of $\mathbb{R}^{p}$ in $\mathbb{R}^{p}$. Using (5.2), and noting $\mathbf{B}_{0}$ the solution of the equation, we get for all $k$ of $U$,

$$
I \mathbf{B}(M, k)=-\left\{\frac{\partial H}{\partial \mathbf{B}}\left(M, \mathbf{B}_{0}\right)\right\}^{-1} \cdot \operatorname{IH}\left(M, \mathbf{B}_{0}, k\right),
$$

which means that

$$
I \mathbf{B}(M, k)=-\left\{\sum_{k \in U} \frac{\partial l_{k}}{\partial \mathbf{B}}\left(M, \mathbf{B}_{0}\right)\right\}^{-1} \cdot l_{k}\left(\mathbf{B}_{0}\right) .
$$

This situation is, for example, that of regression coefficients, the $l_{k}$ stemming from normal equations. If the regression is linear, it obtain (classical notations)

$$
l_{k}(\mathbf{B})=\frac{1}{\sigma_{k}^{2}} \cdot \mathbf{x}_{k}\left(y_{k}-\mathbf{z}_{k}^{\top} \mathbf{B}\right)
$$

and finally the linearized vector of $\mathbb{R}^{p}$

$$
\mathbf{z}_{k}=I \mathbf{B}(M, k)=-\left(\sum_{k \in U} \frac{\mathbf{x}_{k} \mathbf{z}_{k}^{\top}}{\sigma_{k}^{2}}\right)^{-1} \cdot \frac{1}{\sigma_{k}^{2}} \cdot \mathbf{x}_{k}\left(y_{k}-\mathbf{z}_{k}^{\top} \mathbf{B}_{0}\right)
$$

which can also be found through the "traditional" approach. In the case of logistic regression, the conventional tools are no longer sufficient and then

$$
\mathbf{z}_{k}=I \mathbf{B}(M, k)=\left\{\sum_{k \in U} \mathbf{x}_{k} \mathbf{z}_{k}^{\top} \cdot f\left(\mathbf{z}_{k}^{\top} \mathbf{B}_{0}\right)\left(1-f\left(\mathbf{z}_{k}^{\top} \mathbf{B}_{0}\right)\right)\right\}^{-1} \mathbf{x}_{k}\left(y_{k}-f\left(\mathbf{z}_{k}^{\top} \mathbf{B}_{0}\right)\right),
$$

where $f(u)=e^{u}\left(1+e^{u}\right)$.
To obtain the linearized variable of a quantile, the functional "distribution function" must be considered:

$$
F(x)=T(M, x)=\frac{1}{N} \cdot \operatorname{Card}\left\{k \in U / x_{k} \leqslant x\right\}=\frac{1}{N} \cdot \int 1_{x_{k} \leqslant x} d M,
$$

whose influence function is

$$
I T(M, x, k)=\frac{1}{N} \cdot\left\{1_{x_{k} \leqslant x}-F(x)\right\} .
$$

Assuming for simplicity that this increasing function is differentiable and invertible, we define the quantile $q_{\alpha}$ of order $\alpha$, where $\alpha \in[0,1]$, using $F\left(q_{\alpha}\right)=\alpha$. Application of the formula (5.2) would then lead to the linearized variable

$$
z_{k}=I q_{\alpha}(M, k)=-\left[N \cdot F^{\prime}\left(q_{\alpha}\right)\right]^{-1} \cdot\left(1_{x_{k} \leqslant q_{\alpha}}-\alpha\right) .
$$

Jean-Claude Deville has proposed a clever idea that makes it simple to take into account the fact that $F(x)$ is a step function, which is neither derivable nor invertible.

Let's finish by specifying the linearization of a "Gini Index", which is a standard inequality index. The index considered in the article is defined as

$$
\mathrm{GINI}=\frac{1}{t_{x}} \int x F(x) d M,
$$

where $t_{x}=\sum_{k \in U} x_{k}$. Using (5.3) we obtain

$$
z_{k}=F\left(x_{k}\right) \cdot \frac{x_{k}-\bar{x}_{k, \text { inf }}}{t_{x}}-\mathrm{GINI} \cdot \frac{x_{k}}{t_{x}}
$$

by applying

$$
\bar{x}_{k, \text { inf }}=\frac{\int x 1_{x<x_{k}} d M}{\int 1_{x<x_{k}} d M}
$$

which corresponds to the average of the $x$-values less than $x_{k}$.

## 6. Quota sampling

"Quota" sampling is not a survey method used frequently by national statistical institutes. The key argument is the resulting bias of the estimators, where smaller variances are not made available as opposed to with a well-chosen probabilistic method, given the large sample sizes commonly encountered in official statistics. Another more philosophical argument is the dependence of the quality of the estimate on a model, which is a set of simplifying assumptions of reality, and is sometimes overly simplifying. Jean-Claude Deville often made this second argument, which he considered to reflect a lack of neutrality that was not suitable for a national statistical institute, at least when it is possible to do otherwise (it may be argued that models are systematically used to deal with non-response, but this is unavoidable). It may be because this highly used empirical survey method (particularly in the private sector) is inherently risky and therefore controversial that Jean-Claude Deville felt the need to formalize the question. He appears to have been the first to do so in a comprehensive manner, writing an authoritative article in the journal Survey Methodology in 1991 (Deville, 1991). In this article, two types of models are distinguished: one model deals with sampling, another with the variable of interest. In each configuration, an estimator is given, its bias studied, and when possible, the author produces a theoretical expression of variance as well as an unbiased estimator of the variance.

### 6.1 The general framework

Recall the principle of quota sampling, by, for simplicity, restricting to the case of one or two so-called quota variables. Auxiliary information is made up of subpopulation sizes defined by the modalities of the
quota variables, which are therefore qualitative in nature. If there is a single qualitative variable with $I$ modalities to define these subpopulations, we have the population count $N_{i}$ for each modality $i$ varying between 1 and $I$. If there are two qualitative variables with respectively $I$ and $J$ modalities, denoting by $N_{i, j}$ the population count of the cell $(i, j)$, we know the marginal count $N_{i,=}=\sum_{j=1}^{J} N_{i, j}$ for any $i$ varying between 1 and $I$ as well as the marginal count $N_{., j}=\sum_{i=1}^{I} N_{i, j}$ for all $j$ varying between 1 and $J$ ("marginal" quotas). It should be noted that availability of the cross counts $N_{i, j}$ ("cross" quotas) brings us back to the case of a single qualitative variable. The selection of the sample is done empirically, without a sampling frame, following a few collection guidelines to randomize the composition of the overall sample of size $n$ as much as possible while imposing constraints on the $n_{i}$ 's (case of a quota variable) or assigning the $n_{i, j}$ 's (case of two quota variables). These constraints, which will be referred to as "quota constraints", are set at the discretion of the survey statistician, but in practice, the most common method is the method of proportional quotas, where the idea is to ensure that the sample has the same structure as the population, i.e., $n_{i}=n \frac{N_{i}}{N}$ in the case of a single variable, or $n_{i, n}=n \frac{N_{i,}}{N}$ for all $i$ and $n_{,, j}=n \frac{N_{, j}}{N}$ for all $j$ in the context of cross quotas ("proportional quota" constraints). Although this is a reassuring standard, it does not correspond to the intuitive optimal situation, since it is always preferable to increase the sample sizes in cells with the greatest dispersion (refer to the Neyman optimum in stratified sampling - which differs from proportional allocation but is preferable).

### 6.2 A sampling model

The case of quotas on a single variable can be treated by pretending that we are in the case of a stratified sampling design with simple random sampling in each stratum, each stratum being associated with a modality $i$ of the quota variable, whereby the sample size $n_{i}$ is imposed. It is difficult to imagine other alternatives, so there is nothing original to add in this simple context.

The interesting case is that of marginal quotas. The model has two phases. First, the selection can be viewed as a simple random draw under constraint, the constraints being those imposed on the $n_{i, .}$ and on the $n_{, j, j}$, i.e., the constraints of quotas. That is a pretty daring assumption, that assumes that interviewers are completely neutral in selecting their respective samples. Technically, a simple random draw under constraint reduces to considering that any sample not meeting quotas has a zero probability of selection and that all samples meeting quotas have the same probability of selection. Practically, it could be implemented by conducting successive and independent simple random draws, while rejecting samples until quota constraints are satisfied (rejective selection). Secondly, the existence of these constraints is somewhat forgotten.

Let $P_{i, j}$ be the weight $\frac{N_{i, j}}{N}$ of the cell $(i, j)$. The objective is to estimate these weights. Indeed, the average $\bar{Y}$ of any variable of interest $Y$ defined in the population is written $\bar{Y}=\sum_{(i, j)} \frac{N_{i, j}}{N} \cdot \bar{Y}_{i, j}$ and, under the postulated model, the true average $\bar{Y}_{i, j}$ in the cell $(i, j)$ is estimated without bias by the simple average $\bar{y}_{i, j}$ in the sample intersecting this cell, so that the natural estimator by quotas of $\bar{Y}$ will be $\hat{\bar{Y}}_{\text {quota }}=\sum_{(i, j)} \hat{P}_{i, j} \cdot \bar{y}_{i, j}$ where $\hat{P}_{i, j}$ is an estimate of $P_{i, j}$ - as much as possible without bias. With a standard simple random draw
selection, the sample sizes per cell $(i, j)$, i.e., $n_{i, j}$, follow a multinomial distribution. In the second stage of modeling, we postulate that, subject to the quota constraints, the distribution of the $n_{i, j}$ 's remains mutinomial. In this context, the estimates $P_{i, j}$ are obtained, for example, using maximum likelihood estimation, which leads to maximizing the objective function $\prod_{(i, j)} P_{i, j}^{n_{i, j}}$. Solutions must be compatible with known marginal information, which also requires

$$
\sum_{j=1}^{J} P_{i, j}=\frac{N_{i, .}}{N} \text { for all } i \text { and } \sum_{i=1}^{I} P_{i, j}=\frac{N_{, . j}}{N} \text {, for all } j \text {. }
$$

The solutions obtained are in the form $\hat{P}_{i, j}=\frac{n_{i, j}}{n}\left(a_{i}+b_{j}\right)^{-1}$ (the unknown coefficients $a_{i}$ and $b_{j}$ are the Lagrange coefficients associated with the constraints). Since there are $I+J-1$ independent constraints, for example, by imposing the identifier constraint $b_{J}=0$, we obtain the $a_{i}$ and the $b_{j}$, by solving the system (S)

$$
\begin{gathered}
\sum_{j=1}^{J} n_{i, j}\left(a_{i}+b_{j}\right)^{-1}=n \frac{N_{i, .}}{N}, \text { for all } i=1, \ldots, I, \\
\sum_{i=1}^{I} n_{i, j}\left(a_{i}+b_{j}\right)^{-1}=n \frac{N_{-, j}}{N}, \text { for all } j=1, \ldots, J-1 .
\end{gathered}
$$

At this stage, it is clear that if quotas are proportional (it should be remembered that this is the scenario almost systematically chosen in practice), the system is resolved by systematically choosing $a_{i}=1$ and $b_{j}=0$, in which case $\hat{P}_{i, j}=\frac{n_{i j}}{n}$ and $\hat{\bar{Y}}_{\text {quota }}=\bar{y}$, the customary sample mean. This provides a well-known result. The case of non-proportional quotas provides a complex estimator, whose analytical form is not explicit but asymptotically unbiased, since the specification of the model - not discussed in the founding article - is not in question. This point should prompt caution because, if proportional quota constraints are actually respected on average with a simple random selection, it seems to us that the sampling model becomes a priori more fragile when quotas significantly move away from proportional quotas, because of the fact that the second phase of this model is more difficult to accept. If the first phase of the model is probably the only way to build a theoretical basis for treating quotas, we could on the other hand - this would be another exercise - try to maximize the density of the $n_{i, j}$ under the quota constraints to obtain the $\hat{P}_{i, j}$.

Jean-Claude Deville proposed a variance expression for $\hat{\bar{Y}}_{\text {quota }}$, the sampling design being the only randomization mechanism that generates variability here. It begins by breaking down the true averages per cell $\bar{Y}_{i, j}$ according to $\bar{Y}_{i, j}=A_{i}+B_{j}+E_{i, j}$ by imposing the constraints $B_{J}=0$,

$$
\sum_{j=1}^{J} N_{i, j} E_{i, j}=0, \text { for all } i=1, \ldots, I
$$

and

$$
\sum_{i=1}^{I} N_{i, j} E_{i, j}=0, \text { for all } j=1, \ldots, J-1
$$

Then, he introduces the coefficients $a_{i}^{0}$ and $b_{j}^{0}$ ensuring the equalities

$$
E_{p}\left\{\frac{n_{i, j}}{n}\left(a_{i}+b_{j}\right)^{-1}\right\}=\frac{N_{i, j}}{N}\left(a_{i}^{0}+b_{j}^{0}\right)^{-1} .
$$

Indeed, $\frac{n_{i, j}}{n}\left(a_{i}+b_{j}\right)^{-1}$ is an estimator of $P_{i, j}=\frac{N_{i, j}}{N}$ but it is not unbiased of $P_{i, j}$ in general, and the coefficients $a_{i}^{0}$ and $b_{j}^{0}$ satisfy the equations of the system (S)

$$
\begin{gathered}
\sum_{j=1}^{J} N_{i, j}\left(a_{i}^{0}+b_{j}^{0}\right)^{-1}=N_{i,}, \text { for all } i=1, \ldots, I, \\
\sum_{i=1}^{I} N_{i, j}\left(a_{i}^{0}+b_{j}^{0}\right)^{-1}=N_{., j}, \text { for all } j=1, \ldots, J-1 .
\end{gathered}
$$

If quotas are proportional, then $a_{i}=1$ and $b_{j}=0$ and given the multinomial model, we also have $a_{i}^{0}=1$ and $b_{j}^{0}=0$ (these values are solutions of the previous non-linear system in all cases, but they are not the appropriate values for non-proportional quotas). By applying

$$
S_{i, j}^{2}=\frac{1}{N_{i, j}} \sum_{k \in(i, j)}\left(y_{k}-\bar{Y}_{i, j}\right)^{2},
$$

a variance expression is obtained:

$$
V\left(\hat{\bar{Y}}_{\text {quota }}\right)=\frac{1}{n} \sum_{(i, j)} \frac{N_{i, j}}{N}\left\{E_{i, j}^{2}+\left(a_{i}^{0}+b_{j}^{0}\right)^{-1} S_{i, j}^{2}\right\} .
$$

The optimal strategy is not one of proportional quotas, but one that "inflates" the quotas $n_{i .}$ and $n_{. j}$ for the modalities $i$ and $j$ corresponding to the high dispersions $S_{i, j}^{2}$. This rule, which is very intuitive, does not appear clearly in its principle if we stick to this variance calculation. On the other hand, it becomes evident when we condition on the sizes $n_{i, j}$.

When the quotas are proportional, we therefore have

$$
V\left(\hat{\bar{q}}_{\text {quota }}\right)=\frac{1}{n} \sum_{(i, j)} \frac{N_{i, j}}{N}\left(E_{i, j}^{2}+S_{i, j}^{2}\right) .
$$

It is clearly in our interest to have $E_{i, j}=0$ for all $(i, j)$, therefore an additive model. The article also proposes a variance estimator with low bias that is not complicated when quotas are proportional because it is obtained in this case from a classic residuals calculation in a certain standard linear regression.

### 6.3 Models for the variable of interest

This whole part is about taking a radically different view from the previous one, because this time the model pertains to the variable of interest $Y$. This variable is considered to be random using the model-based approach developed by Royall (1970, 1976a, b, 1988) (see also Valliant, Dorfman and Royall, 2000;

Chambers and Clark, 2012). All the model versions considered in the article fall within the general framework of the linear model, of the type $\mathbf{Y}=\mathbf{X B}+\mathbf{u}$, where $\mathbf{B}$ is a vector of unknown real coefficients, $\mathbf{X}$ a matrix whose columns are explanatory variables, and $\mathbf{u}$ is a vector of errors with a mean equal to 0 and an unknown model variance $\mathbf{V}$.

In this context, a vector of values $y_{k}, k=1, \ldots, N$, is composed of two vectors: $\mathbf{Y}_{s}$ of size $n$ that includes the observed values $y_{k}$-those for which $k \in s-$ and $\mathbf{Y}_{r}$ of size $N-n$ that includes the unobserved values $y_{k}$. To define an optimal estimator of the true total $t_{y}=\sum_{k \in U} y_{k}$, it is natural to try to minimize the mean square error $\mathrm{E}_{m}\left(\hat{t}-t_{y}\right)^{2}$, where $\mathrm{E}_{m}(\cdot)$ denotes the expectation under the randomization with respect to the values $y_{k}$ (which therefore has nothing to do with the sampling randomization). We also impose an unbiasedness condition of the form $\mathrm{E}_{m}\left(\hat{t}-t_{y}\right)=0$. Finally, we are seeking a simple estimator, therefore linear, i.e., of the form of $\hat{t}=\mathbf{g}_{s}^{\top} \mathbf{Y}_{s}$, where $\mathbf{g}_{s}$ is a vector (to be found) of size $n$. The solution to this problem leads to

$$
\hat{t}_{y, \mathrm{opti}}=\mathbf{1}_{s}^{\top} \mathbf{Y}_{s}+\mathbf{1}_{r}^{\top}\left\{\mathbf{X}_{r} \hat{\mathbf{B}}+\mathbf{V}_{r s} \mathbf{V}_{s s}^{-1}\left(\mathbf{Y}_{s}-\mathbf{X}_{s} \hat{\mathbf{B}}\right)\right\},
$$

where $\mathbf{1}_{s}$ (respectively $\mathbf{1}_{r}$ ) is a vector of size $n$ (respectively $N-n$ ) of 1's, the terms $\mathbf{X}_{s}, \mathbf{X}_{r}, \mathbf{V}_{s s}$ and $\mathbf{V}_{r s}$ representing the sub-matrices of $\mathbf{X}$ and $\mathbf{V}$ formed by the rows and columns associated with the sets of indices $s$ and $r$. The estimated coefficient $\hat{\mathbf{B}}$ is the general least squares estimator, i.e., $\hat{\mathbf{B}}=$ $\left(\mathbf{X}_{s}^{\top} \mathbf{V}_{s s}^{-1} \mathbf{X}_{s}\right)^{-1}\left(\mathbf{X}_{s}^{\top} \mathbf{V}_{s s}^{-1} \mathbf{Y}_{s}\right)$. We will place ourselves a priori in the case where $\mathbf{V}_{r s}=0$, a simplifying assumption that is easy to accept unless the selection involves clusters. As a result, the optimal estimator simplifies itself greatly since

$$
\hat{t}_{y, \text { opti }}=\mathbf{1}_{s}^{\top} \mathbf{Y}_{s}+\mathbf{1}_{r}^{\top}\left(\mathbf{X}_{r} \hat{\mathbf{B}}\right)=\sum_{k \in s} y_{k}+\sum_{k \in r} \hat{y}_{k},
$$

where $\hat{t}_{y k}$ is the $k^{\text {th }}$ coordinate of the vector $\mathbf{X}_{r} \hat{\mathbf{B}}$, that is, the optimal predictor of the unknown value $y_{k}$. It is this expression that will subsequently be used to express the estimators arising from the quota method. It should be noted that if 1 is one of the regressors (which is extremely common), we have $\sum_{k \in s} y_{k}=$ $\sum_{k \in s} \hat{y}_{k}$, so that $\hat{t}_{y, \text { opti }}=\sum_{k \in U} \hat{y}_{k}$.

The estimator $\hat{y}_{y, \text { opti }}$ is unbiased by construction in the sense that $\mathrm{E}_{m}\left(\hat{t}_{y, \text { opti }}-t_{y}\right)=0$. The assessment of its accuracy also relies on its variance, defined by $\mathrm{E}_{p} \mathrm{E}_{m}\left(\hat{t}_{y, \text { opti }}-t_{y}\right)^{2}$, where $\mathrm{E}_{p}(\cdot)$ the expectation with respect to the sampling design. It is also equal to $\mathrm{E}_{m} \mathrm{E}_{p}\left(\hat{t}_{y, \text { opti }}-t_{y}\right)^{2}$.

The combination of a sampling randomization and a model randomization is not easy to treat if the model on $y_{k}$ changes when we have information on the membership - or non-membership - of the individual $k$ in the selected sample. Otherwise, any calculation would become unmanageable, and moreover, we would not even be able to credibly formalize the model's dependence on the sample, which we recall, is random. Since it is not desirable (and not possible practically speaking) to seek to refine the assumptions beyond a certain degree, it is considered that the selection law of $s$ and the one that generates $Y$ are independent.

This is a non-informative model. Under these conditions, the model applied on $y_{k}$ applies blindly to any individual $k$ without there being any need to know if $k \in s$ or if $k \notin s$. This bias is essential to make variance calculations, but unfortunately, it is questionable. Typically, in the case of empirical surveys, one can have doubts, at least when the operation is not conducted under very tight control of the interviewers' practice, that there is no relationship between the values of $Y$ and whether they belong to the sample. This is what creates the main risk of bias in the estimators and what feeds the traditional criticism of empirical sampling. But let's move on from this risk, accept it, and summarize, under these conditions, the theory set out in Jean-Claude Deville's article (1991).

The case of quotas based on a single qualitative variable and the case of cross-quotas are based on the same linear model, which is very simple: noting $i$ the modality of the quota variable and $k$ the individual identifier, we have $y_{i, k}=m_{i}+u_{i, k}$ with $\mathrm{E}_{m}\left(u_{i, k}\right)=0$ and $\mathrm{E}_{m}\left(u_{i, k}^{2}\right)=\sigma_{i}^{2}$. This reflects the expected natural situation, where the qualitative variable (or the cross-classification of the two variables) explains the variable of interest well. The sampling technique makes it so that the sample sizes $n_{i}$ in each modality are independent of the selected sample $s$. We have $\hat{y}_{i, k}=\hat{m}_{i}=\bar{y}_{i}$, the simple average of $y_{i, k}$ calculated on individuals in the sample belonging to the modality $i$. In this case, it is easy to check that the optimal unbiased estimator is

$$
\hat{t}_{y, \mathrm{opti}}=\sum_{i=1}^{I} N_{i} \bar{y}_{i} .
$$

This is exactly the expression of the classical post-stratified estimator. It can be shown that

$$
\mathrm{E}_{m} \mathrm{E}_{p}\left(\hat{t}_{y, \text { opti }}-t_{y}\right)^{2}=\sum_{i=1}^{I} N_{i}^{2}\left(1-\frac{n_{i}}{N_{i}}\right) \frac{\sigma_{i}^{2}}{n_{i}}
$$

by using the fact that the $n_{i}$ 's are independent of $s$, and then obtain an unbiased estimate of this variance.
Now let's look at marginal quotas in the context of an additive model, that is, for an individual $k$ belonging to the modalities $i$ and $j$ of the two quota variables respectively: $y_{i, j, k}=\alpha_{i}+\beta_{j}+u_{i, j, k}$ with $\mathrm{E}_{m}\left(u_{i, j, k}\right)=0$ and $\mathrm{E}_{m}\left(u_{i, j, k}^{2}\right)=\sigma_{i}^{2}+\gamma_{j}^{2}$. The residuals are also considered to be mutually independent. The unbiased optimal estimator becomes

$$
\hat{t}_{y, o p \mathrm{opi}}=\sum_{i, j} n_{i, j} \bar{y}_{i, j}+\sum_{i, j}\left(N_{i, j}-n_{i, j}\right)\left(\hat{\alpha}_{i}+\hat{\beta}_{j}\right) .
$$

Since the estimators of $\alpha_{i}$ and $\beta_{j}$ are complicated to obtain in the general framework presented here, JeanClaude Deville suggests using the least squares estimators associated with an ordinary model. In this context, after writing out the normal equations, and thus easily obtaining the $\hat{\alpha}_{i}$ and $\hat{\beta}_{j}$, a bit of algebra leads to

$$
\hat{t}_{y, \text { opti }}=\sum_{i} N_{i .} \hat{\alpha}_{i}+\sum_{j} N_{. j} \hat{\beta}_{j} .
$$

In the case of proportional quotas, we check that $\hat{t}_{y, \text { opti }}=N \bar{y}$. This is the estimator that is used classically in quota sampling, which is often used due to its simplicity. In the case of non-proportional quotas, we can always express analytically $\hat{t}_{y, \text { opti }}$, particularly as a linear combination of simple estimators per cell $\bar{y}_{i, j}$, i.e., $\hat{t}_{y, \text { opti }}=\sum_{i, j} \hat{N}_{i, j} \bar{y}_{i, j}$, but the expressions of $\hat{N}_{i, j}$ are complex.

When quotas are proportional, the true variance $V\left(\hat{t}_{y, \text { opti }}\right)$ can be expressed in terms of $N_{i,}, N_{. j}, \sigma_{i}^{2}$ and $\gamma_{j}^{2}$, and a variance estimator can also be obtained as

$$
\hat{\mathrm{V}}\left(\hat{t}_{y, \text { opti }}\right)=N^{2} \frac{1-\frac{n}{N}}{n} \sum_{i, j} \frac{n_{i, j}}{n} s_{i, j}^{2},
$$

where $s_{i, j}^{2}$ is the standard dispersion of the $y_{i, j, k}$ 's collected in the cell $(i, j)$. It is unbiased in the sense that $\mathrm{E}_{p} \mathrm{E}_{m} \hat{\mathrm{~V}}\left(\hat{t}_{y, \text { opi }}\right)=\mathrm{V}\left(\hat{t}_{y, \text { opi }}\right)$. The feasibility of a variance calculation and its unbiased estimate is closely linked to the proportionality aspect of quotas: without this property, the development of the variance expression is no longer possible because the estimator $\hat{t}_{y, \text { opii }}$ depends on the sample sizes per cell $n_{i, j}$ and these are themselves depend on the drawn sample... based on an unknown probability, by definition of an empirical selection.

Finally, the article discusses the more general case of marginal quotas in a context where the model loses its additive aspect and becomes $y_{i, j, k}=\alpha_{i}+\beta_{j}+\gamma_{i, j}+u_{i, j, k}$. The optimal estimator is $\hat{t}_{y, \text { opti }}=\sum_{i, j} N_{i, j} \bar{y}_{i, j}$ but past this stage, we find ourselves in a context where not only do we not know $N_{i, j}$ but it does not even seem possible to produce a "natural" estimator. If the quotas are proportional, the practice - and clearly the only possible outcome - is to revert to the simple, standard expression $\hat{t}=N \bar{y}$. The price to pay is that of bias, which is

$$
\mathrm{E}_{p} \mathrm{E}_{m}\left(N \bar{y}-t_{y}\right)=\mathrm{E}_{p}\left\{\sum_{i, j}\left(N \frac{n_{i, j}}{n}-N_{i, j}\right) \gamma_{i, j}\right\} .
$$

There is no reason for it to be zero, but when the instructions are well designed and the collection meets them, we can hope that the ratios $n_{i, j} / n$ and $N_{i, j} / N$ will be numerically close, and therefore the bias will be small.

### 6.4 Operational conclusion

The conclusion that can be drawn on the treatment of quota sampling is as follows. If one uses a single quota variable or if one uses cross-classified quotas (which is the best!) by using multiple quota variables, there is no difficulty in performing point estimation without bias and then calculating a variance. For marginal quotas, with a sampling model, an estimator is obtained in all circumstances. It is asymptotically unbiased and its variance can always be estimated. With a model on the variable of interest, we do not know how to manage non-additive models. If the model is additive, an unbiased estimator can always be produced, but if the quotas are not proportional, then we don not know how to estimate the variance.

## 7. Conclusion

This article does not cover all of the original developments made by Jean-Claude Deville in the field of survey statistics. He has written numerous articles and made numerous presentations at conferences on other less significant aspects. As a result, his major works have themselves known a variety of fates. Calibration was certainly the most significant, best known and most widely used breakthrough worldwide, to the point that, today, there is no survey that is not calibrated on some external data, at least in the world of official statistics. The weight-share method is dependent on fairly specific survey conditions, but is nonetheless widely used, in particular, when repeated surveys are implemented over time and more specifically in the form of rotational sampling. The analytical development of variance expression for complex estimators regularly finds applications but relies on a rather complicated theory and has some serious competitors, namely the resampling methods. Balanced sampling has also had some success, but its use in official statistics is developing more slowly than calibration methods, probably because changing a sampling design is a much more consequential decision than changing an estimation procedure. Tillé (2011) already listed a list of applications 10 years after the method was developed. Interest in balanced sampling has since continued to grow. In France, the National Institute of Statistics and Economic Studies now makes a wide use of balanced sampling designs. As for the theory of quotas, this is mainly a clarification operation that helped somewhat demystify this technique that official statistics still looks at with some suspicion. JeanClaude Deville has undoubtedly touched upon it all, and with brilliance, something that the scientific community readily recognized by awarding him the prestigious Waksberg Prize in 2018, just three years before his death.

## References

Antal, E., Langel, M. and Tillé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. Invited Lecture at the $58^{\text {th }}$ Congress of the International Statistical Institute.

Bellhouse, D.R. (1988). A brief history of random sampling methods. Handbook of Statistics Volume 6: Sampling, (Eds., P.R. Krishnaiah and C.R. Rao), New York, Amsterdam. Elsevier/North-Holland, 1-14.

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

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. Journal of Official Statistics, 3, 141-153.

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

Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 34-42.

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. Survey Methodology, 22, 1, 17-22. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14389-eng.pdf.

Binder, D.A., and Kovačević, M.S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. Survey Methodology, 21, 2, 137-145. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1995002/article/14396-eng.pdf.

Boistard, H., Lopuhaä, H.P. and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. Electronic Journal of Statistics, 6, 1967-1983.

Bousabaa, A., Lieber, J. and Sirolli, R. (1999). La Macro Cube. Technical report, Rennes: ENSAI.
Brewer, K. (2013). Three controversies in the history of survey sampling. Survey Methodology, 39, 2, 249262. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11883eng.pdf.

Chambers, R.L., and Clark, R.G. (2012). An Introduction to Model-Based Survey Sampling with Applications. Oxford: Oxford University Press.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika, 95, 555-571.

Chauvet, G. (2009). Stratified balanced sampling. Survey Methodology, 35, 1, 115-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10888-eng.pdf.

Chauvet, G., Bonnéry, D. and Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. Journal of Statistical Planning and Inference, 141(2), 984-994.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. Biometrika, 98, 459-471.

Chauvet, G., Haziza, D. and Lesage, É. (2015). Examining some aspects of balanced sampling in surveys. Statistica Sinica, 25, 313-334.

Chauvet, G., and Tillé, Y. (2006a). A fast algorithm of balanced sampling. Journal of Computational Statistics, 21, 9-31.

Chauvet, G., and Tillé, Y. (2006b). Fastcube SAS-IML Macro. Université de Neuchâtel.
Chopin, N., and Ducrocq, G. (2021). Fast compression of MCMC output. Entropy, 23(8).

Choudhry, G.H., and Singh, M.P. (1979). Sampling with unequal probabilities and without replacement A rejective method. Survey Methodology, 5, 2, 162-177. Paper available at https://www150.statcan.gc.ca/ n1/en/pub/12-001-x/1979002/article/54834-eng.pdf.

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Statistics and Probability Letters, 18(5), 353-362.

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

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. Survey Methodology, 30, 1, 17-26. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004001/article/6991-eng.pdf.

Demnati, A., and Rao, J.N.K. (2010). Linearization variance estimators for model parameters from complex survey data. Survey Methodology, 36, 2, 193-201. Paper available at https://www150.statcan.gc.ca/n1/en/ pub/12-001-x/2010002/article/1 1381-eng.pdf.

Devaud, D., and Tillé, Y. (2019). Deville and Särndal's calibration: Revisiting a 25 years old successful optimization problem. TEST, 4, 1033-1065.

Deville, J.-C. (nda). $15^{\text {ème }}$ round. cette fois y'a vraiment qêqchose. Manuscript internal note, Paris: Insee.

Deville, J.-C. (ndb). Comparaison des plans à probabilités inégales avec and sans remise. Manuscript internal note, Paris: Insee.

Deville, J.-C. (ndc). Échantillonnage à entropie max (rédaction rapide). Manuscript internal note, Paris: Insee.

Deville, J.-C. (1989). Une théorie simplifiée des sondages. Les Ménages: Mélanges en L'honneur de Jacques Desabie, (Eds., P. L’Hardy and C. Thélot), Paris: Insee, 191-214.

Deville, J.-C. (1991). A theory of quota surveys. Survey Methodology, 17, 2, 163-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14504-eng.pdf.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys, Örebro. Statistics Sweden, 1-18.

Deville, J.-C. (1998a). La correction de la nonréponse par calage ou par échantillonnage équilibré. Actes $d u$ Colloque de la Société Statistique du Canada, Sherbrooke, Canada.

Deville, J.-C. (1998b). Les enquêtes par panel : en quoi différent-elles des autres enquêtes? suivi de: Comment attraper une population en se servant d'une autre. Actes des Journées de Méthodologie Statistique, INSEE Méthodes, Paris: Insee, No. 84-85-86, 63-82.

Deville, J.-C. (1998c). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Working paper no 9804, Statistical Methodology, Paris: Insee.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey Methodology, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n 1/en/ pub/12-001-x/1999002/article/4882-eng.pdf.

Deville, J.-C. (2000a). Generalized calibration and application to weighting for non-response. CompStat, Proceedings in Computational Statistics: $14^{\text {th }}$ Symposium Held in Utrecht, The Netherlands, New York: Springer, 65-76.

Deville, J.-C. (2000b). Note sur l'algorithme de Chen, Dempster et Liu. Manuscript internal note, Rennes: CREST-ENSAI.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. Actes des Journées de Méthodologie Statistique, Paris: Insee-Méthodes, 3-20.

Deville, J.-C. (2014). Échantillonnage équilibré exact poissonnien. $8^{\text {ème }}$ Colloque Francophone sur les Sondages, Université de Bourgogne, Dijon, 1-6.

Deville, J.-C. (2015). Quelques éléments de géométrie et d'algèbre pour comprendre la nature d'un échantillonnage équilibré. Actes des Journées de Méthodologie Statistique, Paris: Insee-Méthodes, 1-8.

Deville, J.-C., and Dupont, F. (1993). Non-réponse : principes et méthodes. Actes des Journées de Méthodologie Statistique, Paris: Insee-Méthodes, 53-69.

Deville, J.-C., and Grosbras, J.-M. (1987). Algorithmes de tirage. Les Sondages, (Eds., J.-J. Droesbeke, B. Fichet and P. Tassi), Paris: Economica, 209-233.

Deville, J.-C., Grosbras, J.-M. and Roth, N. (1988). Efficient sampling algorithms and balanced sample. CompStat, Proceedings in Computational Statistics $8^{\text {th }}$ Symposium held in Copenhagen, Heidelberg: Physica Verlag, 255-266.

Deville, J.-C., and Jacod, M. (1996). Replacing the traditional French census by a large scale continuous population survey. Annual Research Conference. US Bureau of Census.

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.

Deville, J.-C., and Maumy-Bertrand, M. (2006). Extension of the indirect sampling method and its application to tourism. Survey Methodology, 32, 2, 177-185. Paper available at https://www150.statcan. gc.ca/n1/en/pub/12-001-x/2006002/article/9552-eng.pdf.

Deville, J.-C., and Qualité, L. (2005). Échantillonnage multidimensionnel (de plusieurs échantillons à la fois) à entropie maximum : définition, propriétés, algorithmes et programmes. Actes des Journées de Méthodologie Statistique, Paris: Insee-Méthodes, 1-8.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87(418), 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling. Journal of the American Statistical Association, 88, 1013-1020.

Deville, J.-C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. Biometrika, 85, 89-101.

Deville, J.-C., and Tillé, Y. (2000a). Échantillonnage équilibré par la méthode du cube, variance et estimation de variance. Actes des Journées de Méthodologie Statistique, Paris: Insee-Méthodes, 15-35.

Deville, J.-C., and Tillé, Y. (2000b). Selection of several unequal probability samples from the same population. Journal of Statistical Planning and Inference, 86, 215-227.

Deville, J.-C., and Tillé, Y. (2001). Échantillonnage équilibré par la méthode du cube, variance et estimation de variance. Enquêtes, Modèles et Applications, (Eds., J.-J. Droesbeke and L. Lebart), Paris: Dunod, 444362.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. Biometrika, 91, 893912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. Journal of Statistical Planning and Inference, 128, 569-591.

Dupačová, J. (1979). A note on rejective sampling. Contribution to Statistics (Jaroslav Hájek Memorial Volume), Academia Prague, 71-78.

Dupont, F. (1993). Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. Actes des Journées de Méthodologie Statistique, 15-16 December, 1993, Insee-Méthodes No 56-57-58, Complement, 9-42.

Durr, J.-M., and Dumais, J. (2002). Redesign of the french census of population. Survey Methodology, 28, 1, 43-49. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002001/article/6414eng.pdf.

Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. Panel Surveys, (Eds., D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh), New York: John Wiley \& Sons, Inc., 135-159.

Ernst, L.R., Hubble, D.L. and Judkins, D.R. (1984). Longitudinal family and household estimation in sipp. Proceedings of the Survey Research Methods Section, American Statistical Association, 682-687.

Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in twophase sampling. Journal of Official Statistics, 18, 233-255.

Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. International Statistical Review, 74, 127-147.

Eustache, E., Vallée, A.-A. and Tillé, Y. (2022). Balanced donor imputation handling Swiss cheese nonresponse. Statistica Sinica, accepted.

Falorsi, P., Piersante, A. and Bako, B. (2016). Indirect sampling, a way to overcome the weakness of the lists in agricultural survey. Proceedings ICAS VII: Seventh International Conference on Agricultural Statistics, Rome, 24-26 October, 2016.

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. The Annals of Statistics, 454-469.

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

Fuller, W.A., Legg, J.C. and Li, Y. (2017). Bootstrap variance estimation for rejective sampling. Journal of the American Statistical Association, 112(520), 1562-1570.

Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). Annali di Statistica, Series 6, 4, 1-107.

Graf, É., and Tillé, Y. (2014). Variance estimation using linearization for poverty and social exclusion indicators. Survey Methodology, 40, 1, 61-79. Paper available at https://www150.statcan.gc.ca/n1/en/ pub/12-001-x/2014001/article/14000-eng.pdf.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Compositional Data Analysis: Theory and Applications, (Eds., V. Pawlowsky-Glahn and A. Buccianti), Chichester: Wiley, 114-127.

Grafström, A., and Lisic, J. (2019). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.5.

Grafström, A., Lundström, N.L.P. and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. Biometrics, 68(2), 514-520.

Grafström, A., and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics, 14(2), 120-131.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics, 35, 1491-1523.

Hájek, J. (1981). Sampling from a Finite Population. New York: Marcel Dekker.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. and Stahel, W.A. (1985). Robust Statistics: The Approach Based on the Influence Function. New York: John Wiley \& Sons, Inc.

Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. Computational Statistics and Data Analysis, 74, 81-94.

Hasler, C., and Tillé, Y. (2016). Balanced $k$-nearest neighbor imputation. Statistics, 105, 11-23.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. Statistical Science, 32(2), 206-226.

Haziza, D,. and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics, 32(1), 129-145.

Holt, D., and Smith, T.M.F. (1979). Post-stratification. Journal of the Royal Statistical Society, A142: Part 1, 33-46.

Huang, E.T., and Fuller, W.A. (1978). Non-negative regression estimation for sample survey data. Proceedings of the Social Statistics Section, American Statistical Association, 300-305.

Huang, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the income survey development program. Proceedings of the Social Statistics Section, American Statistical Association, 670-675.

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

Jauslin, R., Eustache, E., Panahbehagh, B. and Tillé, Y. (2021). StratifiedSampling: Different Methods for Stratified Sampling. R Foundation for Statistical Computing, Vienna, Autricha. R package version 0.3.0.

Jauslin, R., Eustache, E. and Tillé, Y. (2021). Enhanced cube implementation for highly stratified population. Japanese Journal of Statistics and Data Science, 4, 783-795.

Jauslin, R., and Tillé, Y. (2020a). Spatial spread sampling using weakly associated vectors. Journal of Agricultural, Biological and Environmental Statistics, 25(3), 431-451.

Jauslin, R., and Tillé, Y. (2020b). WaveSampling: Weakly Associated Vectors Sampling. R package version 0.1.1 http://CRAN.R-project.org/package=WaveSamplinghttp://CRAN.R-project.org/package=WaveSampling.

Judkins, D., Hubble, D., Dorsch, J., McMillen, D. and Ernst, L. (1984). Weighting of persons for sipp longitudinal tabulations. Proceedings of the Social Statistics Section, American Statistical Association, 676-687.

Kalton, G., and Brick, J. (1995). Weighting schemes for household panel surveys. Survey Methodology, 21, 1, 33-44. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1995001/article/14412eng.pdf.

Kiær, A.N. (1896). Observations et expériences concernant des dénombrements représentatifs. Bulletin de l'Institut International de Statistique, 9, 176-183.

Kiær, A.N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. Bulletin de l'Institut International de Statistique, 11, 180-185.

Kiær, A.N. (1903). Sur les méthodes représentatives ou typologiques. Bulletin de l'Institut International de Statistique, 13, 66-78.

Kiær, A.N. (1905). Discours sans intitulé sur la méthode représentative. Bulletin de l'Institut International de Statistique, 14, 119-134.

Kiesl, H. (2010). Selecting kindergarten children by three stage indirect sampling. Proceedings of the Survey Research Methods Section, American Statistical Association, 2730-2738.

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

Kott, P.S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. Journal of the American Statistical Association, 105(491), 1265-1275.

Langel, M., and Tillé, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. Metron, 69, 45-65.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. Survey Methodology, 21, 1, 25-32. Paper available at https://www150.statcan. gc.ca/n1/en/pub/12-001-x/1995001/article/14413-eng.pdf.

Lavallée, P. (2001). La méthode généralisée du partage des poids et le calage sur marges. Enquêtes, Modèles et Applications, (Eds., J.-J. Droesbeke and L. Lebart), Paris: Dunod, 396-403.

Lavallée, P. (2002). Le Sondage Indirect ou la Méthode Généralisée du Partage des Poids. (Ed., l'Université de Bruxelles, Paris: Ellipses.

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

Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. Survey Methodology, 36, 1, 69-79. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/ 11249-eng.pdf.

Lemel, Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondages. Annales de l'Insee, 22-23, 273-281.

Lesage, É., Haziza, D. and D'Haultfoeuille, X. (2019). A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. Journal of the American Statistical Association, 114, 906-915.

Leuenberger, M., Eustache, E., Jauslin, R. and Tillé, Y. (2022). Balancing a sample almost perfectly. Statistics and Probablity Letters, 180, 109229.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics, 15, 305-327.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. and Puech, P. (2023). Many-toone indirect sampling with application to the french postal traffc estimation. The Annals of Applied Statistics, 17(1), 838-859.

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

Nedyalkova, D., and Tillé, Y. (2012). Bias robustness and efficiency in model-based inference. Statistica Sinica, 22, 777-794.

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

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57, 377-387.

Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. Biometrika, 63, 605-614.

Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. Journal of the American Statistical Association, 71, 657-664.

Royall, R.M. (1988). The prediction approach to sampling theory. Handbook of Statistics, Sampling, (Eds., P.R. Krishnaiah and C.R. Rao), Amsterdam: Elsevier, Volume 6, 399-413.

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

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. Survey Methodology, 33, 2, 99-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488eng.pdf.

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

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

Sautory, O., and Le Guennec, J. (2003). La macro CALMAR2 : redressement d'un échantillon par calage sur marges - documentation de l'utilisateur. Technical report, Paris: Insee.

Stephan, F.F. (1942). An iterative method of adjusting sample frequency data tables when expected marginal totals are known. Annals of Mathematical Statistics, 13, 166-178.

Thionet, P. (1953). La Théorie des Sondages. Institut National de la Statistique et des Études Économiques, Études théoriques vol. 5, Paris: Imprimerie nationale.

Thomsen, I. (1978). A second note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. Statistisk Tidskrift, 16, 278-285.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: An appraisal. Survey Methodology, 37, 2, 215-226. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/ 11609-eng.pdf.

Tillé, Y. (2016). The legacy of Corrado Gini in survey sampling and inequality theory. Metron, 74(2), 167174.

Tillé, Y., and Matei, A. (2021). Sampling: Survey Sampling. R package version 2.9.

Vallée, A.-A., and Tillé, Y. (2019). Linearization for variance estimation by means of sampling indicators: Application to nonresponse. International Statistical Review, 87(2), 347-367.

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

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer.
Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association, 66, 411-414.

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.

Yates, F. (1949). Sampling Methods for Censuses and Surveys. London: Charles Griffin.

Zhang, S., Han, P. and Wu, C. (2022). Calibration techniques encompassing survey sampling, missing data analysis and causal inference. International Statistical Review, accepted for publication.

# Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" 

Guillaume Chauvet ${ }^{1}$


#### Abstract

In this discussion, I will present some additional aspects of three major areas of survey theory developed or studied by Jean-Claude Deville: calibration, balanced sampling and the generalized weight-share method.


Key Words: Calibration; Balanced sampling; Weight share.

## 1. Introduction

In this discussion, I will provide some additional information on calibration in Section 2, on the applications of balanced sampling in Section 3 and on the weight-share method in Section 4. In Section 5, I will conclude with some more personal details about my work with Jean-Claude and his influence on my research.

## 2. Calibration

The unification of adjustment methods on auxiliary information in the form of calibration estimators (Deville and Särndal, 1992; Deville, Särndal and Sautory, 1993) is certainly Jean-Claude Deville's most important contribution. Let $U$ be a population of size $N$. For a unit $k \in U$, let $\mathbf{x}_{k}$ be a $J$-vector of auxiliary variables of the total $t_{\mathbf{x}}=\sum_{k \in U} \mathbf{X}_{k}$. For a sample $S$ with sampling weights $d_{k}=1 / \pi_{k}$, Deville and Särndal (1992) showed that finding new weights $w_{k}$, as close as possible to the weights $d_{k}$ that satisfy the calibration constraints on $t_{x}$, leads to the solution

$$
\begin{equation*}
w_{k}=d_{k} F\left(q_{k} \hat{\lambda} \mathbf{x}_{k}\right), \tag{2.1}
\end{equation*}
$$

with $F(\cdot)$ a distance-dependent calibration function, $\hat{\lambda}$ a parameter adjustment vector and $q_{k}$ a scaling factor. In most cases, $q_{k}$ is chosen equal to 1 , but with $J=1$ and a single positive auxiliary variable $x_{k}$, the choice of $q_{k}=x_{k}^{-1}$ finds the ratio estimator with any function $F(\cdot)$.

The choice of the calibration function $F(x)=1+x$ (linear method) finds the generalized regression estimator (GREG). Deville and Särndal (1992, Result 5) showed that, for a general calibration function $F(\cdot)$, the calibration estimator $\hat{t}_{y, C}$ obtained is asymptotically equivalent to the GREG estimator and shares its properties: negligible bias and same asymptotic variance. However, it should be noted that this result requires a few assumptions about the function $F(\cdot)$. In particular, it must verify $F(0)=F^{\prime}(0)=1$, which
guarantees that its first-order Taylor expansion is the same as in the linear method (Deville and Särndal, 1992, Result 4).

A remarkable aspect of calibration estimators is that they require very little auxiliary information. The auxiliary variables $\mathbf{x}_{k}$ must be observed in the sample, and only the population totals $t_{\mathbf{x}}$ must be known. Even auxiliary totals estimated using a baseline survey can be used-see for example Renssen and Nieuwenbroek (1997); Rancourt (2001); Berger, Muñoz and Rancourt (2009); Dever and Valliant (2010). Using estimated totals instead of true totals leads to increased variance, which remains limited if the baseline survey is much larger (Dever and Valliant, 2016). With well-chosen auxiliary variables, this increase may be smaller than the benefit linked to calibration (e.g., Ceccarelli and Guandalini, 2014).

Calibration remains a very active area of research. In the case of many auxiliary variables, it is possible to modify the calibration equation so that certain constraints are only approximately met. This principle has led to the development of ridge calibration methods (Chambers, 1996), principal component calibration (Cardot, Goga and Shehzad, 2017) or penalized calibration (Guggemos and Tillé, 2010); see also Breidt and Opsomer (2017) for a review. In this large-scale case, Chauvet and Goga (2022) have also proposed a bootstrap criterion for choosing calibration variables.

## 3. Balanced sampling

The cube method is one of the finest technical innovations in survey theory in the last 25 years. A sample design is said to be balanced on a $q$-vector of auxiliary variables $\mathbf{z}_{k}$, if all samples $S$ with a non-zero probability of selection satisfy balancing condition

$$
\begin{equation*}
\sum_{k \in S} d_{k} \mathbf{z}_{k}=\sum_{k \in U} \mathbf{z}_{k} \tag{3.1}
\end{equation*}
$$

The cube method (Deville and Tillé, 2004), which selects (approximately) balanced samples, is based on two ingenious innovations. One is the splitting method, which breaks down a sample design into simpler designs. The other is a geometric representation that allows us to see a balanced sampling step as a random walk in a hyperplane of dimension $N-q$.

According to the equation (3.1), balancing can be viewed as a calibration that is integrated into the sampling design, without having to modify the weights $d_{k}$. The drawback is that the variables $\mathbf{z}_{k}$ must be known for each individual in the population, while the calibration would only require knowledge of the auxiliary totals $t_{\mathbf{z}}$. Another practical problem is that balancing is generally destroyed by unit non-response. For this reason, the method is particularly interesting in a sampling context with low non-response, like for the selection of primary units for multi-stage sampling; see, for example, Costa, Guillo, Paliod, Merly-Alpa, Vincent, Chevalier and Deroyon (2018) for the sample design associated with the selection of the NAUTILE master sample from the Institut national de la statistique et des études économiques.

Balanced sampling is also very useful for dealing with item non-response. Suppose that in sample $S$, a variable of interest $y_{k}$ is only observed in a sub-sample $S_{r}$, and missing in additional sample $S_{m}$. The imputed estimator of the total is written as

$$
\hat{t}_{y I}=\sum_{k \in S_{r}} d_{k} y_{k}+\sum_{k \in S_{m}} d_{k} y_{k}^{*},
$$

with $y_{k}^{*}$ an imputed value for $k \in S_{m}$. Imputation can be based on the model

$$
\begin{equation*}
m: y_{k}=f\left(\mathbf{z}_{0 k}^{\top} \beta\right)+\epsilon_{k}, \tag{3.2}
\end{equation*}
$$

with $f(\cdot)$ a known function, $\mathbf{z}_{0 k}$ a vector of known variables for any $k \in S, \epsilon_{k}$ a random noise and $\beta$ a vector of parameters to be estimated. The imputed value can be generated using the random imputation mechanism

$$
\begin{equation*}
I: y_{k}^{*}=f\left(\mathbf{z}_{0 k}^{\top} \hat{\beta}\right)+\epsilon_{k}^{*}, \tag{3.3}
\end{equation*}
$$

with $\hat{\beta}$ an estimator of $\beta$, and $\epsilon_{k}^{*}$, a random term obtained by randomly drawing from the residuals $e_{k}=y_{k}-f\left(\mathbf{z}_{0 k}^{\top} \hat{\beta}\right)$ observed for $k \in S_{r}$. This imputation mechanism preserves the distribution of $y_{k}$ at the cost of an additional variance for the imputed estimator $\hat{t}_{y I}$ because of the random term $\sum_{k \in S_{m}} d_{k} \epsilon_{k}^{*}$. A modification of the Cube Method (Chauvet, Deville and Haziza, 2011) allows sampling among residuals $e_{k}, k \in S_{r}$, ensuring that the variability of $\sum_{k \in S_{m}} d_{k} \epsilon_{k}^{*}$ is (almost) zero. The distribution of the imputed variable is maintained, while avoiding an increase in variance for $\hat{y}_{y l}$. Most random imputation methods allow a balanced version, which has the great advantage of reducing variance without requiring additional information.

## 4. Weight-share method

The generalized weight-share method provides an elegant and practical solution for surveying a population $U^{B}$ using another population $U^{A}$, for which there is a sampling frame. As the authors pointed out, this is an essential tool for producing cross-sectional estimates in longitudinal surveys, where households present at date $t+1$ are captured using individuals selected at time $t$ and tracked over time (e.g., Ardilly and Lavallée, 2007).

The key element of the method lies in the ability to transform any variable $y^{B}$ from $U^{B}$ into a synthetic variable $y^{A}$ from $U^{A}$, with the same total. Let $l_{j k}$ be the link variable between the two populations, equal to 1 if units $j \in U^{A}$ and $k \in U^{B}$ are linked, and 0 otherwise. Let $L_{k}^{A}=\sum_{j \in U^{A}} l_{j k}$ be the total number of links between $k$ and $U^{A}$, assumed to be $>0$ for all $k \in U^{A}$. For any $j \in U^{A}$, the synthetic variable is written as

$$
\begin{equation*}
y_{j}^{A}=\sum_{k \in U^{B}} \frac{l_{j k} y_{k}^{B}}{L_{k}^{B}}, \tag{4.1}
\end{equation*}
$$

which means that for any $k \in U^{B}$, each $y_{k}^{B}$ is divided equally between all units in $U^{A}$ linked to it. The conservation property of the total $\sum_{k \in U^{B}} y_{k}^{B}=\sum_{j \in U^{A}} y_{j}^{A}$ allows for using a Horvitz-Thompson estimator on $U^{A}$, which simplifies variance estimation in particular.

The weight-share method is traditionally used for discrete populations (e.g., individuals, households, businesses). The same principle applies to more complex cases, where the survey population and the sample population are different in nature. In forest inventories (e.g., Gregoire and Valentine, 2007), the usual practice is to select a sample of points in a continuous universe $\mathcal{U}^{A}$ and use plots at these points to capture trees, which constitute the (discrete) population of interest $U^{B}$. To switch to a total estimate of $U^{B}$, a variable $y^{B}$ can be transported on $\mathcal{U}^{A}$ in a so-called local density variable, following the same principle used in equation (4.1); see Stevens and Urquhart (2000) and Mandallaz (2007). Chauvet, Bouriaud and Brion (2023) have shown that various methods proposed in the forest inventory literature can be seen as stemming from an extension of the weight share method.

## 5. In conclusion

My first collaboration with Jean-Claude was on my doctoral thesis, which I wrote under his supervision between 2004 and 2007. The thesis dealt with the use of bootstrap methods in survey theory, and more specifically with the Gross method (1980), which Jean-Claude had already considered in a review article on replicate-based variance estimation techniques (Deville, 1987). As part of my thesis, I studied the application of the Gross method to sample designs with unequal probabilities (including Poisson and rejective sampling), balanced designs and multi-stage sampling.

Ultimately, I only co-wrote four articles with Jean-Claude, which isn't very many given the number of years we spent together at the Laboratoire de Statistique d'Enquête. However, my research was strongly influenced by our discussions. A significant part of my work deals with balanced sampling methods, and more specifically the theoretical properties and applications of the pivotal method. My thesis work on bootstrap left me with a guilty pleasure for variance estimation methods, as much for analytical techniques as for bootstrap. I have had the opportunity to use the linearization approach on several occasions in the epidemiological field, for estimating treatment effects using inverse probability weighted estimators, very similar in essence to the estimators used to deal with unit non-response. The link between the weight-share method and forestry methods is a new discovery for me and has inspired great research prospects. The methods developed by Jean-Claude still have a bright future ahead of them.

## Acknowledgements

My sincere thanks to Jean-François Beaumont for offering to write a discussion of the article by Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé, tracing Jean-Claude Deville's contributions in five areas of survey theory. As the authors have pointed out, Jean-Claude has made major contributions that have become standard practice at national statistical institutes.

## References

Ardilly, P., and Lavallée, P. (2007). Weighting in rotating samples: The SILC survey in France. Survey Methodology, 33, 2, 131-137. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10490-eng.pdf.

Berger, Y.G., Muñoz, J.F. and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals - An application to the extended regression estimator and the regression composite estimator. Computational Statistics \& Data Analysis, 53(7), 2596-2604.

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. Statistical Science, 32(2), 190-205.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. Statistica Sinica, 243-260.

Ceccarelli, C., and Guandalini, A. (2014). Increasing the accuracy of it-silc estimates through the use of auxiliary information from labour force survey. Italian Journal of Applied Statistics, 24(1), 103-115.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. Journal of Official Statistics, 12, 3-32.

Chauvet, G., Bouriaud, O. and Brion, P. (2024). An extension of the weight share method when using a continuous sampling frame. Survey Methodology, forthcoming.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. Biometrika, 98(2), 459-471.

Chauvet, G., and Goga, C. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. Journal of Statistical Planning and Inference, 217, 177-187.

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

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

Dever, J.A., and Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. Journal of Survey Statistics and Methodology, 4(3), 289-318.

Deville, J.-C. (1987). Réplications d'échantillons, demi-échantillons, jackknife, bootstrap. Les Sondages, (Eds., J.-J. Droesbeke, B. Fichet and P. Tassi), Paris: Economica, 147-171.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87(418), 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. Journal of the American Statistical Association, 88(423), 1013-1020.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. Biometrika, 91(4), 893912.

Gregoire, T.G., and Valentine, H.T. (2007). Sampling Strategies for Natural Resources and the Environment. CRC Press.

Gross, S.T. (1980). Median estimation in sample surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 181-184.

Guggemos, F., and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. Journal of Statistical Planning and Inference, 140(11), 3199-3212.

Mandallaz, D. (2007). Sampling Techniques for Forest Inventories. CRC Press.

Rancourt, E. (2001). La régression étendue: un ensemble de pratiques d'estimation qui poussent constamment la théorie. Recueil Enquêtes, Modèles et Applications, Dunod, 334-343.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. Journal of the American Statistical Association, 92(437), 368-374.

Stevens, J., and Urquhart, N. (2000). Response designs and support regions in sampling continuous domains. Environmetrics, 11(1), 13-41.

# Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" 

Marc Christine ${ }^{1}$


#### Abstract

This article discusses and provides comments on the Ardilly, Haziza, Lavallée and Tille's summary presentation of Jean-Claude Deville's work on survey theory. It sheds light on the context, applications and uses of his findings, and shows how these have become engrained in the role of statisticians, in which Jean-Claude was a trailblazer. It also discusses other aspects of his career and his creative inventions.


Key Words: Sampling; Calibration; Balancing; Weight sharing; Quota; Statistical methodology; Statistician.

The announcement of Jean-Claude Deville's death in October 2021 was met with great emotion at the Institut national de la statistique et des études économiques (INSEE) and the Service statistique public (SSP) (France's official statistical system), as well as in international circles (e.g., Canada, European partners). He was a great personality - both personally and intellectually - and his colleagues were deeply saddened by his death, but his body of work speaks to his outstanding qualities. Yves Tillé compiled a bibliography of Jean-Claude Deville's work, available on INSEE's Journées de méthodologie statistique website (French only), which also contains the proceedings of all the Journées from 1991 to 2022, including the papers mentioned in this article.

Ardilly, Haziza, Lavallée and Tille's article is a remarkable account of some of Jean-Claude Deville's major innovations in survey theory and statistical survey practice, particularly at the SSP, but also internationally - balanced sampling, calibration, weight sharing, calculation of complex estimator variance, quota sampling.

For those unfamiliar with Jean-Claude, the article seems to suggest that his career was essentially devoted to developing advanced fields of theoretical investigation, complex mathematical formulations and highly specialized research. However, as mentioned in the introduction, it is important to remember that Jean-Claude Deville worked at INSEE for several decades, part of which he spent working only with the Statistical Methods Unit, where survey questions had a key role. After this, he began working at the survey statistics laboratory of the École nationale de la statistique et de l'analyse de l'information (France's National School of Statistics and Information Analysis) in Rennes, France.

All this to say that he was driven by the need to find practical solutions to the real problems he or other statisticians faced in their day-to-day work, without necessarily knowing how to solve them. He also had a remarkable ability to describe and broaden the scope of the problems brought to his attention as well as to establish links between different disciplines.

Jean-Claude Deville also worked on subjects other than those strictly related to survey theory, including demography and data analysis - further proof of his diverse research. He was also the founder of and resident theoretical expert on a major overhaul of the French census system, transforming it into a continuous sample census that was implemented in 2004. Although this article does not discuss his work, it is worth remembering.

He also taught at the École nationale de la statistique et de l'analyse économique (France's National School of Statistics and Economic Analysis) (e.g., second-order random processes, in-depth surveys).

Ardilly, Haziza, Lavallée and Tillé produced an eloquent, yet highly detailed summary of the technical aspects of Jean-Claude Deville's work. This discussion cannot begin without first highlighting the quality and comprehensive nature of their article.

In addition, what follows is not intended as a commentary on the quality of this general article or on the intrinsic benefit of Jean-Claude Deville's work, but rather as a way to shed light on some of his work to recontextualize it according to the daily questions of the statisticians who gave rise to it or who applied it.

## A man and his career

Discussing Jean-Claude Deville's work using the overview presented in the article also means discussing his day-to-day work as a statistician, which - as mentioned above - fuelled his theoretical reflections.

Here are a few examples that can help to remember the variety of his work and his pragmatism. He was an expert on many subjects. The following is a random list drawn from my personal experience:

Using the Vacances des Français survey (survey on vacations taken by the French), he helped answer a question from a group of Breton communes, which wanted to show that tourism in their areas had dropped following the sinking of the Amoco-Cadiz in 1978 and the resulting oil spill, with the aim of obtaining compensation in criminal court (1984).

In 1985, for the CAMME project (transforming the household economic survey into a monthly survey), he devised a method for drawing the telephone sample from the paper directory - the only one available at the time.

From 1993 to 1998, he worked extensively to transform the French employment survey - which was initially annual - into a quarterly, then ultimately monthly, survey. He was disappointed that not all his recommendations were followed, but he did support and defend his colleagues, which was appreciated. For example, when a service provider with no subject-matter knowledge wanted to get involved ("I appreciate your vigilant tone", he told me when I protested against this outside initiative).

He was the architect of the modernized (continuous) census of population (1999-2000) and led a working group where he enjoyed participating in "spirited" scientific debate.

He closely followed the OCTOPUSSE project (Organisation Coordonnée du Tirage Optimisé Pour Une mise en œuvre StatiStique des Échantillons, i.e., France's coordinated optimized selection for statistical sample processing project) to build the new master sample from the continuous census, while ironically
illustrating his concern (in the form of a "parable of local leaders") that INSEE's regional directors would be asked for their opinion on the samples once they had been selected (2009).

## Calibration

This corpus should be first because, on the one hand, these studies are now more than 30 years old and, on the other, it could be said that calibration procedures are the most popular theories in that they have become both commonplace and widespread in most statistical surveys. In addition, of the studies cited in Ardilly, Haziza, Lavallée and Tillé, it is undoubtedly the oldest theory that has been developed, implemented and normalized in tools, such as the SAS CALMAR macro.

The value of calibration is well established - not only does it reduce estimator variance, but it also ensures the coherence of estimated totals based on a statistical survey with external data deemed to be accurate and indisputable, on a certain number of sociodemographic, geographic and economic reference variables.

To use it requires only available auxiliary information. Such information exists in most statistical systems and is usually correlated with survey variables of interest, but the calibration variables and those observed in the survey need to be conceptually coherent.

However, one downside is the apparent ease with which the technique can be used and the temptation to calibrate to "everything" available. The price to pay is a more delicate adjustment of calibration weights and the risk of obtaining outlier weights, hence the need to limit the weights or, more specifically, the ratio between calibration weights and initial survey weights. It is also important to consider the definition of the most relevant calibration variables, without trying to multiply them, as there is a risk of introducing random collinearity.

For example, at INSEE, the "Groupe Marges" (calibration working group) is working to define a policy on survey calibration methods, both in terms of the variables to use - whether mandatory or optional - and the sources best suited to providing reference margins.

The method is so popular that it can be used "without weights": empirical data can be calibrated to totals to make them "more representative" of the reference universe, without knowing the exact process of origin, then it is implicitly agreed to act as if there were initial survey weights that are all equal.

Calibration can also be done on "nothing", if no auxiliary information is available, which is rare. Generally, there are two dual formulas to keep in mind:

$$
\left\{\begin{array}{c}
\sum_{i=1}^{N} \pi_{i}=E(\operatorname{card} s) \\
E\left(\sum_{i \in s} \frac{1}{\pi_{i}}\right)=N
\end{array}\right.
$$

for a sample $s$ selected from a population of size $N$, with inclusion probabilities $\pi_{i}$. At the very least, we can therefore calibrate to the size of the population to have weights that add up exactly to this size.

Calibration procedures are used when there are several levels of nested statistical units, e.g., individualshouseholds, establishments-businesses, when calibration data on the different levels are available and when it is necessary to ensure the weightings between these levels are consistent.

Calibration has also often been used as an implicit way to correct non-response within a general procedure commonly referred to as "survey adjustment," which eliminates the need for an explanatory nonresponse model, which requires knowing explanatory variables at the individual level for both respondents and non-respondents. Françoise Dupont (Dupont, 1993) illustrated the equivalence of an integrated method and a two-stage method (non-response correction, then calibration) under certain assumptions.

More specifically, the two approaches were combined in the generalized calibration method for treatment of non-response. With generalized calibration, non-response in a survey can be adjusted even when the individual characteristics most correlated with non-response are known only in the sample of people who responded to the questionnaire. This is particularly true when the non-observance of auxiliary calibration variables among non-respondents is the result of a delay between the information in the frame and the survey itself.

The method consists more specifically in writing the calibration equations as follows: $\sum_{k \in r} d_{k} F\left(z_{k}^{\prime} \lambda\right) x_{k}=X$, where $z_{k}$ denotes the vector of non-response explanatory variables, known in the sample $r$ of respondents, and $x_{k}$ denotes the vector of calibration variables, well correlated with the variables of interest, known in the sample and for which the totals (vector $X$ ) are known in the population. $F$ is a calibration function.

As a result, the method can correct non-response even when the explanatory variables are observed only in the sample of respondents, particularly when these variables are variables of interest. This has been programmed in a macro named "Calmar 2" (Le Guennec and Sautory, 2022), which is less popular than the original Calmar macro.

## Balanced sampling

In some respects, balanced sampling - first formalized in the work of Jean-Claude Deville and Yves Tillé in the early 2000s - may appear to be a mirror image of calibration. It goes back to the idea JeanClaude Deville himself put forth as a statistician's classic dream of building a "representative" sample, a perfect reduced model of the population from which it was drawn. While it was already known that a stratified sampling plan with proportional allocation conformed to the population structure in the sample, in terms of a certain number of descriptive characteristics, the aim was also to work with representativeness with regard to continuous variables.

Compared with calibration, this technique also has the advantage of not modifying the design weights, as calibration does, but of imposing balancing conditions for a given set of inclusion probabilities, and therefore, design weights.

However, it should be noted that a survey design, as a law on the set of possible samples in a finite population, cannot be reduced to obtaining a fixed set of inclusion probabilities $\pi_{i}$. Selecting balanced
sampling that respects the set of the $\pi_{i}$, as opposed to one that does not, means modifying the initial design while respecting the values of the $\pi_{i}$ and imposing so-called balancing constraints, which could give the impression of altering the randomness of the process.

Naturally, in extreme cases, too many constraints would reduce the number of admissible samples, or even eliminate them all. Conversely, obtaining a balanced sample may be only approximate, hence the need to put variables into order of their importance in balancing so as to eventually relax constraints.

For more than 20 years now, balancing techniques have spread through the statistical community and become commonplace. Some of the first applications in France involved selecting rotation groups of communes for the continuous census (to which Jean-Claude Deville contributed significantly), with various constraints on population counts. The technique was then used to draw primary units from the master sample from the census for household surveys, and it has also been used to ensure "geographical" balancing, but it would be impossible to list all its current applications. It's also thanks to the fact that the initial macro, called CUBE, saw many developments, extensions and kinetic improvements, first in SAS, then in R.

As with calibration, it can be said that balancing can be carried out "even on nothing", i.e., when no particular variable of interest is available or relevant to the survey topic. As a result, we use basic properties of balancing:

- Balancing on the constant variable equal to 1 is equivalent to balancing the sample on population size N.
- Balancing on the inclusion probability variable is equivalent to ensuring a fixed-size design.
- A stratified design with simple random sampling within each stratum is balanced on the stratum membership indicator variables, and, with constant inclusion probabilities per stratum, balancing on these variables amounts to stratified sampling with a fixed sample size per stratum.
- With constant inclusion probabilities in the population, balancing on stratum membership indicator variables is equivalent to a stratified sampling with a fixed sample size per stratum and proportional allocation.

In addition, balanced sampling using the Cube macro often replaces the basic sampling procedures described above. However, if variables of interest are added, the balancing method obviously becomes more specific than a "basic" selection.

## Weight sharing

The weight sharing method seems to be old (1980s) yet fairly intuitive. Philippe Brion (Clairin and Brion, 1997) cites examples of its application: the "weighted segment" method used in agricultural statistics involves selecting a sample of segments (determined from natural boundaries, e.g., rivers, roads), then surveying each farm with land in the segment. It then becomes clear that a farm can be reached several times from several segments and that this repetition must be taken into account to ensure the weights are calculated appropriately. This method was inspired by Julien and Maranda (1989).

Other examples are also given: to survey a population of nomadic herders, a list of the country's water points must be drawn up and a sample drawn, then we survey the individuals who visited these water points. The risk of double counting can be eliminated by retaining only the first visit during a given period, which requires knowledge of the details of these visits, including how many and when.

On a different note, "basic" weight sharing is used when there is a survey of individuals drawn from a sample of households, with characteristics relating to each of the two levels, and we want to produce consistent statistics from the two levels (e.g., a distribution of individuals according to a given qualitative characteristic must be consistent with a household-level statistic that counts individuals with this characteristic). Therefore, the household and individual weights need to be articulated via a weight sharing method. This field was then extensively investigated and theorized by Pierre Lavallée in collaboration with Jean-Claude Deville.

Jean-Claude Deville - a Breton both by adoption and at heart - had a keen interest in tourism in Brittany. One field of application for the weight sharing method was surveys of visitors to tourist sites. By surveying tourists at a given site for a given period, information about their stays (e.g., location, duration, accommodation), sociodemographic profiles and spending can be obtained. However, conducting these surveys at several sites could result in potentially interviewing the same tourist multiple times, thereby considerably biasing the resulting estimates.

By identifying links (in this case, the places where a tourist surveyed at one location had previously been elsewhere), using the weight sharing method corrects this risk of bias. However, it is very important to clearly and precisely identify all desirable links when using this method and anticipate what questions should be added to the questionnaire for this purpose, whether in these or other surveys.

Examples of similar uses of the method include surveys of homeless people (interviewed at a food establishment or accommodation) and attendance at cultural festivals (e.g., the opera festival in Aix-enProvence or Beaune, the short film festival in Clermont-Ferrand), where audiences may attend several shows, and therefore be surveyed multiple times.

## Calculating the variance of complex estimators

As a reminder, below are a few areas in which Jean-Claude Deville's techniques can be applied:

- fractiles;
- poverty indicators, estimated using SILC (Statistics on Income and Living Conditions) surveys, which have a major political impact;
- concentration indexes (e.g., GINI).

He addressed the following other areas on this theme in an innovative way, but they were not cited in Ardilly, Haziza, Lavallée and Tillé:

- Eigenvalues in a principal component analysis of sample data;
- Estimators derived from selection points from an infinite set, where discrete probabilities are replaced by measures in the sense of measure theory.


## Quotas

Jean-Claude Deville's work on this subject should be seen more as an exercise in style. He was very reluctant about empirical versus random surveys, as the former were suitable only for private institutions that lacked appropriate frames, whereas official statistics had the means to manage such frames, draw samples from them, and calculate estimators and their variances based on the definition of a survey design.

However, he wanted to build a theoretical approach to the quota method and define best practices.
There are two possible approaches:

- A model approach on the variable of interest, which is the best way of theoretically justifying the quota method. However, Jean-Claude Deville was, by nature, very reticent of models and of the risks official statistics would be taking by resorting to them. A model is based on assumptions, but are they verified, how do we know and can they be imposed on the user?
- We try to place quotas in the context of design-based sampling and pretend it was stratified simple random sampling.

In all cases, the key issue is the variance of the selected estimators. Can it be calculated theoretically and estimated and how can rules be derived for choosing between a random method and a quota method when either is possible but with different cost constraints and sample size scope?

## Conclusion

In all the studies mentioned, and many others, Jean-Claude Deville was not only an innovator, but he also paved the way for many other areas of future development and research, which many researchers are now pursuing.

We cannot recall Jean-Claude Deville's memory without mentioning another major project he was involved in, one that goes hand in hand with his theoretical work: INSEE's Journées de méthodologie statistique. He started this event in March 1991 and organized them until 1998. They are still held but are organized by other teams. Michel Glaude, Director of Demographic and Social Statistics at INSEE, highlighted the following at the end of the seventh edition of the Journées on December 5, 2000:
"There was a passing of the baton between a team [Jean-Claude's] that had organized the Journées de Méthodologie Statistique in the past and had devoted a great deal of energy and affection to it, and a new one. [...] I think it's a fine lesson in how to pass on and pursue a goal."

Jean-Claude was obviously a major contributor to each Journées event, and some studies mentioned in Ardilly, Haziza, Lavallée and Tillé were presented there for the first time (jms-insee.fr). His last two papers are from 2015 and 2018 (Deville, 2015; Deville, 2018).

To understand his motivations and objectives in creating the Journées, we must go all the way back to the introduction he gave in 1991, which appeared in the Proceedings of the first edition:
"The first goal of the Journées was to give the statistical community a tangible way to express themselves by bringing it together in one place for two days.

The second was to make "practising statisticians"feel that statistical methods are useful for doing statistics. The third was to put on an event of a high scientific level, even if it meant setting the bar a little high for some of the public.

Inspired by the example and the contacts he had established with Statistics Canada, Jean-Claude Deville's Journées were vector of knowledge sharing and transfer and an opportunity to invite leaders in international statistics to come and have lively, productive debates with them. The events have become a place for statisticians from the public and private sector, professional and academic statisticians, and French and foreign statisticians to strike up a conversation and innovate.

Had it not been for the COVID-19 pandemic, he might have been able to take part in the 30th anniversary in March 2021, but the most recent edition had to be postponed to March 2022.

Jean-Claude Deville had an inquisitive mind, always seeking to expand his knowledge and looking for improvements. He paved the way for others and proved that statistics is far from a static discipline.

## References

Clairin, R., and Brion, P. (1997). Manuel de sondages. Applications aux pays en voie de développement, Insee-Ceped, 1996. PDF version at: https://www.pseau.org/outils/ouvrages/ceped_manuel_de_ sondages_1993.pdf.

Deville, J.-C. (2015). Quelques éléments de géométrie et d'algèbre pour comprendre la nature d'un échantillonnage équilibré. Actes des Journées de méthodologie statistique, Insee, Paris.

Deville, J.-C. (2018). Calage à poids bornés : Que fait-on au juste ? Actes des Journées de méthodologie statistique, Insee, Paris.

Dupont, F. (1993). Calage et redressement de la non-réponse totale, validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. Actes des Journées de méthodologie statistique, Insee, Paris.

Julien, C., and Maranda, F. (1989). Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988. Working paper No. BSMD-89-012F, Statistics Canada, Ottawa.

Le Guennec, J., and Sautory, O. (2002). CALMAR 2 : une nouvelle version de la macro Calmar de redressement d'échantillon par calage. Actes des Journées de méthodologie statistique, Insee, Paris.

# Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" 

Françoise Dupont ${ }^{1}$


#### Abstract

Many things have been written about Jean-Claude Deville in tributes from the statistical community (see Tillé, 2022a; Tillé, 2022b; Christine, 2022; Ardilly, 2022; and Matei, 2022) and from the École nationale de la statistique et de l'administration économique (ENSAE) and the Société française de statistique. Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé provide an in-depth look at Jean-Claude Deville's contributions to survey theory. To pay tribute to him, I would like to discuss Jean-Claude Deville's contribution to the more day-to-day application of methodology for all the statisticians at the Institut national de la statistique et des études économiques (INSEE) and at the public statistics service. To do this, I will use my work experience, and particularly the four years (1992 to 1996) I spent working with him in the Statistical Methods Unit and the discussions we had thereafter, especially in the 2000s on the rolling census.


Key Words: Methodology; Calibration; Balanced sampling; Automatic coding; Harmonic analysis; Jean-Claude Deville.

I first learned of Jean-Claude Deville's work on harmonic analysis of data on the trajectories of women's fertility for my studies at ENSAE in the late 1980s. Jean-Claude Deville had examined these methods in the 1970s.

In his 1974 article (Deville, 1974), Jean-Claude Deville explained that after examining fertility in previous studies, he was looking for a satisfactory exploratory method to use rich individual data since many were available for multiple dates, which was quite rare at the time. He used the 1962 family survey, which provided information on the life trajectory of female respondents regarding the birth of children. In his introduction "Statistician seeking an analytical method," he explains that he wants to analyze the gradual formation of families and the calendar of births based on the number of years married and that his entire thinking stems from a concrete problem with the data to analyze. He explains in detail why he does not want to use econometric methods that presuppose a model - and therefore a preconceived idea of causal links. We already see his reservations about econometrics, which increased later on. Next were the mathematical developments to show how to adapt the general principle of harmonic analysis to time series data. This is an extension of harmonic analysis, as it is usually practised. An application of this was illustrated in the article by Deville (1977).

This is a typical example of the research process in applied statistics: we start with a tangible case in a work environment for which we do not yet have a suitable tool, then conduct bibliographical research by integrating our predecessors' contributions, and lastly, a mathematical reflection to propose a new method to test out in real life, refine and popularize. This long, open process, the example of Pierre Thionet's career (1916-2002, architect of survey methodology at INSEE from 1946, to whom we owe the introduction of random surveys at INSEE, together with Raymond Lévy-Bruhl, who came up with methodological
reflections involving many ideas that went on to be developed) and whom Jean-Claude admired (Deville, 2003; Armatte, 2003; Ardilly, 2022), explains why he has always defended career-based logic for methodologists so that ideas can come from accumulated experience and feed on theoretical readings, by proposing improved methods in order to innovate.

For him, the elegance of the mathematical solution and its successful application were a source of great satisfaction. He spoke passionately about the findings he was proud of, such as drawing 1 out of 20 people for the 1990 census, which already comprises the idea of balanced sampling that Tillé (2022a) spoke about in his tribute to Jean-Claude Deville at the 2022 Journées de Méthodologie Statistique conference (Statistical Methodology Days). Another source of pride was the survey designs for external requesters for which he proposed creative, custom-made solutions.

He was the architect of two "institutional" creations that expanded over time: a Statistical Methodology Unit and the Journées de Méthodologie Statistique. In both cases, "problems and solutions" were prioritized over "theory."

## Creation of the Statistical Methodology Unit

In 1994, the Statistical Methodology and Survey Division saw its missions expanded and its staff grow, going on to become the Unité Méthodes Statistiques (UMS) (Statistical Methodology Unit), which was responsible for the statistical methods used to produce demographic and social statistics. He led the UMS until 1998.

## Journées de Méthodologie Statistique

Jean-Claude Deville created the Journées de Méthodologie Statistique, based on the concept of the annual scientific meetings that take place at the U.S. Census Bureau and Statistics Canada. In his view, there were a number of objectives, which he presented in his introduction of the third edition in 1993 (Deville, 1993):
"There are many general statisticians in regional directorates and departmental statistical services with few opportunities to meet to talk shop." [translation]. The purpose was to lead a network of statisticians and maintain their skills.
"The second, harder idea [...] was to give visibility to methodology work." [translation]
"Lastly, the third idea: ask foreign colleagues to give their thoughts on a specific problem and comment on INSEE's work." [translation]

He concludes as follows:
"In the end, the success of the Journées is easily explained: official statistics, as a discipline, exist in both France and around the world. It requires effort; the Journées de la statistique officielle are the periodic scientific meeting. It is easy to think of what is missing: an associated structure, as well as training and research structures in universities. All this should result in a more systematic publication of the work in more accessible forms than the current ones." [translation]

Of the four years (1992 to 1996) I spent in the Statistical Methods and Survey Division and in the UMS, in two different positions, I witnessed many changes to statistical practices, always based on two pillars: theory and practice.

In 1992, when I returned to the UMS to work on non-response, the theoretical groundwork of calibration was established and the SAS CALMAR macro (Deville, Särndal and Sautory, 1993) developed by Olivier Sautory could already be used to apply this method. On one hand, the tangible work of INSEE statisticians consisted in applying an iterative calibration procedure by successively calibrating each margin until the process converges to adjust their surveys.

Thanks to the first article on calibration published in 1992 in JASA (Deville and Särndal, 1992), he scored very high points with his INSEE colleagues and gave survey methodology its first distinctions at the institute. It had yet to be proven that this method simultaneously corrected non-response bias and improved the accuracy of the estimator, as per the intuition of practitioners and their survey adjustment practices. This is the work I completed with Jean-Claude Deville, which resulted in publications for the 1993 Journées de méthodologie statistique (Deville and Dupont, 1993; and Dupont, 1993).

We also needed to expand the use of the SAS CALMAR macro among practitioners. This is why the UMS gave many presentations on the CALMAR macro and offered to advise survey managers. Thanks to the presentations, the availability of the tools and accompanying advice, and management of the adjustment of complex surveys, we were able to gain the trust of survey managers and gradually modify their practices. This led to more in-depth reflection on the different cases of adjustment and to moving toward more homogeneous practices and a more sophisticated solution, when applicable (Dupont, 1995).

In the same spirit, other practices were developed with the help of his UMS colleagues. The growing success of the methodology days demonstrated the desire and need to talk about and share statistical methodology in all fields, without prioritizing subjects according to their mathematical content, but rather according to their actual usefulness, in a field wider than econometrics, the only statistical discipline truly recognized at INSEE at the time. He helped to bolster the UMS's credibility and advance its recognition in official statistics in France and around the world.

This recognition was confirmed with the development of the unit in charge of demographic and social statistical methodology, whose human and financial resources expanded in 1994, while opening up to data collection methods with the creation of a small team to explore this, which he entrusted to me.

Under his leadership, the UMS conducted purely methodological operations along with current collection operations. Work was initiated to quantify the interviewer effect (Berthier, Deville and Néros, 1998) and two other projects to analyze the interviewer effect more qualitatively in surveys being developed on the duration of work and situations of disadvantage, another on the effect of mandatory response (Berthier and Dupont, 1999), and lastly, a more thorough project measured the impact of manual editing by managers on paper questionnaires. This involved examining the reality of the manual edits done in regional directorates based on instructions from survey designers in order to examine the evolution of edits with a view to transitioning surveys to computer-assisted personal interview (CAPI) mode, which includes the
possibility of automatic edits. By looking at the differences between the instructions given and the actual work done, this project illustrated the difficulty involved in manual edits.

Many other projects began during this period in order to advance practices through turnkey tools provided to statisticians, as is already done at Statistics Canada.

After the CALMAR tool, management of the automatic coding tool SICORE, designed by Pascal Rivière and Eric Meyer (Rivière, 1993) and led by Jean-Claude Deville, was transferred to the UMS to develop and systematize its use for coding occupations and countries. The tool was also adopted by foreign countries through INSEE's cooperation activities.

In 1994, a Comité du label (approval committee) was created (Christine and Roth, 2020), with the purpose of looking at survey design in terms of sampling, adjustment, questionnaire design and, more generally, the information collection system from a methodology perspective. The UMS was represented by an expert in charge of reviewing how best practices in terms of methods are applied. This systematic review of the methodology of all surveys helped to gradually develop the role of advising survey designers on the adjustment and estimation phases, as well as on questionnaire design (Bilocq, 1996). He contributed to advancing tangible practices and developing the advisory role of the UMS in survey development, which was formalized a few years later in the organization.

The UMS took on the management of a common set of questions in household surveys about household composition, which had just been developed by a joint effort led by an experienced survey designer. The challenge was to systematize its use by offering services, as for a tool, in order to harmonize practices inasmuch as possible. Variables were added to produce data on immigrants and foreigners that could be analyzed in all household surveys. Managing the conversion to CAPI and automatic coding has made it more attractive to survey designers, and has enabled us to take pooling in surveys further, thanks to a turnkey tool.

In line with the idea of a turnkey tool and advice on using it, the design of a precision calculation software package called POULPE (Caron, Deville and Sautory, 1998) was launched.

In the same period, the methodology of the employment survey was undergoing a major change, with the introduction of an ongoing employment survey. The UMS played a key role in the discussions on the various aspects of methods and organization (Lagarenne and Schuhl, 1995; Détour, Thiesset and Schuhl, 1995) that led to the introduction of the ongoing employment survey, which is still conducted today.

The need for local statistics incited the UMS to support the implementation of a local population estimation method (Decaudin and Labat, 1996).

At the same time in 1994, Jean-Claude Deville and other INSEE colleagues were thinking about making methodological changes to the census, which was to later become the "rolling census" (Deville and Jacod, 1996). This unique method, though complex with regard to estimation and variance calculation, had many benefits for the organization, particularly smoothing the budget and the workload, which is why it was adopted (Durr, 2005). This transformation of the census was the first opportunity to use balanced sampling
when it was introduced in the 2000s, thanks to the CUBE macro (Tillé, 2022a). In practice, it revealed other qualities; for example, it contributed to implementing a cycle of ongoing improvements to the processes, which was not part of its initial objectives.

Everyone knows Jean-Claude Deville differently, based on their experience working with him. I especially wanted to show his constant desire to bring about changes to the work at INSEE, be it for "hard" methodology (estimation, adjustment, sampling and variance calculation) or "soft" methodology, i.e., collection methods that are harder to analyze, such as a topic of study, and are published much less often.

Thanks to his perseverance and the solutions he implemented, he helped survey methodology at INSEE gain more recognition. For Jean-Claude Deville, recognition of the role of methodology remained imperfect because the methodology used for businesses was the responsibility of another unit at the time, despite the potential for pooling resources.

It was his passion for theoretical mathematics and his unwavering belief in the day-to-day application that contributed to all the technical progress mentioned above and paved the way to an ambitious, recognized methodology at INSEE.

## References

Ardilly, P. (2022). Hommage à Jean-Claude Deville. Statistique et Société, 10, 121-123, www.statistique-et-societe.fr.

Armatte, M. (2003). Pierre Thionet et l'introduction en France des méthodes de sondage aléatoire. Journal de la société française de statistique, 144, 227-255, http://eudml.org/doc/199757.

Berthier, C., Deville, J.-C. and Néros, B. (1998). Une méthode de mesure de l'effet enquêteur. Actes des Journées de Méthodologie Statistique, 133-143, Insee, Paris.

Berthier, C., and Dupont, F. (1999). L'incidence du caractère obligatoire des enquêtes. Enquêtes et sondages : méthodes modèles et applications, Dunod.

Bilocq, F. (1996). Conception et évaluation de questionnaire. Actes des Journées de Méthodologie Statistique 1996, Insee, Paris.

Caron, N., Deville, J.-C. and Sautory, O. (1998). Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE. Methodological document on the POULPE software, Insee, Paris.

Christine, M. (2022). Discours introductif en hommage à Jean-Claude Deville. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Christine, M., and Roth, N. (2020). Le Comité du label, Un acteur de la gouvernance au service de la qualité des statistiques publiques. Courrier des statistiques, 5, 39-52.

Decaudin, G., and Labat, J.-C. (1996). Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de populations. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Détour, C., Thiesset, C. and Schuhl, P. (1995). Contrôle de qualité de l'enquête trimestrielle emploi : résultats de l'enquête transitions sur le marché du travail. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Annales de l'Insee, 15, 3-101.

Deville, J.-C. (1977). Analyse harmonique du calendrier de constitution des familles en France, disparités sociales et évolution de 1920 à 1960. Population, 32, 17-63.

Deville, J.-C. (1993). Introduction. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Deville, J.-C. (2003). In Memoriam Pierre Thionet 1915 - 2002. The Survey Statistician, 47, 2.

Deville, J.-C., and Dupont, F. (1993). Non-réponse: Principes et méthodes. Actes des Journées de Méthodologie Statistique, 53-69, Insee, Paris.

Deville, J.-C., and Jacod, M. (1996). Replacing the traditional French census by a large-scale continuous population survey. Proceedings of the Annual Research Conference, U.S. Census Bureau, Washington.

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

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling. Journal of the American Statistical Association, 88, 1013-1020.

Dupont, F. (1993). Calage et redressement de la non-réponse totale : Validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. Survey Methodology, 21, 2, 125-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1995002/article/14397-eng.pdf.

Durr, J.-M. (2005). The French new rolling census. Statistical Journal of the United Nations Economic Commission for Europe, 22, 3-12, DOI:10.3233/SJU-2005-22102.

Lagarenne, C., and Schuhl, P. (1995). Contrôle de qualité de l'enquête trimestrielle emploi : résultats de l'enquête protocole. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Matei, A. (2022). In memory of Jean-Claude Deville. International Association of Survey Statisticians, Webinar 17.

Rivière, P. (1993). SICORE: Système général de chiffrement automatique. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Tillé, Y. (2022a). Session spéciale en hommage à Jean-Claude DEVILLE : L'échantillonnage équilibré. Actes des Journées de Méthodologie Statistique, Insee, Paris.

Tillé, Y. (2022b). On calibration and balanced sampling, Webinar in memory of Jean-Claude Deville, Statistical matching of survey data trough calibration, optimal transport and balanced sampling. International Association of Survey Statisticians, Webinar 17.

# Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" 

# Jean-Claude Deville: Mathematics lover, high-flying researcher, and visionary 

Camelia Goga and Anne Ruiz-Gazen ${ }^{1}$


#### Abstract

Jean-Claude Deville is one of the most prominent researcher in survey sampling theory and practice. His research on balanced sampling, indirect sampling and calibration in particular is internationally recognized and widely used in official statistics. He was also a pioneer in the field of functional data analysis. This discussion gives us the opportunity to recognize the immense work he has accomplished, and to pay tribute to him. In the first part of this article, we recall briefly his contribution to the functional principal analysis. We also detail some recent extension of his work at the intersection of the fields of functional data analysis and survey sampling. In the second part of this paper, we present some extension of Jean-Claude's work in indirect sampling. These extensions are motivated by concrete applications and illustrate Jean-Claude's influence on our work as researchers.


Key Words: Electricity consumption estimation; Functional data analysis; Generalized weight share method; Indirect sampling; Postal traffic estimation.

## 1. Introduction

During his long career as "administrateur" at Insee (Institut national de la statistique et des études économiques), the French Statistical Office, Jean-Claude Deville encountered various statistical studies, some of them raising new challenges in statistics or in survey sampling theory. One of his greatest concerns was to solve the problems of practitioners, as he himself stated in the introduction of Deville (1974): "The following article was born from a very concrete problem: the study of the progressive constitution of families, of the calendar of births according to the duration of marriage." (the original French version is: "L'article qui suit est né d'un problème tout à fait concret: l'étude de la constitution progressive des familles, du calendrier des naissances en fonction de la durée de mariage"). He did not hesitate to go to the field and participate in the drawing of a sample for example. He went one day, at 5 o'clock in the morning, in a sorting center to assist in the drawing of letters carried out by the French postal service. On this occasion he said: "We can feel statistics live!" (source: personal communication with Alain Dessertaine, research engineer at La Poste, the French postal service, who collaborated with Jean-Claude Deville on surveys conducted by La Poste, see also Section 3 of the present paper). His taste for applications, combined with his taste for mathematics and his extraordinary mathematical culture, led to the development of new theories and methods in statistics and survey sampling that answered the initial problems and much more.

[^2]Our collaboration with Jean-Claude was long-lasting (since the 2000s). We attended many of his fascinating talks and courses, had a lot of long and exciting discussions around mathematics (sometimes on small pieces of paper), and published several papers with him. He greatly influenced our working methods as researchers by encouraging us to start from applied problems, and develop general theoretical results even though we did not work in official statistics. He was very inspiring and passed on to us his passion for survey theory and more generally statistics and mathematics, but also his taste for applications. In this discussion, we would like to illustrate how Jean-Claude influenced our work, and as a result, the work of our PhD students. To do this, we will examine two concrete applications that we have encountered in our career: electricity consumption curves and postal traffic estimation. These applications allowed us to make some (small) advances in the theory of survey sampling that are extensions of Jean-Claude's work in areas where he was a precursor: functional data analysis and indirect sampling.

The Section 2 of this discussion is dedicated to Jean-Claude's work on independent functional data, during the 1970s, before he started working on survey sampling. This section gives also an extension of the functional data analysis approach to dependent survey data from a population of curves, such as the electricity consumption curves, and the aim is to estimate the total electricity consumption curve. In the Section 3, we describe a problem of indirect sampling encountered in practice at La Poste. We focus on a particular type of links structure between the frame and the target populations called "Many-to-One" and show how the results in Deville and Lavallée (2006) can be extended. We also consider a particular double indirect sampling where the number of links to observe can be greatly reduced. Section 4 concludes the discussion with some personal thoughts on our collaboration with Jean-Claude.

## 2. Deville and functional data analysis: A visionary of big data

### 2.1 Independent functional data

The seminal paper from 1974, "Méthodes statistiques et numériques de l'analyse harmonique", unfortunately published only in French, develops a new kind of statistics, the statistics from functional data, largely popularized later by the book of Ramsay and Silverman (2005). Deville's 1974 paper relates to a study carried out in 1962 on 240,000 women aged under 70 , with the main aim of individually reconstructing the history of the constitution of each family. The date of birth of each woman along with the dates of birth of her children were registered, as well as the level of education, occupation, geographical location of the family in 1962, and further information about the husband, etc. As Deville claims in Deville (1977), which details the application on family data of the method developed in Deville (1974), "the statistician feels a certain perplexity in front of such rich data". Deville was at the dawn of big data analysis and the use of traditional methods such as principal component analysis (PCA) led to uninterpretable results. Statistical experiments based on a division of time at a finer step, quarters instead of years, led Deville to think that the "continuous nature of time was not without influence on the choice of a method of analysis" (Deville, 1974, Chapter 1). He considered that each individual $k$ was characterized by a temporal curve or trajectory $Y_{k}(t)$ for $t$ varying in a closed interval $[0, \mathcal{T}]$, "instead of being caracterized by a finite dimensional vector"
(Deville, 1977) and he extended principal component analysis to this new functional setting. In the family study, 0 represents the wedding year and $\mathcal{T}=20$ years since the wedding year. Deville considers this period of time after the wedding date as the maximum laps of time for having children since very few children are born after 20 years of marriage. The value $Y_{k}(t)$ represents the number of children at year $t$ (since the wedding year). The study was conducted on only 100,000 so-called complete families, i.e., not dissolved after 20 years of marriage and the wife being under 45 years of age at the time of marriage.

The observations $Y_{k}=\left(Y_{k}(t)\right)_{t \in[0, \tau]}$ are curves or functions, and handling such objects requires tools from the theory of random processes and functional analysis. New and different statistical objectives appear in this functional setting and in his pioneer work, Deville puts the basis of a new kind of statistics, called several years later functional data analysis (Ramsay and Silverman, 2005). In this infinite dimensional space, a first step would be to represent the data as well as possible in a small dimensional space in order to get a description of the functional data that allows interpretation. In order to do that, Deville (1974) supposed that $Y_{k}$ are independent random functions belonging to the Hilbert space $L^{2}[0, \mathcal{T}]$, i.e., the space of square integrable functions defined on the closed interval $[0, \mathcal{T}]$, equipped with the inner product $<f, g>=\int_{0}^{\tau} f(t) g(t) d t$ and the induced norm

$$
\|f\|=\left(\int_{0}^{\tau} f^{2}(t) d(t)\right)^{1 / 2}
$$

for $f, g \in L^{2}[0, \mathcal{T}]$. He showed that $Y_{k}$ may be represented as follows:

$$
Y_{k}(t)=\mu(t)+\sum_{j \geq 1}<Y_{k}-\mu, v_{j}>v_{j}(t), \quad t \in[0, \mathcal{T}],
$$

where $\mu(t)$ is the mean of $Y_{k}(t)$ (i.e., the average number of children for the study families at instant $t$ ), $v_{j}(\cdot), j \geq 1$ are orthonormal functions from $L^{2}[0, \mathcal{T}]$ depending on $t$ but not on the individual $k$, and called the harmonics of the process in Deville (1974) by analogy with harmonical analysis of periodical/stationary signals on Fourier basis. The components $\left\langle Y_{k}-\mu, v_{j}\right\rangle, j \geq 1$ depend on the individual $k$ but not on the time, are not correlated with each other, and are of variance equal to $\lambda_{j}$, with $\lambda_{1} \geq \lambda_{2} \geq \ldots$. The functions $Y_{k}$ may be approximated in a small $q$-dimensional space as follows:

$$
\begin{equation*}
Y_{k}(t)=\mu(t)+\sum_{j=1}^{q}<Y_{k}-\mu, v_{j}>v_{j}(t)+R_{q, k}(t), \quad t \in[0, \mathcal{T}], \tag{2.1}
\end{equation*}
$$

where the remainder $R_{q, k}(t)$ is the smallest possible according to an integrated variance criterion. Relation (2.1) allows us to represent as well as possible the deviation of the curves $Y_{k}$ from their mean function $\mu$ in a small $q^{\text {th }}$ dimensional space. Expression (2.1) is the functional extension of the classical decomposition on the first $q$ principal components, and Deville (1974, Chapter 4) viewed it as a functional principal component analysis (FPCA), a term used later by many authors (Ramsay and Silverman, 2005). The functions $v_{j}(\cdot)$ and the quantities $\lambda_{j}, j \geq 1$ are closely linked to the covariance operator $\Gamma$ of $Y_{k}$, the functional equivalent of the covariance matrix. The covariance operator $\Gamma$ maps $L^{2}[0, \mathcal{T}]$ to $L^{2}[0, \mathcal{T}]$ and is defined by:

$$
\begin{equation*}
\Gamma a(r)=\int_{0}^{\tau} \gamma(r, t) a(t) d t \tag{2.2}
\end{equation*}
$$

for any function $a \in L^{2}[0, \mathcal{T}]$, and $\gamma(r, t)=\operatorname{cov}\left(Y_{k}(r), Y_{k}(t)\right)$, the covariance between $Y_{k}(r)$ and $Y_{k}(t)$ for $r, t \in[0, \mathcal{T}]$. Then, $\lambda_{j}$ and $v_{j}(\cdot)$ are respectively the eigenvalues and eigenfunctions of $\boldsymbol{\Gamma}$ :

$$
\begin{equation*}
\boldsymbol{\Gamma} v_{j}(t)=\lambda_{j} v_{j}(t) \tag{2.3}
\end{equation*}
$$

with $\lambda_{j} \geq 0, j \geq 1$, supposed to be sorted in decreasing order $\lambda_{1} \geq \lambda_{2} \geq \ldots$. In a $p$-dimensional space, $\boldsymbol{\Gamma}$ is the classical covariance matrix of size $p \times p$, and $\lambda_{j}$ and $v_{j}$ are the usual eigenvalues and eigenvectors. The reader familiar with multivariate PCA (Jolliffe, 2002), will recognize in $\left\langle Y_{k}-\mu, v_{j}\right\rangle$ the scores of the $j^{\text {th }}$ principal component with variance equal to $\lambda_{j}$ where $\lambda_{j}$ interpreted as the part of the total variance explained by the $j^{\text {th }}$ principal component. The same interpretation of $\lambda_{j}$ is valid in the new functional setting (Deville, 1974). The $q^{\text {th }}$ dimensional space spanned by the eigenvectors $v_{j}, j=1, \ldots, q$ gives a representation of the main modes of variation along time $t$ of the data around the mean $\mu$. Deville (1977) implemented this new statistical approach and analyzed in-depth the family trajectories.

### 2.2 Dependent functional data from sample surveys

The method developed in Deville (1974) may be applied to various contexts, and Deville gave several examples in Chapter 8 of the same article. He had forward-looking ideas since studies requiring these functional tools soon appeared in various domains such as chemometrics (Hastie and Mallows, 1993), economics (Kneip and Utikal, 2001), climatology (Besse, Cardot and Stephenson, 2000), biology (Chiou, Müller and Wang, 2003), to name just a few among the numerous references on this field.

A new statistical age appeared with the spread of automatic process for data collection as well as with increasing storage capacities. Due to such smart devices, it is now very common to have very large data sets. Electricity load curves is an example of such data. More exactly, the ERDF (Electricité et Réseaux de France) installed more than 30 million smart meters in every household and company. These meters are able to send individual electricity consumption measures at very fine time scales (every 30 minutes or even every minute or second). The discretization scheme being very fine, the statistical units can be considered as functions of time. For illustration, let us consider a test population of $N=18,902$ French companies for which the electricity consumption has been measured every half an hour over a period of two weeks. A sample of 20 load curves extracted from this test dataset is drawn in Figure 2.1 as well as the mean and the median profiles. In this example, $Y_{k}(t)$ is the electricity consumption of individual $k$ at instant $t$.

Nevertheless, in the presence of technical and budgetary constraints due to limited passband or storage costs of huge databases, the analysis of the whole dataset may be very difficult or very costly. In Chiky (2009), it is shown that if we are only interested in simple indicators, such as total or mean trajectories, even very simple survey sampling techniques, such as simple random sampling without replacement, are attractive alternatives to signal compression techniques since they permit to obtain precise estimates at a reasonable cost. Based on these findings, the ERDF considered implementing effective sampling strategies to estimate the total electricity consumption leading to several research works combining functional data
analysis and sampling theory. A research collaboration has been set up between ERDF (Alain Dessertaine) and the Université de Bourgogne (Hervé Cardot and Camelia Goga) and several PhD theses were dedicated to related topics (Etienne Josserand, Pauline Lardin-Puech, Anne De Moliner).

Figure 2.1 A sample of 20 electricity consumption curves measured every half an hour over a period of one week.


The mean consumption curve in the population is drawn in bold blue line and the median curve in red.

The first work combining functional data and survey sampling techniques was Cardot et al. (2010) who conducted FPCA with survey data. Next, several papers considered the estimation of the total or mean curve with sampling designs (Cardot and Josserand, 2011; Cardot, Degras and Josserand, 2013). An important issue with this new type of datasets is how to build asymptotic confidence bands with desired coverage rates (Cardot and Josserand, 2011). Cardot, Dessertaine, Goga, Josserand and Lardin (2013) made a comparison, in terms of precision of the estimators for the mean of electricity consumption, of different approaches that can take auxiliary information into account. They also compared the width of the confidence bands. The conclusion of the empirical study was that incorporating the auxiliary information in the sampling design or at the estimation stage improves a lot the performance of estimators. In particular, the width of confidence bands is greatly reduced. Theoretical justification of these results are established in Cardot, Goga and Lardin (2013) and Cardot, Goga and Lardin (2014).

We present briefly below FPCA for survey data. This work extends the results from Deville (1974) for functional but non-independent data, as well as the influence function based approach from Deville (1999).

### 2.2.1 FPCA with survey data

Let us consider a population $U$ of size $N$, on which we observe, for each unit $k$, a deterministic function of time $Y_{k}=\left(Y_{k}(t)\right)_{t \in[0, \tau]}$. The population total curve is defined by $t_{Y}=\sum_{k \in U} Y_{k}$ and the mean trajectory by $\mu_{N}=t_{Y} / N$. The value of $t_{Y}$ or $\mu_{N}$ at a measurement point $t \in[0, \mathcal{T}]$ is obtained directly as $t_{Y}(t)=$ $\sum_{k \in U} Y_{k}(t)$ and $\mu_{N}(t)=\sum_{k \in U} Y_{k}(t) / N$, respectively. Figure 2.1 exhibits the mean electricity curve (in blue) computed over the study population.

To perform FPCA, it is first required to estimate the covariance function of the data at the population level. For $r$ and $t$ in $[0, \mathcal{T}]$, the covariance function $\gamma(r, t)$ between $\left(Y_{k}(r)\right)_{k \in U}$ and $\left(Y_{k}(t)\right)_{k \in U}$ is given by:

$$
\gamma(r, t)=\frac{1}{N} \sum_{k \in U}\left(Y_{k}(r)-\mu_{N}(r)\right)\left(Y_{k}(t)-\mu_{N}(t)\right), \quad(r, t) \in[0, \mathcal{T}] \times[0, \mathcal{T}],
$$

and the associated covariance operator $\Gamma$ is given in (2.2). The covariance operator has the following equivalent forms:

$$
\Gamma=\frac{1}{N} \sum_{k \in U}\left(Y_{k}-\mu_{N}\right) \otimes\left(Y_{k}-\mu_{N}\right)=\frac{1}{N} \sum_{k \in U} Y_{k} \otimes Y_{k}-\mu_{N} \otimes \mu_{N},
$$

where the tensor product of two elements $a$ and $b$ of $L^{2}[0, \mathcal{T}]$ is defined by:

$$
a \otimes b(y)=<a, y>b, \text { for all } y \in L^{2}[0, \mathcal{T}] .
$$

The eigenelements of $\boldsymbol{\Gamma}$ are given by:

$$
\begin{equation*}
\boldsymbol{\Gamma} v_{j}(t)=\lambda_{j} v_{j}(t), \quad t \in[0, \mathcal{T}], \quad j=1, \ldots, N . \tag{2.4}
\end{equation*}
$$

### 2.2.2 Design-based estimators of FPCA

The mean trajectory $\mu_{N}$ or the variance operator $\boldsymbol{\Gamma}$, as well as the eigenelements are unknown since we do not have access to all units $k$ from the population $U$. They are estimated using a sample $s$ selected from $U$ according to a sampling design $p(\cdot)$, with first-order inclusion probabilities $\pi_{k}, k \in U$. Skinner, Holmes and Smith (1986) studied some properties of multivariate PCA in a survey framework, and Deville (1999) determined the asymptotic variance of eigenelements of a PCA using the influence function approach. All these study parameters are non-linear functions of the total $t_{Y}$. The eigenvalues and eigenfunctions of $\boldsymbol{\Gamma}$ are defined through the implicit equation (2.3) and are also non-linear functions. The approach developed in Deville (1999) consists in writing the parameter of interest as a functional $T$ of the discrete measure $M$ defined on the space $L^{2}[0, \mathcal{T}]$ :

$$
M=\sum_{k \in U} \delta_{Y_{k}},
$$

where $\delta_{Y_{k}}$ is the Dirac measure at $Y_{k}$ with $k \in U$. Then, following (Cardot, Chaouch, Goga and Labruère, 2008, 2010), $\mu_{N}$ and $\Gamma$ can be written as functionals of $M$ :

$$
\begin{gather*}
\mu_{N}=\frac{\int Y d M}{\int d M}  \tag{2.5}\\
\Gamma=\frac{\int\left(Y-\mu_{N}\right) \otimes\left(Y-\mu_{N}\right) d M}{\int d M} . \tag{2.6}
\end{gather*}
$$

The eigenelements of $\boldsymbol{\Gamma}$, defined by the implicit equation (2.4), are also functionals of $M$. The measure $M$ can be estimated by:

$$
\hat{M}=\sum_{k \in s} \frac{1}{\pi_{k}} \delta_{Y_{k}}
$$

and the estimators of $\mu_{N}$ and $\Gamma$ are obtained by plugging $\hat{M}$ in (2.5) and (2.6), and obtain:

$$
\begin{equation*}
\hat{\mu}_{\text {Haj }}=\frac{\hat{t}_{Y}}{\hat{N}}, \tag{2.7}
\end{equation*}
$$

where $\hat{N}=\sum_{k \in s} 1 / \pi_{k}$ is the Horvitz-Thompson estimator of $N$. The variance operator $\Gamma$ is estimated by

$$
\begin{equation*}
\hat{\boldsymbol{\Gamma}}=\frac{1}{\hat{N}} \sum_{k \in s} \frac{\left(Y_{k}-\hat{\mu}_{\text {Haj }}\right) \otimes\left(Y_{k}-\hat{\mu}_{\text {Haj }}\right)}{\pi_{k}}=\frac{1}{\hat{N}} \sum_{k \in s} \frac{Y_{k} \otimes Y_{k}}{\pi_{k}}-\hat{\mu}_{\text {Haj }} \otimes \hat{\mu}_{\text {Haj }} . \tag{2.8}
\end{equation*}
$$

The estimators $\hat{\lambda}_{j}$, of $\lambda_{j}$, and $\hat{v}_{j}$ of $v_{j}$ are the eigenelements of $\hat{\Gamma}$, namely

$$
\begin{equation*}
\hat{\boldsymbol{\Gamma}} \hat{v}_{j}(t)=\hat{\lambda}_{j} \hat{v}_{j}(t), \quad t \in[0, \mathcal{T}], \quad j=1, \ldots, N . \tag{2.9}
\end{equation*}
$$

### 2.2.3 Asymptotic approximation and variance estimators

All population parameters of interest can be written as $T(M)$, where $T$ is a functional of degree 0 , namely $T(M / N)=T(M)$. Under mild assumptions on the sampling design and the functions $Y_{k}$, Cardot et al. (2010) show that the design-based estimators $T(\hat{M})$ given in (2.7)-(2.9) are asymptotically designunbiased and consistent for $T(M)$. To obtain the asymptotic variance of $T(\hat{M})$, Cardot et al. (2010) give a von-Mises expansion of $T(\hat{M} / N)=T(\hat{M})$ around $M / N$ as follows:

$$
\begin{aligned}
T(\hat{M})-T(M) & =\int I T\left(\frac{M}{N}, Y_{k}\right) d\left(\frac{\hat{M}}{N}-\frac{M}{N}\right)+R_{T} \\
& =\sum_{k \in U} I T\left(M, Y_{k}\right)\left(\frac{I_{k}}{\pi_{k}}-1\right)+R_{T},
\end{aligned}
$$

where $I_{k}=1$ if unit $k$ is selected in a sample and zero otherwise; $I T\left(M / N, Y_{k}\right)=N \cdot I T\left(M, Y_{k}\right)$ where $I T\left(M, Y_{k}\right)$ is the influence function of $T$ at point $Y_{k}$ as defined in Deville (1999) and called the linearized variable of $T ; R_{T}$ is a remainder term. The influence function of $\mu_{N}$ is $I \mu_{N}\left(M, Y_{k}\right)=\left(Y_{k}-\mu_{N}\right) / N$ (Deville, 1999). Using perturbation theory arguments, Cardot et al. (2010) show that $I \Gamma\left(M, Y_{k}\right)=$ $\left(\left(Y-\mu_{N}\right) \otimes\left(Y-\mu_{N}\right)-\Gamma\right) / N$ and

$$
\begin{aligned}
& I \lambda_{j}\left(M, Y_{k}\right)=\frac{1}{N}\left(<Y_{k}-\mu_{N}, v_{j}>^{2}-\lambda_{j}\right) \\
& I v_{j}\left(M, Y_{k}\right)=\frac{1}{N} \sum_{\ell \neq j} \frac{<Y_{k}-\mu_{N}, v_{j}><Y_{k}-\mu_{N}, v_{\ell}>}{\lambda_{j}-\lambda_{\ell}} v_{\ell},
\end{aligned}
$$

provided that $\lambda_{j} \neq \lambda_{\ell}$ for all $j \neq \ell$. Similar expressions have been obtained for $I \lambda_{j}$ and $I v_{j}$ by Deville (1999) for the non-functional case. Cardot et al. (2010) show that $R_{T}=o_{p}\left(n^{-1 / 2}\right)$ for $T$ given by (2.7)-(2.9), and determine the asymptotic variance of $T(\hat{M})$. Cardot et al. (2010) give also variance estimators which are consistent under assumptions on higher-order inclusion probabilities.

### 2.2.4 Why use FPCA?

The principal components scores $\left\langle Y_{k}-\mu_{N}, v_{j}\right\rangle$, for $j=1, \ldots, q$, are indicators of the deviation of curve $Y_{k}$ from its mean function $\mu_{N}$ and allow us to reconstruct rather well the functions $Y_{k}, k \in U$ in a small $q^{\text {th }}$ dimensional space. When auxiliary variables are available, Cardot et al. (2010) suggest using them to improve the estimation of $t_{\gamma}$. FPCA may be also used when auxiliary information is very large; for example, the electricity consumption recorded every 30 minutes during a certain period may be used as auxiliary information in model-assisted or calibration type estimators of $t_{Y}$ (Cardot, Goga and Shehzad, 2017). Such auxiliary variables are functional, and FPCA allows us to determine new uncorrelated variables and to perform calibration (Deville and Särndal, 1992) on them to improve the estimation of $t_{Y}$ as in Cardot et al. (2017).

## 3. Revisiting and going beyond Deville's ideas on indirect sampling

The question was submitted to us by Alain Dessertaine and Pauline Puech from La Poste. They were concerned by the loss of precision of the national postal traffic estimators after changing the sampling design. To understand their problem in depth, we worked on the theory with a PhD student, Estelle Médous, and derived interesting methodological advances. As explained in Section 4 of Ardilly, Haziza, Lavallée and Tillé (2023), Jean-Claude Deville played a key role in naming and deriving the general principles and properties of indirect sampling. The principal results are given in Deville and Lavallée (2006) and Lavallée (2007). In this section, we give some extensions of Jean-Claude Deville's research work on indirect sampling that are detailed in Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine, and Puech (2023) and Medous (2023).

### 3.1 The survey sampling problem of La Poste

The measurement of monthly French postal traffic is important in particular for the accounting followup at La Poste. However, only part of the postal traffic goes though an automatized processing and the monthly postal traffic has to be estimated through a probability-based survey. In 1994, this survey was called SYCI (Système de collecte de l'information), and consisted in drawing samples from populations of
postman rounds according to a stratified, two-stage and balanced sampling design, and using calibrated estimators. Jean-Claude Deville, then "Head of the Statistical Methods" at Insee, participated in the validation of the methodology of this survey, and in particular studied the precision of the estimators.

Since 2008, the organization of postman rounds has changed and is no longer stable over time. Sampling directly from the target population has become impossible, and the sampling design has evolved to an indirect sampling design, called SYCI2, where the target population is still the population of postman rounds but the frame population is the population of postal addresses. As detailed in Section 4 on indirect sampling of Ardilly et al. (2023), as soon as the link weights are standardized (their sum equals one), Generalized Weight Share Method (GWSM) estimators are unbiased for the parameter of interest. But their precision is not easy to compare with the direct approach. At La Poste, the indirect sampling of SYCL2, compared to the previous direct sampling method of SYCI, led to a precision loss of the estimators, with an increase of the estimated standard deviations by a factor between 2 and 3 . The study summarized below aimed to understand in depth this loss of precision.

The frame population is made of postal addresses while the target population is made of postman rounds. The links structure between these two populations is of a particular type: every address belongs to one and only one postman round. Such links are called Many-to-One (or "All-to-One" in Deville and Lavallée, 2006) and will be abbreviated by "MtO" from now on. As detailed below, for this type of links structure, it is possible to go beyond Deville and Lavallée (2006) in the theoretical study of the GWSM estimators.

### 3.2 MtO indirect sampling

Let us focus on study parameters that are totals over the target population $U_{T}$. Following Deville and Lavallée (2006), we assume that a sampling frame exists for a population $U_{F}$, that is related to $U_{T}$ in such a way that any unit in $U_{T}$ is linked to at least one unit in $U_{F}$. Moreover, we focus on the MtO situation where every unit in $U_{F}$ is linked to one and only one unit in $U_{T}$. Let $y$ be the variable of interest measured on $U_{T}$, and let $y_{k}$ be its value for the $k^{\text {th }}$ unit in $U_{T}$. We are interested in estimating $t_{y}=\sum_{k \in U_{T}} y_{k}$, the total of $y$ over $U_{T}$. A sample $s_{F}$ is drawn from $U_{F}$ according to a sampling design $p_{F}(\cdot)$. We can associate to $s_{F}$ the vector $\left(I_{1}, \ldots, I_{N_{F}}\right)^{\prime}$ where $I_{i}$ is the sample membership indicator of the individual $i$ from $U_{F}$ defined as $I_{i}=1$ if $i$ is selected and $I_{i}=0$ otherwise. We denote by $\pi_{i}=\operatorname{Pr}\left(i \in s_{F}\right)$ the first-order inclusion probability of unit $i$ and by $\pi_{i i^{\prime}}=\operatorname{Pr}\left(i, i^{\prime} \in s_{F}\right)$ the second-order inclusion probability of units $i$ and $i^{\prime}$. We suppose that all of the units $i$ have a positive inclusion probability $\pi_{i}>0$ and we denote by $d_{i}=1 / \pi_{i}$ their sampling weights. The sample $s_{F}$ in $U_{F}$ leads to a sample $s_{T}$ in $U_{T}$ made of the units in $U_{T}$ linked to at least one unit in $s_{F}$.

Consider, for all $i \in U_{F}$ and $k \in U_{T}$, some non negative link weight $\theta_{i k}$ such that $\theta_{i k}$ is positive when $i \in U_{F}$ and $k \in U_{T}$ are linked, and $\theta_{i k}=0$ otherwise. We define the standardized link weights $\tilde{\theta}_{i k}=\theta_{i k} / \sum_{i^{\prime} \in U_{F}} \theta_{i^{\prime k}}$ which satisfy the constraint $\sum_{i^{\prime} \in U_{F}} \tilde{\theta}_{i^{\prime} k}=1$. If, for a given unit $k \in U_{T}, U_{F k}$ denotes the set of units $i$ in $U_{F}$ that are linked to $k$, the link weights $\theta_{i k}$ are equal to zero when $i$ does not belong to $U_{F k}$, and the standardized link weights can be written as $\tilde{\theta}_{i k}=\theta_{i k} / \sum_{i^{\prime} \in U_{F k}} \theta_{i^{\prime} k}$. The total $t_{y}$ can be written as the total on $U_{F}$ of the variable $\sum_{k \in U_{T}} \tilde{\theta}_{i k} y_{k}, i \in U_{F}$, as follows:

$$
\begin{equation*}
t_{y}=\sum_{k \in U_{T}} y_{k}=\sum_{k \in U_{T}}\left(\sum_{i \in U_{F}} \tilde{\theta}_{i k}\right) y_{k}=\sum_{i \in U_{F}}\left(\sum_{k \in U_{T}} \tilde{\theta}_{i k} y_{k}\right) . \tag{3.1}
\end{equation*}
$$

The GWSM estimator of $t_{y}$ studied in Deville and Lavallée (2006) is given by

$$
\begin{equation*}
\hat{t}_{y 1}=\sum_{i \in s_{F}} d_{i}\left(\sum_{k \in s_{T}} \tilde{\theta}_{i k} y_{k}\right), \tag{3.2}
\end{equation*}
$$

and it estimates unbiasedly the total $t_{y}$, provided that the link weights are standardized.
To compute $\hat{y}_{y 1}$, the link weights have to be standardized for all $k \in s_{T}$, which implies that the sum $\sum_{i^{\prime} \in U_{F}} \theta_{i^{\prime} k}=\sum_{i^{\prime} \in U_{F k}} \theta_{i^{\prime} k}$ has to be known for the units $k$ sampled in $U_{T}$. In Figure 3.1, we consider the case of 10 observations in $U_{F}$ linked to one observation in $U_{T}$ with link weights $\theta_{i k}$. The observation $i$ surrounded by an ellipse is sampled in $U_{F}$. It implies that the unit in $U_{T}$ linked with $i$ is sampled, and that 10 links have to be observed to compute the sum of the link weights.

Figure 3.1 Small example with 10 MtO links for a unit sampled in the target population from a unit sampled in the frame population.


Number of observed links: 10

Like the direct Horvitz-Thompson estimator, GWSM estimators are unbiased for the total of interest. However, it is not possible to compare their variance with a direct Horvitz-Thompson estimator in a general framework. Indeed, in Subsection 4.1 of Ardilly et al. (2023), the authors recall that for indirect sampling and given first-order inclusion probabilities in the frame population, it is usually not possible to derive the inclusion probabilities in the target population.

Comparing the direct and indirect approach is therefore difficult in general. Some results are obtained for MtO links and Poisson sampling in Medous et al. (2023). For Poisson sampling in the frame population and MtO links, the inclusion probabilities in the target population can be derived from those in the frame population, and the precision of the Horvitz-Thompson estimator of a total under direct sampling can be compared to the precision under indirect sampling. More precisely, Proposition 2.3 in Medous et al. (2023) states that the variance of the Horvitz-Thompson estimator for the direct sampling is always smaller than or equal to the variance of the GWSM estimator under the associated indirect sampling design. For Poisson sampling and MtO links, there is a loss of precision when using indirect sampling and a GWSM estimator compared to the associated direct sampling design and the Horvitz-Thompson estimator.

Moreover, for MtO links, and a general condition on the sampling design, called the $\Delta$-property, Medous et al. (2023) show that the weak optimal link weights derived in Deville and Lavallée (2006) are also strongly optimal link weights for the GWSM estimator (see Section 4.5 in Ardilly et al., 2023, for more details). The $\Delta$-property is a property of the covariance matrix $\Delta=\left(\delta_{i i^{\prime}}\right)_{i, i^{\prime} \in U_{F}}$ of the random sample membership indicators of the sampling design $p_{F}($.$) . For a given unit k \in U_{T}$, let us recall that $U_{F k}$ is the set, of size $N_{F k}$, of units $i$ in $U_{F}$ that are linked to $k$. Because of the MtO links structure, the size $N_{F}$ of $U_{F}$ is equal to the sum $\sum_{k \in U_{T}} N_{F k}$. In what follows, we consider that units in $U_{F}$ are ordered according to the $U_{F k}$, so that the matrix $\Delta$ of size $N_{F} \times N_{F}$ can be written in blocks $\Delta_{k k^{\prime}}, k, k^{\prime} \in U_{T}$, of size $N_{F k} \times N_{F k}$. The number of blocks is equal to the size $N_{T}$ of the population $U_{T}$. For MtO links and ordered indices in the population $U_{F}$ as detailed above, a sampling design $p_{F}($.$) satisfies the \Delta$-property if:
(i) $\Delta_{k k}$ is invertible, for all $k \in U_{T}$,
(ii) $\Delta_{k k^{\prime}}=c_{k k^{\prime}} 1_{k} 1_{k^{\prime}}^{t}$ for $k \neq k^{\prime} \in U_{T}$, where $1_{k}$ is the $N_{F k}$-dimensional vector of ones.

The $\Delta$-property is a technical property that allows us to simplify the expression of the variance of the GWSM estimator in the particular case of MtO links. This property is satisfied for Poisson sampling and simple random sampling without replacement because the covariance matrix $\Delta$ of these sampling designs have constant off-diagonal terms. But the property is also true for example for stratified simple random sampling without replacement with equal sampling rates, or with unequal sampling rates when the units of $U_{F}$ linked to the same unit in $U_{T}$ belong to the same stratum (see Medous et al., 2023, or details). This result was already derived for Poisson sampling and simple random sampling without replacement in Deville and Lavallée (2006) but not stated clearly as a property of the MtO links structure. Furthermore and still under the $\Delta$-property, Medous et al. (2023) provide an explicit and simple expression for the difference of the variances between any GWSM estimator and the optimal GWSM estimator.

The sampling design of SYCI2 is actually more complex than a simple indirect sample and the loss in precision observed at La Poste when using SYCI2 compared to SYCI had to be studied further. SYCI2 is making use of a double indirect sampling design with an intermediate population of outgoing mail sorting boxes. The properties of double indirect sampling designs are studied in Medous et al. (2023) and Medous (2023) and explained briefly in the next subsection.

### 3.3 MtO-MtO double indirect sampling

As explained in Section 4 of Ardilly et al. (2023), the total of the link weights has to be known for each observation sampled indirectly in the target population in order to standardize the link weights and get unbiased estimators. In the case of MtO links, the number of links between the frame population and the target population may be very large and enumerate the links may not be feasible. At La Poste, it means that all addresses delivered during a sampled postman round must be known. On average, there are approximately 500 addresses per postman round, and it is not possible to enumerate all of the addresses before the departure of the postman. To address this issue, La Poste had to consider a double indirect sampling design and the associated double GWSM estimator. The idea of double indirect sampling is to introduce an intermediate population $U_{M}$ between the frame and the target populations, and to use the same principles as in simple indirect sampling.

The use of an intermediate population was already introduced in Deville and Lavallée (2006) (see Subsection 3.3 for the transitivity property and 6.1 for the factorization) in order to simplify the derivation of optimal link weights. At La Poste, an intermediate population of mail sorting boxes is used to reduce the number of links to observe. The idea is that by introducing an intermediate population, we obtain more freedom in the choice of the link weights as detailed below. For the sake of simplicity, we focus on the case where not only the links between $U_{F}$ and $U_{M}$ but also between $U_{M}$ and $U_{T}$ are MtO. We consider nonnegative link weights $\theta_{i k}^{\mathrm{FT}}$ for $i \in U_{F}$ and $k \in U_{T}$ (resp. $\theta_{i j}^{\mathrm{FM}}$, for $i \in U_{F}, j \in U_{M}, \theta_{j k}^{\mathrm{MT}}$ for $j \in U_{M}$, $k \in U_{T}$ ) associated with the links between $U_{F}$ and $U_{T}$ (resp. $U_{F}$ and $U_{M}$, and $U_{M}$ and $U_{T}$ ). In order to express the total $t_{y}=\sum_{k \in U_{T}} y_{k}$ as a total on $U_{F}$ as in equation (3.1), we need to compute standardized link weights $\tilde{\theta}_{i j}^{\mathrm{FM}}$ and $\tilde{\theta}_{j k}^{\mathrm{MT}}$ such that

$$
\begin{equation*}
\sum_{i \in U_{F}} \sum_{j \in U_{M}} \tilde{\theta}_{i j}^{\mathrm{EM}} \tilde{\theta}_{j k}^{\mathrm{MT}}=1 . \tag{3.3}
\end{equation*}
$$

We have:

$$
t_{y}=\sum_{k \in U_{T}} y_{k}=\sum_{k \in U_{T}}\left(\sum_{i \in U_{F}} \sum_{j \in U_{M}} \tilde{\theta}_{i j}^{\mathrm{FM}} \tilde{\theta}_{j k}^{\mathrm{MT}}\right) y_{k}=\sum_{i \in U_{F}}\left(\sum_{k \in U_{T}} \sum_{j \in U_{M}} \tilde{\theta}_{i j}^{\mathrm{FM}} \tilde{\theta}_{j k}^{\mathrm{MT}} y_{k}\right) .
$$

A double GWSM estimator, unbiased for $t_{y}$, is:

$$
\begin{equation*}
\hat{t}_{y 2}=\sum_{i \in s_{F}} d_{i}\left(\sum_{k \in s_{T}} \sum_{j \in s_{M}} \tilde{\theta}_{i j}^{\mathrm{FM}} \tilde{\theta}_{j k}^{\mathrm{MT}} y_{k}\right) \tag{3.4}
\end{equation*}
$$

where the sample $s_{M}$ in $U_{M}$ (resp. $s_{T}$ in $U_{T}$ ) is made of the units in $U_{M}$ (resp. $U_{T}$ ) linked to at least one unit in $s_{F}$ (resp. $s_{M}$ ).

The standardization of the link weights given by (3.3) is called "global standardization". By standardizing each set of links $\tilde{\theta}_{i j}^{\mathrm{FM}}$ and $\tilde{\theta}_{j k}^{\mathrm{MT}}$ separately, which is called "double standardization", it is easy to check that the link weights are also globally standardized. But as illustrated in the small example below, considering double standardized link weights, can result in a gain in terms of the number of links to observe, compared to simple indirect sampling. This property is not necessarily true for other types of globally standardized link weights.

In Figure 3.2, the links between the frame population $U_{F}$ and the target population $U_{T}$ are the same as in Figure 3.1, but an intermediate population is introduced with 2 units, both linked to the only unit in $U_{T}$, and each one being linked to 5 units in $U_{F}$. In order to compute the estimator $\hat{t}_{y 2}$ from (3.4), we need the link weights to be globally standardized (see (3.3)). One possibility to standardize the sum $\sum_{j \in U_{M}} \theta_{i j}^{\mathrm{FM}} \theta_{j k}^{\mathrm{MT}}$ is to divide it by its sum over $i \in U_{F k}$, where $U_{F k}$ is the population of units in $U_{F}$ linked with $k \in U_{T}$. On the left panel of Figure 3.2, this double standardization leads to observe 12 links since there are 10 observations in $U_{F}$ linked to the observation in $U_{T}$ and 2 additional links between $U_{M}$ and $U_{T}$. This situation is even worse than the 10 links to observe in the simple indirect case (see Figure 3.1).

On the contrary, using a double standardization for the link weights is less costly, and the reason is a bit subtle. The standardization of the $\theta_{i j}^{\mathrm{FM}}$ (resp. $\theta_{j k}^{\mathrm{MT}}$ ) consists in dividing them by their sum over $i \in U_{F j}$ (resp. $j \in U_{M k}$ ), where $U_{F j}$ (resp. $U_{M k}$ ) is the population of units in $U_{F}$ (resp. $U_{M}$ ) linked with $j \in U_{M}$ (resp. $k \in U_{T}$ ). It means that the links have to be known only for $i \in U_{F j}$ for the double standardization instead of $U_{F k}$ for the global standardization proposed just before. On the right panel of Figure 3.2, the 2 links between $U_{M}$ and $U_{T}$ have to be observed, but only 5 links (among 10) have to be observed between $U_{F}$ and $U_{M}$, which leads to 7 links to observe in total. As detailed in Medous et al. (2023), for MtO links between $U_{F}$ and $U_{M}$ and between $U_{M}$ and $U_{T}$, double indirect sampling with double standardization may lead to a reduction in terms of the number of links to observe. At La Poste, on average, there are 500 postal addresses per postman round with 50 addresses per box and 10 boxes per round. Using double indirect sampling with double standardization leads on average to 60 links to observe compared to 500 with simple indirect sampling.

The advantage of double indirect sampling in terms of number of links to observe is clear for La Poste, but questions on the choice of double standardized link weights arise. Can the choice of link weights have a large impact on the efficiency of GWSM estimators? Is it possible to find double standardized optimal link weights? The answer to both questions is affirmative. In Medous et al. (2023), we used some Monte Carlo simulations to illustrate that there may be situations where the loss of efficiency is huge. Moreover, results on historical postal data allow us to find a factor between 2 and 3 of efficiency loss between SYCI and SYCI2. Very recently, Medous (2023), in a preprint that will be available soon, derived the optimal doubly standardized link weights. These link weights allow us to observe fewer links compared to simple indirect sampling design, while obtaining the GWSM estimator of minimal variance (for any variable of interest).

Figure 3.2 Same example as on the previous figure, but with an intermediate population. On the left panel, 12 links necessary for global standardization. On the right panel, only 7 links necessary for double standardization.


Number of observed links: 12


Number of observed links: 7

## 4. Conclusion

The two directions we decided to focus on in this discussion are two examples of Jean-Claude's legacy. We also worked with Jean-Claude Deville on the estimation of complex parameters, on coordinated samples using partial influence function (Goga, Deville and Ruiz-Gazen, 2009), or were inspired by his ideas on non-parametrics for survey sampling (Goga and Ruiz-Gazen, 2014).

Jean-Claude was a man of exchange and transmission, with fruitful research collaborations, and fascinating outreach presentations. We remember in particular some of his presentations for the students of the Master in Statistics and Econometrics at Toulouse School of Economics, and his paper in "Pour la science" Deville (2006). They were full of captivating anecdotes on the renovated French census, on the tourism survey in Brittany (with stories of bakeries and freeway tolls), on the survey for the evaluation of the cost of renovating the documents of the François Mitterand library, to name a few.

With Jean-Claude disappears a fount of knowledge and intelligence from which we thought we could drink endlessly, but also a "father" and a friend (we were part of the "copains" as he used to say). We miss him terribly.

## Acknowledgements

We thank Alain Dessertaine for his many discussions and memories about Jean-Claude Deville.

## References

Ardilly, P., Haziza, D., Lavallée, P. and Tillé, Y. (2023). Jean-Claude Deville's contributions to survey theory and official statistics. Survey Methodology, 49, 2, 257-298. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00017-eng.pdf.

Besse, P.C., Cardot, H. and Stephenson, D. (2000). Autoregressive forecasting of some functional climatic variations. Scandinavian Journal of Statistics, 27, 673-682.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2008). Functional principal components analysis with survey data. Functional and Operatorial Statistics, (Eds., S. Dabo-Niang and F. Férraty), Heidelberg: Springer-Verlag.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. Journal of Statistical Planning and Inference, 140, 75-91.

Cardot, H., Degras, D. and Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. Bernoulli, 19, 2067-2097.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E. and Lardin, P. (2013). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. Survey Methodology, 39, 2, 283-301. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11888-eng.pdf.

Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. Electronic Journal of Statistics, 7, 562-596.

Cardot, H., Goga, C. and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. Scandinavian Journal of Statistics, 41, 516-534.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. Statistica Sinica, 27, 243-260.

Cardot, H., and Josserand, E. (2011). Horvitz-Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. Biometrika, 98, 107-118.

Chiky, R. (2009). Résumé de Flux de Données Distribués. PhD thesis, Paris: Sup Telecom.

Chiou, J.-M., Müller, H. and Wang, J. (2003). Functional quasi-likelihood regression models with smooth random effects. Journal of the Royal Statistical Society, B, 65, 405-423.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Ann. Insee, 15, 3-104.

Deville, J.-C. (1977). Analyse harmonique du calendrier de constitution des familles en france. Disparités sociales et évolution de 1920 à 1960. Population (French Edition), 32(1), 17-63.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey Methodology, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/en/ pub/12-001-x/1999002/article/4882-eng.pdf.

Deville, J.-C. (2006). Peut-on croire aux sondages ? Pour la Science, (344), 58-65.

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.

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

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample data. Biometrika, 96(3), 691-709.

Goga, C., and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. Journal of the Royal Statistical Society, B, 76, 113-140.

Hastie, T., and Mallows, C. (1993). A discussion on "A statistical view of some chemometrics regression tools" by L.E. Frank and J.H. Friedman. Technometrics, 35, 140-143.

Jolliffe, I.T. (2002). Principal Component Analysis. Springer Series in Statistics, second edition. New York: Springer-Verlag.

Kneip, A., and Utikal, K. (2001). Inference for density families using functional principal component analysis. Journal of the American Statistical Association, 96, 519-542.

Lavallée, P. (2007). Indirect Sampling. New York: Springer Science \& Business Media.

Medous, E. (2023). Optimal Weights for double Many-To-One Generalized Weight Share Method. Preprint.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. and Puech, P. (2023). Many-toOne indirect sampling with application to the French postal traffic estimation. The Annals of Applied Statistics, 17(1), 838-859.

Ramsay, J.-O., and Silverman, B.-W. (2005). Functional Data Analysis, second edition. New York: Springer Series in Statistics.

Skinner, C.J., Holmes, D. and Smith, T. (1986). The effect of sample design on principal components analysis. Journal of the American Statistical Association, 81, 789-798.

# Comments on "Jean-Claude Deville's contributions to survey theory and official statistics" 

Carl-Erik Särndal ${ }^{1}$


#### Abstract

In recent decades, many different uses of auxiliary information have enriched survey sampling theory and practice. Jean-Claude Deville contributed significantly to this progress. My comments trace some of the steps on the way to one important theory for the use of auxiliary information: Estimation by calibration.


Key Words: Auxiliary information; Generalized regression estimation; Calibration estimation.

## Introduction

I am honored to reflect on the article by Ardilly, Haziza, Lavallée and Tillé on Jean-Claude Deville and his many contributions to survey theory and practice.

I refer to the article as Ardilly, Haziza, Lavallée and Tillé (AHLT). One part of it is devoted to estimation of the finite population total by calibration theory. I concentrate on it here. More than thirty years have passed since the publication of the influential article on calibration by Deville and Särndal (1992). It was followed a year later by an important sequel, also published in the Journal of the American Statistical Association (JASA).

A forerunner - and a special case - of the calibration estimator is "the well-known estimator by generalized linear regression", as AHLT phrase it. Given in their formula (3.5), it is popularly known as the (linear) generalized regression (GREG) estimator. This gives me reason here to examine how the GREG estimator paved the way for the calibration estimator. The GREG estimator evolved in form and shape in the mid-to-late seventies and in the eighties; it is fair to say that it is the progenitor of the estimator by calibration.

The GREG construction was a product of "the prediction argument": To predict as well as possible the study variable values for the non-observed population units. On the other hand, a somewhat later stream in the literature dwells on "the weighting argument": To conceive and compute the estimator of the population total through an appropriate weighting of the study variable values $y_{k}$ observed for units $k$ in the probability sample $s$, preferably a weighting more "information laden" than simply the inverse inclusion probability weighting $d_{k}=1 / \pi_{k}$.

My comments here trace my gradual understanding over several decades of estimation in the presence of auxiliary information. They are an incomplete testimony to a period of development where Jean-Claude Deville played an important part. Many others contributed and should have been recognized in my notes.

## Meeting

I first met Jean-Claude in the Swiss Alps. Not on skis, but at the traditional yearly "winter school" in statistics held at the resort Les Diablerets. We were both invited speakers there in 1987. I had been told beforehand that I was going to meet a French statistician, relatively new to the survey sampling field, who had made himself known by some recent interesting contributions.

I did meet Jean-Claude there, in person, and in spirit. Contact and mutual understanding happened from the beginning. As a graduate of the prestigious Ecole Polytechnique, Jean-Claude relied on a strong mathematical background and ability; in addition, his work at Institut national de la statistique et des études économiques (INSEE) made him familiar with national statistical agencies and their efforts to produce accurate national statistics.

A unique combination seemed to me to characterize him, as I got to know him: There was a mathematical insight that he applied in sometimes astonishing ways, yet with a clear background in national statistics production. He was very sensitive to resolving problems embedded in a greater practical environment, such as a national statistical agency.

As he explained to me, at already well over 30, he had decided to change fields of research interest to survey science. The idea of probability sampling intrigued him: to select, with known probabilities, from a finite set of identifiable units. His curiosity was stimulated by the amazing variety of ways to do that, and how to estimate the population parameters on data from a probability sample using auxiliary variables.

Seven years his senior, I had behind me a longer exposure to the field. He told me how he had systematically "taken in" the field of survey science, had studied the significant material, from the classical sampling books of the fifties to the recent work, in the newborn vigor that survey sampling theory found from around 1970. For example, he admired the work of the Czech sampling theoretician Jaroslav Hàjek. He had studied the 1977 Wiley book that I had co-authored, Foundations of Inference in Survey Sampling. In the years that followed, he came to Canada several times and we worked together at the Université de Montréal.

## Randomization theory

I believe that Jean-Claude's ideological preference agreed, essentially, with the randomization theory, the design-based fold of survey science. That is, an approach where the inference about the finite population is built on the probability sampling design, using the known inclusion probabilities of the sampled units.

This positioning was by no means obvious, for someone who was learning the field in those days. The 1970's had brought a flux of ideas, some of them conflicting. A formidable challenger to the traditional design-based randomization theory was the model-based theory, which maintained instead that inference was to be based on the stochastic structure stipulated in the assumed model.

So it seems to me that when Jean-Claude set out to learn the field, he might well have felt attracted to, and become a convinced disciple of, the model-based camp. But he did not.

## Auxiliary information

Estimation built around ideas of supplementary information and predicted $y$-values for non-observed units seems to have been studied, or attempted, as early as in the 1950's, in places, such as the national statistical agencies, where insightful survey methodologists dwelled. At INSEE in France, P. Thionet and Y. Lemel were among persons whose important work influenced Jean-Claude.

An important instance of auxiliary information occurs when values are known for all population units on one or more variables, $x$-variables, thought to be related to the survey variable, the $y$-variable. The Scandinavian countries, equipped as they are with high quality population registers of different kinds, were testing grounds for this methodology. The uses of auxiliary information led, beginning in the 1970's, to a strong development in survey statistics research and practice.

## Generalized regression estimator

The GREG estimator rests on "the prediction argument"; the $y$-values for the non-sampled units are predicted, on the basis of a perceived relationship between the study variable $y$ and one or more correlated auxiliary variables. The linear GREG estimator, as we know it today, given in formula (3.5) in AHLT, evolved gradually, from the mid-to-late 1970's onward. Several took a part in this; a gradual refinement took place. The first time that I used "generalized regression estimator" as a term and as a construction principle - incomplete at the time - was, as far as I recall now, in a co-authored article in Biometrika, Cassel, Särndal and Wretman (1976). Among contributions to "the regression thinking", the work at Iowa State by W.A. Fuller and his students stands out.
"Generalized" meant essentially that the early form of the GREG estimator extended what Cochran and other texts from the 1950's had presented as a method to reduce variance, by attaching a linear regression adjustment term to the basic design-unbiased estimator. The adjustment, small in magnitude in large samples, was a function of the auxiliary variables.

More generally, let $\hat{y}_{k}$ be the predicted value, by a linear or non-linear assisting model, of the study variable value $y_{k}$. The GREG construction is $\hat{Y}_{\text {GREG }}=\sum_{s} d_{k} y_{k}+\left(\sum_{U} \hat{y}_{k}-\sum_{s} d_{k} \hat{y}_{k}\right)$, with $U$ representing the population and $s$ representing the probability sample from $U$. The residuals $y_{k}-\hat{y}_{k}$ left by the model fit become more apparent when we write it as $\hat{Y}_{\text {GREG }}=\sum_{U} \hat{y}_{k}+\sum_{s} d_{k}\left(y_{k}-\hat{y}_{k}\right)$.

In papers from the late seventies and early eighties, I had occasion to examine the linear GREG estimator, that is, where $\hat{y}_{k}$ is the result of a linear regression fit. I remember the 1976 co-authored Biometrika article especially because of an administrative curiosity.

Although it was exceptional for an ordinary submission to Biometrika, the chief editor, D.R. Cox, wished to accompany the article with a special discussion, something to which we agreed. The article was seen, apparently, as an unorthodox mixture of an appeal to a model and its properties, and, at the same time, to the randomization distribution, $p$, arising from the probability sampling.

The discussant was T.M.F. (Fred) Smith, a highly respected survey theoretician with whom I came to enjoy, as time went by, much fruitful and friendly contact. He wrote: "This paper raises a fundamental issue in finite population inference via a superpopulation. ... The authors impose the further constraint that (the regression estimator) $T$ should be $p$-unbiased ... Why should the selection probabilities, $p$, take any precedence over the model $\xi$ ?" It was a well-motivated question from his model-based point of view: If the regression model was trustworthy for building the estimator, why should it then be abandoned, in favour of the randomization distribution, when it came to evaluating basic properties, such as bias and variance?

This was an illustration of an exchange that could occur in those days, between a member of the new model-based camp and one of the traditional design-based camp. In due time, the perspective in that article developed into "model-assisted design-based estimation", i.e., assisted by the model but not dependent on its "truth". The asymptotic $p$-unbiasedness served as protection against a possibly false or improper model. The "truth" or not of the model was not the primary issue. If the model does not hold, the estimator is still asymptotically design-consistent.

## Reoriented thought process

The literature from 1980 and later suggested a "reorientation of the thought process"; some of the attention shifted away from the prediction argument to an alternative argument centered on the weights assigned to the observed sample $y$-values.

In the prediction approach, I and others had built an estimator by predicting the unobserved $y$-values as well as possible, via a regression fit of some kind, and with a use of the auxiliary variables. In the interest of accurate estimation, it was clear that the predictions $\hat{y}_{k}$ must reflect as accurately as possible the unknown $y_{k}$ for the non-observed units; the residuals $\hat{y}_{k}-y_{k}$ should be small.

The weighting argument focuses instead on estimation by an appropriate weighting of the $y$-values observed in the sample. In the interest of accurate estimation, the weight given to a sampled unit should reflect what is known, and what is particular, about a unit in the population. "Good weights" could or should be sample-dependent functions of the auxiliary vector values $\mathbf{x}_{k}$. They might entail just a small but important adjustment to the basic design weights $d_{k}=1 / \pi_{k}$, as is the case with the weights implied by the GREG estimator formula.

In Särndal (1982), I write the linear GREG estimator as a weighted sum over the sample, with the observed study variable value $y_{k}$ receiving the weight $d_{k} g_{k}$, where the sampling design weight $d_{k}=1 / \pi_{k}$ undergoes a small adjustment $g_{k}$, slightly away from "one". Bethlehem and Keller (1987) also show the linear weighting interpretation of the GREG estimator, and they note the calibration property of the weights.

## Model-assisted survey sampling

The thick manuscript of the book with that title, by Särndal, Swensson and Wretman, was nearing completion when I and Jean-Claude co-operated in the late 1980's. I described to him the spirit of the book.

Published in 1992 by Springer-Verlag; it became popularly known as the Yellow Book. It advocated a design-based outlook on inference, more particularly in a vein that became widely known as model-assisted. The role of the model was explained at length, in particular in Chapter 6; the wording was important; it was convincing and inspired others.

Curious as it may seem today, as book authors we were quite sensitive to the survey theory climate. For or against "reliance on models" had been a much debated question in the survey sampling theory literature since the early 1970 's. We had had our part of "the hot feelings", through the case with the above mentioned 1976 Biometrika article on generalized regression estimation.

The Yellow Book used the prediction argument to carefully build and explain the GREG estimator of $Y=\sum_{U} y_{k}$. Predicted values $\hat{y}_{k}$ from a linear regression fit are given by $\hat{y}_{k s}=\mathbf{x}_{k}^{\prime} \mathbf{B}_{s}=$ $\mathbf{x}_{k}^{\prime}\left(\sum_{s} d_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime}\right)^{-1}\left(\sum_{s} d_{k} \mathbf{x}_{k} y_{k}\right)$, where $s$ is the probability sample and $d_{k}=1 / \pi_{k}$. They are computable for all units if $\mathbf{x}_{k}$ is available for all. The linearly weighted form of the GREG estimator is also emphasized:

$$
\hat{Y}_{\text {GREG }}=\sum_{s} d_{k} g_{k} y_{k},
$$

where $g_{k}=1+\left(\sum_{U} \mathbf{x}_{k}-\sum_{s} d_{k} \mathbf{x}_{k}\right)^{\prime}\left(\sum_{s} d_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime}\right)^{-1} \mathbf{x}_{k}$.

## The gee-weights

The "gee-weight", as we used to call $g_{k}$, is a weight factor, equaling one plus a term of minor magnitude in large samples, nevertheless with an important impact. It modifies slightly the design weight $d_{k}=1 / \pi_{k}$ into a total weight $d_{k} g_{k}$, and, as the Yellow Book explains, those weights satisfy

$$
\sum_{s} d_{k} g_{k} \mathbf{x}_{k}=\sum_{U} \mathbf{x}_{k} .
$$

It came to be known as the calibration property. I vividly remember my momentary surprise at my own "discovery" of the property; it was, however, quite evident upon a closer look, and therefore not given any immediate attention. Others active in the field were no doubt also aware of it in the years around 1980. The weights $d_{k} g_{k}$, their function, their calibration property, and their use in variance estimation were at the center of attention in a Biometrika article by Särndal, Swensson and Wretman (1989).

The calibration property of the weights $d_{k} g_{k}$ was, however, a signal for my interest in calibrated weight systems: it must be a more generally fruitful idea. When Jean-Claude and I discussed it in the late 1980's, the property was a well-established fact, a starting point: It must be possible to extend and generalize it.

## The rise of calibration theory

A phrase in AHLT catches my attention: The authors maintain that the calibration theory, as in Deville and Särndal (1992), brought "... one of the most important advances in the field of estimation in the presence of auxiliary information: It is possible to construct the GREG estimator by means of a calibration." In other
words, a "discovery" was that calibration theory is sufficiently broad in scope to admit the important GREG estimator under its umbrella.

It is indeed important to have the GREG estimator as a member of the calibration family; I add, though, that my own insight happened in a different temporal order: Knowing in advance that the GREG weights have the calibration property, it must be possible to extend the idea.

This took form in the theory well described in AHLT, with a measure of distance to minimize, between the survey weights $d_{k}=1 / \pi_{k}$ and new weights $w_{k}$, subject to the calibration constraint $\sum_{s} w_{k} \mathbf{x}_{k}=\sum_{U} \mathbf{x}_{k}$. A resulting calibrated weight $w_{k}$ is a function of the auxiliary vector value $\mathbf{x}_{k}$.

## The calibration theory: Implications and interpretations

I started these comments by noting two ways to proceed, "the prediction argument" as opposed to "the weighting argument". This important distinction has set its mark on a part of theory development in survey sampling over the past several decades.

The former was successfully used in creating the (linear or non-linear) GREG estimator. It seems to me now that the weighting argument, as used in calibration theory, has a broader application, or a wider appeal. To estimate, you just add up the properly weighted $y$-values.

Much has been said on calibration in the literature since the Deville and Särndal (1992) article. It is true that calibration was presented there as a design-based methodology under ideal survey conditions, one hundred percent survey response, and an absence of other survey errors as well. However, time passing, the concept of calibration has shown itself to be an instrument of extraordinary power and flexibility, as witnessed, for example, in the variety known as model calibration, and especially in its extensions to nonresponse adjustment, treated in a large literature of its own.

## An image

I close with a digression on the word "calibration". At INSEE, the French term "calage" was apparently well established long ago. It was used, it seems, in early examples of a weighting that confirms the known population total of an auxiliary variable.

As a French verb, "caler" means to rig, to fix, to stabilize. Elsewhere also, there were no doubt insightful statisticians who derived and computed weights, with the property today called "calibrated", well before there was a special name and a special theory for the procedure.

For the 1992 JASA article, Jean-Claude and I settled on the term "calibration", aware of the meaning it has to some, namely, in the idea of the balance scale, this instrument used in food stores long ago: two plates at either end of beam. To fill a customer's order for, say, 600 grams of coffee or butter or flour, the store clerk placed weights on one plate, amounting to 600 grams, then measured up the desired food item on the opposite plate, until balance occurred. The store had a collection of metal weights, in different denominations; they were certified, calibrated, to guarantee the customer's right to a correct weighting. As
some have fondly reminded me over the years, "calibrated weights" brings up this old mental image of a trustworthy procedure.

## References

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. Journal of Official Statistics, 3, 141-153.

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

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

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. Journal of Statistical Planning and Inference, 7, 155-170.

Särndal, C.-E., Swensson, B. and Wretman J.H. (1989). The weighted residual technique for estimating the variance of the finite population total. Biometrika, 76, 527-537.

Särndal, C.-E., Swensson, B. and Wretman J.H. (1992). Model assisted survey sampling. New York: Springer-Verlag.

# Statistical methods for sampling cross-classified populations under constraints 

Louis-Paul Rivest ${ }^{\mathbf{1}}$


#### Abstract

The article considers sampling designs for populations that can be represented as a $N \times M$ matrix. For instance when investigating tourist activities, the rows could be locations visited by tourists and the columns days in the tourist season. The goal is to sample cells $(i, j)$ of the matrix when the number of selections within each row and each column is fixed a priori. The $i^{\text {th }}$ row sample size represents the number of selected cells within row $i$; the $j^{\text {th }}$ column sample size is the number of selected cells within column $j$. A matrix sampling design gives an $N \times M$ matrix of sample indicators, with entry 1 at position $(i, j)$ if cell $(i, j)$ is sampled and 0 otherwise. The first matrix sampling design investigated has one level of sampling, row and column sample sizes are set in advance: the row sample sizes can vary while the column sample sizes are all equal. The fixed margins can be seen as balancing constraints and algorithms available for selecting such samples are reviewed. A new estimator for the variance of the Horvitz-Thompson estimator for the mean of survey variable $y$ is then presented. Several levels of sampling might be necessary to account for all the constraints; this involves multi-level matrix sampling designs that are also investigated.


Key Words: Balanced sampling; Creel surveys; Cube method; Multi-level sampling; Monte Carlo simulation; Variance estimation.

## 1. Introduction

Sampling from cross-classified populations raises interesting statistical issues, see Juillard, Chauvet and Ruiz-Gazen (2017) for a recent discussion. When each cell of the cross-classification contains a single unit, the population to sample has size $N M$ as it can then be viewed as a $N \times M$ matrix. The sample consists of cells $(i, j)$ of the population matrix; this defines the $N \times M$ matrix $\mathbf{Z}$ of sample indicators, with $Z_{i j}=1$ if cell $(i, j)$ is selected and $Z_{i j}=0$ otherwise. We focus on designs where the number of selections within each row and each column is fixed a priori. We define the $i^{\text {th }}$ row sample size as the number of selected cells within row $i$ while the $j^{\text {th }}$ column sample size is the number of selected cells within column $j$. This work studies matrix sampling designs for which the row and the column sample sizes are predetermined; the row sample sizes vary from one row to the next while all the column sample sizes are equal. This generalizes a stratified sampling design that would apply if the sample sizes for only one dimension, either row or column, were fixed. Multi-level generalizations of the proposed design are also introduced.

Populations having a matrix format occur, for instance, when pooling tourists (Deville and MaumyBertrand, 2006) and in creel surveys (Kozfkay and Dillon, 2010); time, that is days, is one dimension of the matrix and the other one is location, such as venues frequented by tourists and fishing sites. Figure 1.1 presents a sample selected in a $10 \times 20$ population matrix with row sample sizes equal to $m=8$ and column sample sizes equal to $n=4$. The black entries identify the cells $(i, j)$ that are selected, for which $Z_{i j}=1$. Having constraints on the row and the column sample sizes is useful in many contexts. In Ida, Rivest and

1. Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, 1045 rue de la médecine, Québec, Canada, G1V $0 A 6$. E-mail: louis-paul.rivest@mat.ulaval.ca.

Daigle (2018)'s population matrix the rows are sites and the columns are days. The column sample size $n$ is the number of sites that can be visited in one day while the row sample sizes $\left\{m_{i}: i=1, \ldots, N\right\}$ are related to sites' importance. The sample matrix presented in Figure1.1 also applies for planning $N$ repeated surveys in a population of size $M$. The $i^{\text {th }}$ row of $\mathbf{Z}$ identifies the units selected in survey $i$. The row sample sizes are determined by the objectives of the individual surveys. The fixed column sample size creates a dependency between the row samples that ensures their coordination (Matei and Tillé, 2005): the response burden is shared equally among all population units. This work focusses mostly on the first application as it investigates the sampling properties of the Horvitz-Thompson estimator of the mean of the survey variable $y$ over the $N M$ population units: $y_{i j}$ is the value of the survey variable for cell $(i, j), i=1, \ldots, N$; $j=1, \ldots, M$. For instance, $y_{i j}$ is the total number of hours of fishing at site $i$ on day $j$ in Ida et al. (2018). The goal is to estimate the population mean, $\bar{y}=\sum_{i, j} y_{i j} /(N M)$ using the Horvitz-Thompson estimator $\hat{\bar{y}}=\sum_{i=1}^{N} \hat{\bar{y}}_{i \bullet} / N$, where $\hat{\bar{y}}_{i \bullet}$ is the sample mean for row $i$.

Figure 1.1 Sample units, in black, drawn out of a $10 \times 20$ population matrix.


The set of $N \times M 0$-1 matrices $\mathbf{Z}$ with row totals given by $Z_{i \bullet}=m_{i}, i=1, \ldots, N$ and column totals $Z_{\cdot j}=n, j=1, \ldots, M$ is fairly large (Barvinok, 2010). The goal of the matrix sampling design is to select uniformly among that set, thus all the $N \times M$ matrices $\mathbf{Z}$ fulfilling the constraints on the row and the column sample sizes are equally likely to be selected. This can be achieved using the cube algorithm (Deville and Tillé, 2004); the constraints on the margins are then interpreted as balancing constraints. Rivest and Ebouele (2020) also discuss other sampling algorithms such as hypergeometric sampling and sampling through a Markov chain defined on the set of acceptable matrices $\mathbf{Z}$.

The next section summarizes the findings of Rivest and Ebouele (2020) for the sampling design illustrated in Figure 1.1. A new approach, called the conditional approach, for the evaluation of the sampling properties of the Horvitz-Thompson estimator is proposed in Section 3. Section 4 suggests a new estimator for the between row covariances. Section 5 considers a hierarchical sampling design to select the rows of the population matrix. It also investigates the properties of the Horvitz-Thompson estimator of the population mean of the survey variable for this new design.

## 2. The properties of the matrix sampling design with fixed row and column totals

Consider a sampling design that selects the $N \times M$ matrix of sample indicators $\left\{Z_{i j}\right\}$ uniformly among all the $0-1$ matrix with row totals given by $\left\{m_{i}\right\}$ and fixed column total $n$ for $j=1, \ldots, M$ where $\sum_{i=1}^{N} m_{i}=M n$. Given a matrix $\mathbf{Z}$ that meets these constraints, any permutation of the columns of $\mathbf{Z}$ is an acceptable matrix of sample indicators. This implies that the sampling design for selecting the units in row $i$ is without replacement simple random sampling of $m_{i}$ units among $M$ as all possible samples have the same probability of being chosen. This also entails that the sampling designs for selecting $n$ units among $M$ in a column are identical. It is a without replacement design with variable selection probabilities and a fixed sample size.

The unit inclusion probabilities for cell $(i, j)$ of the proposed design are $\pi_{i j}=\gamma_{i}=m_{i} / M$ for $i=1, \ldots, N$ and $j=1, \ldots, M$. The joint inclusion probabilities of two cells $(i, j)$ and $(k, \ell)$ is $\pi_{i j, k}$. They can be expressed in terms of $\gamma_{i k}$ the joint inclusion probabilities when sampling units within a column. They are given by

$$
\pi_{i j, k \ell}=\left\{\begin{array}{cl}
\gamma_{i k} & i \neq k, j=\ell,  \tag{2.1}\\
m_{i}\left(m_{i}-1\right) /\{M(M-1)\} & i=k, j \neq \ell, \\
\frac{m_{i} m_{k}}{M(M-1)}-\frac{\gamma_{i k}}{M-1} & i \neq k, j \neq \ell .
\end{array}\right.
$$

The joint selection probabilities in the same row or in the same column of a matrix are deduced from the row and the column sampling designs discussed above. When $i \neq k, j \neq \ell$ the joint selection probabilities only depend on $(i, k)$ because of the column exchangeability. One has

$$
m_{i} m_{k}=\sum_{j \neq \ell} E\left(Z_{i j} Z_{k \ell}\right)+\sum_{j} E\left(Z_{i j} Z_{k j}\right)=M(M-1) \pi_{i j, k \ell}+M \gamma_{i k} .
$$

Solving this equation gives the general formula for $\pi_{i j, k e}$.
The joint probability for sampling rows $i \neq k$ in two different columns, $j \neq \ell$, is larger than that for sampling $i$ and $k$ in the same column as

$$
\pi_{i j, k \ell}=\frac{\gamma_{i k}}{M-1}\left(M \frac{\gamma_{i} \gamma_{k}}{\gamma_{i k}}-1\right)>\gamma_{i k},
$$

provided that

$$
\begin{equation*}
\gamma_{i} \gamma_{k}>\gamma_{i k} . \tag{2.2}
\end{equation*}
$$

This condition ensures that the Sen-Yates-Grundy variance estimator for the Horvitz-Thompson estimator of a column total is positive. The condition $\gamma_{i} \gamma_{k}>\gamma_{i k}$ is satisfied by the conditional Poisson sampling design (Chen and Dempster, 1994). It is conjectured that (2.2) is true for the design for sampling units within a column as this design converges to a conditional Poisson sampling design when $M$ goes to $\infty$ and $N$ is fixed (Rivest, 2021). If true, (2.2) would also mean that the probability $\gamma_{i k}$ that a column is sampled in both row $i$ and row $k$ is less than $\gamma_{i} \gamma_{k}$, the inclusion probability if the rows were sampled independently. Thus the fixed row and column totals create a negative coordination between row samples, see Grafström and Matei (2015) for a discussion of positive and negative coordination between samples.

Another interesting result is that, for $i \neq k$ and $j \neq \ell$,

$$
\operatorname{Cov}\left(Z_{i j}, Z_{k \ell}\right)=-\frac{1}{M-1} \operatorname{Cov}\left(Z_{i j}, Z_{k j}\right)=-\frac{1}{M-1}\left(\gamma_{i k}-\gamma_{i} \gamma_{k}\right)
$$

This result remains true when $i=k$ provided that $\gamma_{i i}$ is defined as being equal to $\gamma_{i}=m_{i} / M$. Indeed, one has $\operatorname{Cov}\left(Z_{i j}, Z_{i \ell}\right)=-\gamma_{i}\left(1-\gamma_{i}\right) /(M-1)$ which is minus the variance of $Z_{i j}$ divided by $M-1$. This is used to prove that the covariance between $\hat{\bar{y}}_{i \bullet}$ and $\hat{\bar{y}}_{k_{\bullet}}$, the sample means for survey variable $y$ in rows $i$ and $k$ is

$$
\operatorname{Cov}\left(\hat{\bar{y}}_{i \bullet}, \hat{\bar{y}}_{k \bullet}\right)=\Delta_{i k} S_{i k}, \quad i, k=1, \ldots, N
$$

where $\Delta_{i k}=\left\{\gamma_{i k} /\left(\gamma_{i} \gamma_{k}\right)-1\right\} / M$ is the $(i, k)$ entry of a $N \times N$ matrix $\Delta$ and $S_{i k}$ is the covariance between rows $i$ and $k, S_{i k}=\sum_{j=1}^{M}\left(y_{i j}-\bar{y}_{i \bullet}\right)\left(y_{k j}-\bar{y}_{k \bullet}\right) /(M-1), i, k=1, \ldots, N$ and $\bar{y}_{i \bullet}=\sum_{j=1}^{M} y_{i j} / M$. This result is used to evaluate the variance of $\hat{\bar{y}}=\sum_{i=1}^{N} \hat{\bar{y}}_{i \bullet} / N$, the Horvitz-Thompson estimator of the mean of $y$ :

$$
\begin{equation*}
\operatorname{Var}(\hat{\bar{y}})=\frac{\operatorname{tr}(\mathbf{S} \Delta)}{N^{2}}, \tag{2.3}
\end{equation*}
$$

where $\mathbf{S}$ is the $N \times N$ covariance matrix of the $y$ column vectors and $\operatorname{tr}$ is the trace operator. If (2.2) holds, then the off-diagonal entries of $\Delta$ are negative and (2.3) is smaller than $\sum_{i} \Delta_{i i} S_{i i} / N^{2}$, the variance of the stratified estimator obtained when sampling the rows independently when the between row covariances, $S_{i k}$, are positive. Note also that to evaluate (2.3) one needs numerical values for the joint selection probabilities $\gamma_{i k}$. These can be obtained through simulations or using a numerical algorithm to evaluate the joint selection probabilities for the conditional Poisson sampling design, see Tillé (2006) and Rivest (2021), that can be used as an approximation. A conditional variance formula, that does not use the joint selection probabilities $\gamma_{i k}$ is proposed in the next section.

An obvious choice for an estimator of (2.3) would be the Sen-Yates-Grundy variance estimator. Unfortunately, the condition for it to be positive fails. Indeed the joint selection probability in two different rows and two different columns, $\pi_{i j, k \ell}$ satisfies

$$
\pi_{i j, k \ell}-\pi_{i j} \pi_{k \ell}=-\frac{1}{M-1}\left(\gamma_{i k}-\gamma_{i} \gamma_{k}\right)
$$

which as argued in the discussion of (2.2) should be positive. Thus the Sen-Yates-Grundy variance estimator can be negative and an alternative estimator is needed. A plug-in estimator for (2.3) is given in Section 4. It demands the estimation of the $N(N-1) / 2$ between row covariances. The negative coordination between row samples renders the construction of estimators difficult. The proposals in Rivest and Ebouele (2020) are not really satisfactory as they give biased estimators when the column sample sizes are small. Alternative estimators are considered in Section 4.

### 2.1 Extensions to unequal column sample sizes

This section assumes that both the row and the column sample sizes of the matrix $\mathbf{Z}$ vary. They are given by $\left\{m_{i}: i=1, \ldots, N\right\}$ and $\left\{n_{j}: j=1, \ldots, M\right\}$. The set of possible samples consists of $N \times M$ 0-1 matrices with fixed row and column totals. All the algorithms reviewed in Rivest and Ebouele (2020) can be used to select the sample uniformly in that set. The resulting design is however rather complex when $n_{j}$ takes several positive values. There are no closed form expressions for the unit inclusion probability of cell $(i, j)$ and for the joint inclusion of cells $(i, j)$ and $(j, \ell)$ and there does not seem to be a manageable expression for the variance of the Horvitz-Thompson estimator. The limiting sampling design within a column converges, as $M$ goes to infinity, to a generalization of the conditional Poisson sampling design with untractable single inclusion probabilities, see Rivest (2021). Thus, to implement a design with varying column sample sizes, a simple solution is to stratify by column sample size. Independent matrix designs are then used to select the matrix sample in each stratum.

The matrix sampling design of the previous section can be extended to situations where the two possible column sample sizes are either 0 or $n$. Suppose that out of $M$ columns, $M_{0}<M$ have a non-null sample. The row sample sizes $\left\{m_{i}: i=1, \ldots, N\right\}$ satisfy $\sum_{i} m_{i}=M_{0} n$. The selection of a matrix $\mathbf{Z}$ of sample indicators proceeds in two steps. Step 1 uses without replacement simple random sampling to select the $M_{0}$ columns with non-null samples and a matrix design is used at step 2 to select the sampled cells in the $M_{0}$ columns chosen at step 1.

Let $\gamma_{i k}$ be the conditional joint selection probability for rows $i$ and $k$ given that the column has been selected at step 1 . The findings of the previous section are easily generalized to this new design. For instance

$$
\operatorname{Cov}\left(Z_{i j}, Z_{k \ell}\right)=-\frac{1}{M-1} \operatorname{Cov}\left(Z_{i j}, Z_{k j}\right)=-\frac{1}{M-1}\left(M_{0} \gamma_{i k} / M-m_{i} m_{k} / M^{2}\right) .
$$

In addition variance formula (2.3) holds with a matrix $\Delta^{g}$ defined by $\Delta_{i k}^{g}=\left\{M M_{0} \gamma_{i k} /\left(m_{i} m_{k}\right)-1\right\} / M$ for $i, k=1, \ldots, n$, provided that one sets $\gamma_{i i}=m_{i} / M_{0}$.

## 3. A conditional matrix sampling design

This section discusses a conditional sampling design for which the matrix of sample indicators $\mathbf{Z}$ is fixed, up to a random permutation of its columns. It derives a conditional alternative to (2.3), the variance of the Horvitz-Thompson estimator.

Let $\mathbf{Z}_{0}$ be a $0-1$ matrix with row sums equal to $\left\{m_{i}: i=1, \ldots, N\right\}$ and column sums all equal to $n$. Suppose that the random matrix of sample indicators $\mathbf{Z}$ is obtained by randomly permuting the columns of $\mathbf{Z}_{0}$. This conditional sampling design shares many of the properties discussed in Section 2. The design for sampling units within row $i$ is simple random sampling of $m_{i}$ units in a population of size $M$. The design for sampling units within a column is the same for each column. This design gives a probability of $1 / M$ to each of the columns of $\mathbf{Z}_{0}$ and the probability for selecting unit $i$ and $k$ within a column is

$$
\begin{equation*}
\gamma_{i k}^{c}=\sum_{j=1}^{M} Z_{0 i j} Z_{0 k j} / M, \tag{3.1}
\end{equation*}
$$

where $Z_{0 i j}, Z_{0 k j}$ are the entries $(i, j)$ and $(k, j)$ of $\mathbf{Z}_{0}$ and exponent $c$ stands for conditional. For this design (2.1) holds with $\gamma_{i k}$ replaced by the conditional joint selection probability $\gamma_{i k}^{c}$. In some instances, the sampling design proposed in Section 2 is a conditional sampling design. This occurs when the column sample size is either $n=1$ or $n=N-1$ since all the possible sample indicator matrices $\mathbf{Z}$ are then equal up to a permutation of their columns.

The conditional variance of the Horvitz-Thompson estimator $\hat{\bar{y}}$ is a function of the $N \times N$ matrix $\Delta^{c}$ whose ( $i, k$ ) entry is equal to $\Delta_{i k}^{c}=\left\{\gamma_{i k}^{c} /\left(\gamma_{i} \gamma_{k}\right)-1\right\} / M, \gamma_{i k}^{c}$ is defined in (3.1) and $\gamma_{i}=m_{i} / M$, as defined in Section 2. It is given by

$$
\begin{equation*}
\operatorname{Var}_{c}(\hat{\bar{y}})=\frac{\operatorname{tr}\left(\mathbf{S} \mathbf{S}^{c}\right)}{N^{2}} . \tag{3.2}
\end{equation*}
$$

Given a random matrix $\mathbf{Z}$ of sample indicators obtained with one of the algorithms presented in Rivest and Ebouele (2020), one can evaluate the conditional joint inclusion probabilities using (3.1), applied to the matrix $\mathbf{Z}$. Then the conditional variance formula (3.2) is simpler than (2.3) as it does not require the evaluation of the unconditional joint selection probabilities $\gamma_{i k}$. The derivation of a simple variance formula is the main application of the conditional approach.

The conditional matrix sampling design is a low entropy design and it may happen that (3.2) cannot be estimated. This occurs when one of the conditional joint selection probabilities (3.1) is equal to 0 . Consider, for instance, the design in the Monte Carlo simulations of Section 4.1. It has $N=9, M=36, n=2$ and row sample sizes $m_{i}$ varying between 6 and 11. This design involves $N(N-1) / 2=36$ joint selection probabilities and it not possible to find a matrix $\mathbf{Z}_{0}$ with row totals varying between 6 and 11 for which the 36 values of (3.1) are positive. For this design, only (2.3) can be estimated.

## 4. A new estimator for the between row covariance

This section suggests new estimators for the $N \times N$ covariance matrix $\mathbf{S}$ for $y$. The diagonal elements of $\mathbf{S}$ are easily estimated using the row sample variances. Rivest and Ebouele (2020) use the columns that are sampled in both rows, $i$ and $k$, to estimate the covariance $S_{i k}$. The joint sample size for rows $i$ and $k$ is often less than 2, considering the negative coordination between row samples noted in Section 2. Thus many covariances cannot be estimated using this approach and an alternative estimation strategy is proposed
in this section. It gives nearly unbiased estimators of the variances (2.3) and (3.2) of the Horvitz-Thompson estimator

The proposed covariance estimator relies on the following expression for the covariance as a U-statistic,

$$
\begin{align*}
S_{i k} & =\frac{1}{M-1} \sum_{j=1}^{M}\left(y_{i j}-\bar{y}_{i \bullet}\right)\left(y_{k j}-\bar{y}_{k \bullet}\right) \\
& =\frac{1}{2 M(M-1)}\left\{\sum_{j \neq \ell}^{M}\left(y_{i j}-y_{k \ell}\right)^{2}-(M-1) \sum_{j=1}^{M}\left(y_{i j}-y_{k j}\right)^{2}\right\}, \tag{4.1}
\end{align*}
$$

where $i \neq k$. See Appendix for a derivation of (4.1).
The new covariance estimator uses the joint selection probabilities $\gamma_{i k}$ of Section 2 that are assumed to be strictly positive, to construct estimators of the two terms in (4.1). This yields

$$
\begin{align*}
\hat{S}_{i k} & =\frac{1}{2 M(M-1)}\left\{\sum_{j \neq \ell}^{M} \frac{Z_{i j} Z_{k \ell}\left(y_{i j}-y_{k \ell}\right)^{2}}{M \gamma_{i} \gamma_{k} /(M-1)-\gamma_{i k} /(M-1)}-(M-1) \sum_{j=1}^{M} \frac{Z_{i j} Z_{k j}\left(y_{i j}-y_{k j}\right)^{2}}{\gamma_{i k}}\right\}  \tag{4.2}\\
& =\frac{1}{2 M}\left\{\sum_{j \neq \ell}^{M} \frac{Z_{i j} Z_{k \ell}\left(y_{i j}-y_{k \ell}\right)^{2}}{M \gamma_{i} \gamma_{k}-\gamma_{i k}}-\sum_{j=1}^{M} \frac{Z_{i j} Z_{k j}\left(y_{i j}-y_{k j}\right)^{2}}{\gamma_{i k}}\right\} .
\end{align*}
$$

The plug-in unbiased variance estimator for (2.3) is simply $v_{\mathrm{PI}}=\operatorname{tr}(\hat{\mathbf{S}} \Delta) / N^{2}$, where $\hat{\mathbf{S}}$ has entries given by (4.2). Observe that this covariance estimator can be constructed for the unconditional and the conditional sampling designs presented in Sections 2 and 3. For a conditional sampling design, one replaces $\gamma_{i k}$ by $\gamma_{i k}^{c}$, see (3.1), in (4.2) as long as $\gamma_{i k}^{c}>0$.

Covariance estimator (4.2) is very variable as its second term involves a division by $\gamma_{i k}$ that can be very small. This leads to a covariance matrix estimator $\hat{\mathbf{S}}$ which is not positive definite and to an estimator for (2.3) that can, on some rare occasions, be negative. A more stable estimator can be obtained by assuming that all the between row correlations are equal. An estimator of the common correlation is then

$$
\begin{equation*}
\hat{\rho}=\frac{\sum_{i>k} \hat{S}_{i k}}{\sum_{i>k} \sqrt{\hat{S}_{i i} \hat{S}_{k k}}} \tag{4.3}
\end{equation*}
$$

where $\hat{S}_{i i}$ and $\hat{S}_{k k}$ are the sample variances for row $i$ and row $k$ respectively. An alternative covariance estimator is $\hat{S}_{i k}^{\text {(c) }}=\hat{\rho} \sqrt{\hat{S}_{i i} \hat{S}_{k k}}$. It leads to an equal correlation estimator, $v_{\mathrm{ec}}$, for (2.3). A stratified variance estimator, valid if the rows were sampled independently, $v_{\text {str }}=\sum_{i} \Delta_{i i} \hat{S}_{i i} / N^{2}$, is also included as a benchmark in the Monte Carlo simulations that are presented next.

### 4.1 A Monte Carlo investigation of the sampling properties of the new variance estimators

The sampling properties of variance estimator (4.2) and its equal correlation alternative are investigated in two population matrices. The first one has $N=9, M=36$, column sample size is $n=2$ and row sample sizes $m_{i}$ are ( $11,10,7,10,11,5,6,6,6$ ). In the second one, $M$ and the row sample sizes are doubled while
$N=9$ and $n=2$ are unchanged. The $y$ variable in cell $(i, j)$ has a log-normal distribution with expectation $m_{i}$ and variance $1.72 \times m_{i}^{2}$. It is given by

$$
y_{i j}=m_{i} \exp \left(a_{j}+e_{i j}\right) \quad i=1, \ldots N, j=1, \ldots M
$$

where the column effect $a_{j}$ and the errors $e_{i j}$ are independent variables respectively distributed according to a $N\left(-\sigma_{a}^{2} / 2, \sigma_{a}^{2}\right)$ and a $N\left(-\sigma_{e}^{2} / 2, \sigma_{e}^{2}\right)$ distributions where $\sigma_{a}^{2}+\sigma_{e}^{2}=0.8$. Simulations with $\sigma_{a}^{2}=0,0.2$ and 0.4 are reported. This corresponds to a between row correlation $\rho=\left\{\exp \left(\sigma_{a}^{2}\right)-1\right\} /\left\{\exp \left(\sigma_{a}^{2}+\sigma^{2}\right)-1\right\}$ of respectively $0,0.18$ and 0.40 . Table 4.1 uses 6 simulated populations and the moments of variance estimators are calcuated using $10^{4}$ Monte Carlo samples. These Monte Carlo samples are drawn using the MCMC swap algorithm of Oksanen, Blanchet, Friendly, Kindt, Legendre, McGlinn, Minchin, O'Hara, Simpson, Solymos, Stevens, Szoecs and Wagner (2020) discussed in Rivest and Ebouele (2020) to which the reader is referred for more details on the simulations. The results are reported in Table 4.1.

When $\rho=0$ the three variance estimators are nearly unbiased in Table 4.1. The simple stratified variance estimator does not capture the positive correlation between rows and over-estimates the variance when $\rho>0$. The plug-in and equal correlation estimators are unbiased for the 6 populations considered in Table 4.1. The standard deviations are also revealing. That for $v_{\text {str }}$ is small as this estimator does not depend on the covariance estimators. The plug-in estimator is the most variable while the equal correlation that is based on the average covariance has a smaller standard deviation.

Table 4.1
The variance of $\hat{\bar{y}}$ evaluated using (2.3) and the expectations, and standard deviations between parenthesis, of three variance estimators, the stratified estimator (str), the plug-in estimator (PI), the equal correlation estimator (ec).

| $\boldsymbol{M}$ | $\boldsymbol{\rho}$ | $\operatorname{var}(\hat{\overline{\boldsymbol{y}}})$ | $\boldsymbol{v}_{\text {str }}$ | $\boldsymbol{v}_{\text {PI }}$ | $\boldsymbol{v}_{\text {ec }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 36 | 0 | 0.932 | $0.913(0.684)$ | $0.930(0.757)$ | $0.928(0.746)$ |
| 36 | 0.18 | 0.653 | $0.800(0.258)$ | $0.654(0.288)$ | $0.657(0.280)$ |
| 36 | 0.40 | 0.544 | $0.730(0.306)$ | $0.542(0.338)$ | $0.543(0.329)$ |
| 72 | 0 | 0.486 | $0.506(0.225)$ | $0.485(0.252)$ | $0.486(0.245)$ |
| 72 | 0.18 | 0.350 | $0.424(0.136)$ | $0.350(0.151)$ | $0.349(0.145)$ |
| 72 | 0.40 | 0.334 | $0.564(0.182)$ | $0.334(0.190)$ | $0.332(0.182)$ |

To explain the near unbiasedness of $v_{\text {ec }}$ in Table 4.1, one easily checks that this estimator is indeed unbiased when all the row sample sizes are constant and equal to $m$. The matrix $\Delta$ can then be expressed in terms of the $N \times N$ identity matrix $\mathbf{I}_{N}$ and a $N \times 1$ vector of 1 's, $\mathbf{1}_{N}$, as $\Delta=(N-n)\left(\mathbf{I}_{N}-\right.$ $\left.\mathbf{1}_{N} \mathbf{1}_{N}^{\top} / N\right) /\{m(N-1)\}$ and (2.3) only involves $\sum_{i \neq k} S_{i k}$. It is easily checked that the equal correlation estimator of that sum is unbiased as (4.2) is unbiased.

## 5. A multi-level matrix design

This section suggests multi-level generalizations of the matrix sampling designs of Sections 2 and 3. It involves clusters of rows and level 1 selects a sample of clusters for each column. This is done using a level 1 random 0-1 matrix with fixed margins. A population matrix is then created for each level 1 cluster of rows;
the number of rows in the matrix is the size of the cluster and the number of columns is equal to the column total for that cluster in the level 1 sample indicator matrix. Level 2 selection is done independently in the population matrices of each level 1 cluster. Level three sampling is done in a similar way. This section focusses on two-level designs and uses the following notation:

- $\quad N^{(1)}$ is the number row clusters. The size of the level 1 matrix of sample indicators is $N^{(1)} \times M$; the row and column sample sizes are $\left\{m_{i}^{(1)}: i=1, \ldots, N^{(1)}\right\}$ and $n^{(1)}$;
- $\quad N_{i}^{(2)}, i=1, \ldots, N^{(1)}$ are the sizes of the level 1 row clusters. The level 2 matrix of sample indicators for cluster $i$ has dimension $N_{i}^{(2)} \times m_{i}^{(1)}$; the row and column sample sizes are $\left\{m_{k(i)}^{(2)}: i=\right.$ $\left.1, \ldots, N_{i}^{(2)}\right\}$ and $n^{(2)}$;
- The matrices of sample indicators can be combined in a matrix with $N=\sum N_{i}^{(2)}$ rows and $M$ columns. The column sample size is $n=n^{(1)} \times n^{(2)}$ and the sample size for the $k^{\text {th }}$ row of cluster $i$ is $m_{k(i)}^{(2)}$.

Note $\sum_{i} m_{i}^{(1)} / n_{(1)}=M$ while for $i=1, \ldots, N^{(1)}$, the sum of the row totals of cluster $i$, has to be multiple of $n^{(2)}$ as

$$
\begin{equation*}
\sum_{k} m_{k(i)}^{(2)} / n^{(2)}=m_{i}^{(1)} . \tag{5.1}
\end{equation*}
$$

Figure 5.1 gives a sample obtained with two levels of sampling in a matrix similar to that of Figure 1.1. One has $N=10, M=20$, while the row sample sizes are $m=8$ and column sample sizes are $n=4$. The first level of sampling involves three clusters of rows, of respective size 3 , 3 , and 4 . The $4 \mathbf{Z}$ matrices used to construct Figure 5.1 are given in Appendix.

Figure 5.1 A sample drawn from a $10 \times 20$ population matrix using a two level matrix sampling design.


The hierarchical design proposed in this section is useful to accommodate constraints. For instance, when the rows are sites, a cluster is a set of neighboring sites than can be visited on the same day. When planning repeated, say monthly, surveys for the same population over one year, clusters could be used to ensure that a unit is surveyed only either in the first six months or in the last six months of the year.

The hierarchical design shares a key property with the single level sampling designs presented in Sections 2 and 3. Given a matrix of sample indicators that meets the row and the column constraints, such as that given in Figure 5.1, any permutation of the columns of that matrix gives an acceptable sample matrix. Thus, under this hierarchical sampling scheme, the sampling design for selecting units within a row is without replacement simple random sampling. The selection probability for row $k$ of cluster $i$ is $m_{k(i)}^{(2)} / M$. This is the product of the level 1 selection probability for cluster $i, m_{i}^{(1)} / M$ times the level 2 selection probability $m_{k(i)}^{(2)} / m_{i}^{(1)}$ in row $k$ of cluster $i$. All the designs for sampling units within a column are identical. It has two levels and its joint selection probabilities $\left\{\gamma_{k(i), \ell(j)}^{(2)}: i, j=1, \ldots, N^{(1)}, k=1, \ldots, N_{i}^{(2)}\right.$, $\left.\ell=1, \ldots, N_{j}^{(2)}\right\}$, are calculated as follows:

- For rows in the same cluster, $i=j$ then $\gamma_{k(i) \ell(i)}^{(2)}=\left(m_{i}^{(1)} / M\right) \gamma_{i, k, \ell}^{(2)}$ where $\gamma_{i, k, \ell}^{(2)}$ is the joint selection probabilities for rows $k$ and $\ell$ of cluster $i$ at the second level of sampling.
- For rows in different clusters, $i \neq j$, then $\gamma_{k(i) \ell(j)}^{(2)}=\gamma_{i, j}^{(1)}\left(m_{k(i)}^{(2)} / m_{i}^{(1)}\right)\left(m_{\ell(j)}^{(2)} / m_{j}^{(1)}\right)$ is the product of the joint selection probability for these two clusters at level 1 , times the level 2 single selection probabilities for rows $k$ and $\ell$ in clusters $i$ and $j$.

In this construction the joint selection probabilities, for levels 1 and 2, can be approximated by those of conditional Poisson sampling as discussed in Section 2.

It is now convenient to change the notation and to let $i, k=1, \ldots, N$ denote rows of the design matrix, in agreement with Sections 2 and 3. The selection probability in row $i$ is $\gamma_{i}=m_{i} / M$ and the joint selection probabilities are $\left\{\gamma_{i k}^{h}\right\}$, where exponent $h$ means hierarchical. All the findings of Sections 2 and 3 apply to the hierarchical design provided that the joint selection probabilities $\left\{\gamma_{i k}\right\}$ are replaced by their multi-level alternatives, $\left\{\gamma_{i k}^{h}\right\}$. For instance (2.1) and (2.3) hold for this new design, when written is terms of $\left\{\gamma_{i k}^{h}\right\}$ and the matrix $\Delta^{h}$ whose $(i, k)$ entry is $\Delta_{i k}^{h}=\left\{\gamma_{i k}^{h} /\left(\gamma_{i} \gamma_{k}\right)-1\right\} / M$. The variance estimators proposed in Section 4 applies to the hierarchical design proposed in this section provided that all the joint selection probabilities $\left\{\gamma_{i k}^{h}\right\}$ are strictly positive. The estimates of (2.3) and (3.2) are evaluated using the matrix $\Delta^{h}$ for the hierarchical design and the covariance estimator of Section 4.

### 5.1 A Monte Carlo investigation

This section revisits the Monte Carlo simulations of Section 4.1. The populations are sampled using a two level design: within each column the two rows sampled need to belong to the same cluster. The clusters of rows for level 1 consist of $\{1,2,3\},\{4,5,6\}$, and $\{7,8,9\}$ and the row sample sizes are $(11,10,7,10$, $11,5,6,6,6$ ), the same as those used in Section 4.1. At level $1, \mathbf{Z}$ is a $3 \times 36$ matrix. Its column totals are
$n=1$ while from (5.1), its row totals are ( $14,13,9$ ). The three $\mathbf{Z}$ matrices for level 2 have respectively 14 , 13, and 9 columns and three rows; their column total is $n=2$. For this problem the designs of Sections 2 and 3 are the same as the column totals of all the $\mathbf{Z}$ matrices are either 1 or $N-1$.

Since only one cluster is sampled in each column the joint selection probability $\gamma_{i k}^{h}$ is 0 for two rows $(i, k)$ belonging to different clusters. Thus the covariance $S_{i k}$ of the survey variable between these two rows is not estimable. Indeed only 9 of the $36(=8 \times 9 / 2)$ covariances can be estimated in the simulation design. The variances (2.3) and (3.2) are not estimable. The simulation study investigates the performance of the equal correlation estimator $v_{\mathrm{ec}}$ where the common correlation is estimated, through (4.3), using the 9 covariances that are estimable.

Table 5.1
The variance of $\hat{\bar{y}}$ evaluated using (2.3) and the expectations, and standard deviations between parenthesis, of two variance estimators, the stratified estimator (str) and the equal correlation estimator (ec).

| $\boldsymbol{M}$ | $\rho$ | $\operatorname{var}(\hat{\overline{\boldsymbol{y}}})$ | $\boldsymbol{v}_{\text {str }}$ | $\boldsymbol{v}_{\text {ec }}$ |
| :---: | :---: | :---: | :---: | :---: |
| 36 | 0 | 0.888 | $0.922(0.694)$ | $0.968(0.759)$ |
| 36 | 0.18 | 0.582 | $0.796(0.244)$ | $0.684(0.360)$ |
| 36 | 0.40 | 0.587 | $0.738(0.309)$ | $0.521(0.338)$ |
| 72 | 0 | 0.521 | $0.504(0.230)$ | $0.449(0.251)$ |
| 72 | 0.18 | 0.318 | $0.421(0.136)$ | $0.362(0.151)$ |
| 72 | 0.40 | 0.334 | $0.561(0.183)$ | $0.321(0.195)$ |

The population sampled are the same as those investigated in Section 4.1. Even if the sampling designs differ, the expectations and variances of the stratified variance estimator $v_{\text {str }}$ are identical in the two experiments. Indeed the $v_{\text {str }}$ entries for Tables 4.1 and 5.1 are the same up to Monte Carlo errors. The bias of the equal correlation estimator $v_{\mathrm{cc}}$ ranges between $3 \%$ and $20 \%$. It is smaller than that of $v_{\mathrm{str}}$ when $\rho>0$. Thus variance estimation is a problem when the first level column sample size is 1 . Estimator $v_{\text {str }}$ provides an upper bound for the variance while the validity of $v_{\mathrm{ec}}$ rests on an homogeneity assumption that can be verified, at least in part, by comparing the correlation coefficients that are estimable.

### 5.2 A complex example

This section discusses a complex sampling design presented in Ida et al. (2018) that involves a $54 \times 33$ population matrix. The goal is to estimate the fishing effort $\bar{y}$ over $M=33$ days at 9 sites, grouped in three clusters. In Table 5.2 the sites are numbered from 1 to 9 . Sites 1, 3, 7, and 9 have more fishermen; their planned number of visits is about twice that for the other sites. Each day a site can be visited at 6 time points, 2 in the morning (AM), 2 in the afternoon (PM) and 2 in the evening (EV). So for each day there are $N=9 \times 6=54$ site-time-points. The sampling design selects 4 site-time-points on each day, under two constraints: the sites visited must be in the same cluster and the four visits must be selected as two blocks of two visits in either the morning, the afternoon or the evening. Three levels of sampling are needed to
address these constraints. Because of these constraints it is not feasible to have the number of visits to the more important sites, $1,3,7$, and 9 , exactly equal to twice that at the other sites. The row totals given in Table 5.2 correspond to an approximate solution to the determination of the site-time-point sample sizes. Other approximate solutions are possible; those considered in Ida et al. (2018) are obtained by running the cube algorithm, modified for highly stratified populations by Hasler and Tillé (2014), for the three levels of sampling.

Appendix shows how to obtain a matrix of sample indicators for the $54 \times 33$ population matrix with column total $n=4$ and row totals given in Table 5.2. There are three level of sampling. Level 1 selects one cluster for each day (column), level 2 proceeds cluster by cluster and selects two time periods for each day it is visited. Level 3 sampling is applied within each cluster period; it selects the two sites that will be visited at the two time point in the period. Level three sampling is stratified: one site is selected for each time point. This involves $13 \mathbf{Z}$ matrices of sample indicators.

Table 5.2
Vector of 54 row totals for the $54 \times 33$ matrix $Z$ for the design of Ida et al. (2018).

| Cluster | Period | Time-point-site |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $1-1$ | $1-2$ | $1-3$ | $2-1$ | $2-2$ | $2-3$ | Tot |
| 1 | AM | 4 | 1 | 3 | 3 | 2 | 3 | 16 |
| 1 | PM | 3 | 2 | 3 | 3 | 2 | 3 | 16 |
| 1 | EV | 3 | 2 | 3 | 3 | 1 | 4 | 16 |
| 1 | Tot | 10 | 5 | 9 | 9 | 5 | 10 | 48 |
|  |  | $1-4$ | $1-5$ | $1-6$ | $2-4$ | $2-5$ | $2-6$ |  |
| 2 | AM | 1 | 2 | 2 | 2 | 1 | 2 | 10 |
| 2 | PM | 2 | 2 | 1 | 2 | 2 | 1 | 10 |
| 2 | EV | 2 | 2 | 2 | 2 | 2 | 2 | 12 |
| 2 | Tot | 5 | 6 | 5 | 6 | 5 | 5 | 32 |
|  |  | $1-7$ | $1-8$ | $1-9$ | $2-7$ | $2-8$ | $2-9$ |  |
| 3 | AM | 3 | 2 | 4 | 4 | 1 | 4 | 18 |
| 3 | PM | 3 | 2 | 3 | 3 | 1 | 3 | 16 |
| 3 | EV | 3 | 2 | 4 | 4 | 1 | 3 | 18 |
| 3 | Tot | 9 | 6 | 11 | 11 | 5 | 10 | 52 |

In Table 5.2, equation (5.1) means that the total sample size for each cluster is a multiple of 4 while that for each of the 9 cluster-periods is even. Once a matrix of sample indicators has been selected, either with the cube algorithm or by selecting the 13 random $\mathbf{Z}$ matrices implementing the hierarchical design presented in the appendix, the conditional approach of Section 3 is a relatively straightforward method to estimate the variance. This involves two $54 \times 54$ matrices, an estimated covariance matrix $\hat{\mathbf{S}}$ and $\Delta^{\mathbf{c}}$, evaluated using (3.1). To account for the non estimability of some covariances found in the example considered in Section 5.1, the two variance estimators investigated in the simulation study reported in Table 5.1 could be used. Taking $\hat{\mathbf{S}}$ as a diagonal matrix of row variances gives the stratified variance estimator. The equal correlation estimator $\hat{\mathbf{S}}$ can also be evaluated. Thus two methods of variance estimation are available for this complex problem.

## 6. Discussion

Many samplings problems face operational constraints that need to be addressed when designing a survey, see for instance Vallée, Ferland-Raymond, Rivest and Tillé (2015) for a forestry example. One strategy to address these constraints is to use the cube method, possibly within a multi-level design, after a careful specification of the selection probabilities. This paper proposes an alternative strategy for crossclassified populations where the constraints can be expressed in terms of fixed row and column sample sizes. As illustrated in Section 5.2, this strategy involves setting up a population matrix and target row totals for the matrix of sample indicators. Sample selection is done by selecting relatively small sample indicator matrices $\mathbf{Z}$ uniformly over the set of feasible matrices at each level of the design. The hypergeometric sampling algorithm of Rivest and Ebouele (2020) is suited for this problem as it does not need an initial value $\mathbf{Z}_{0}$, that is required by MCMC algorithms. One advantage of this approach is that the constraints on row and column totals are always verified while the cube method sometimes fails to meet them exactly. An interesting feature of the methodology proposed in this paper is the availability of variance estimators. If, at all the levels, the column sample sizes are larger than 2 , then an unbiased variance estimator for the HorvitzThompson estimator of the mean of the survey variable can be constructed.

## Acknowledgements

This project benefitted from the financial assistance of the Canada Research Chair in Statistical Sampling and Data Analysis and from a discovery grant from the Natural Sciences and Engineering Research Council of Canada. The constructive comments of the special editor and of the referees are gratefully acknowledged. I am also grateful to Anne-Sophie Julien, who contributed to the construction of the sample design proposed in Section 5.2.

## Appendix:

## Proof of (4.2)

One has

$$
\begin{aligned}
\frac{1}{2 M(M-1)}\left\{\sum_{j \neq \ell}^{M}\right. & \left.\left(y_{i j}-y_{k \ell}\right)^{2}-(M-1) \sum_{j}^{M}\left(y_{i j}-y_{k j}\right)^{2}\right\} \\
& =\frac{1}{2 M(M-1)}\left\{\sum_{j, \ell=1}^{M}\left(y_{i j}-y_{k \ell}\right)^{2}-M \sum_{j}^{M}\left(y_{i j}-y_{k j}\right)^{2}\right\} \\
& =\frac{1}{M-1}\left\{\sum_{j=1}^{M} y_{i j} y_{k j}-\frac{\sum_{j=1}^{M} y_{i j} \sum_{\ell=1}^{M} y_{k \ell}}{M}\right\} \\
& =S_{i k} .
\end{aligned}
$$

## The matrices of sample indicators needed to construct Figure 5.1

The $3 \times 20$ matrix for selecting clusters $Z 3, Z 2$, and $Z 1$ respectively within each column is given by

$$
\mathbf{Z}_{1}=\left(\begin{array}{llllllllllllllllllll}
1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1
\end{array}\right) .
$$

The three matrices of sample indicators for level 2 sampling in respectively zones 3,2 and 1 are

$$
\begin{aligned}
& \mathbf{z}_{Z 3}=\left(\begin{array}{llllllllllllllll}
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}\right) \\
& \mathbf{z}_{Z 2}=\left(\begin{array}{llllllllllll}
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0
\end{array}\right) \mathbf{z}_{Z 1}=\left(\begin{array}{llllllllllll}
1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0
\end{array}\right) .
\end{aligned}
$$

The $(10,5)$ and $(8,5)$ entries in Figure 5.1 are equal to 1 . This information can be retrieved from $\mathbf{Z}_{1}$ and $\mathbf{Z}_{z 1}$ : in the fifth column of $\mathbf{Z}_{1}$ zones 2 and 1 are sampled. It is the third times that zone 1 is sampled so that the third column of $\mathbf{Z}_{Z 1}$ informs us that the first and third row of $Z 1$ are sampled in that column. This translates into black boxes for entries $(10,5)$ and $(8,5)$ of Figure 5.1.

## Section 5.2: Some matrices of sample indicators for a three level sampling problem

To simplify the presentation we use matrices $\mathbf{Z}_{0}$ where identical columns are pooled together. A matrix of sample indicators is obtained by taking a random permutation of $\mathbf{Z}_{0}$. For the first level of sampling $\mathbf{Z}_{0}$ can be expressed using row vectors of ones, $\mathbf{1}_{n}^{\top}$, and zeros, $\mathbf{0}_{n}^{\top}$, where $n$ is the length of the vector. The level $1 \mathbf{Z}_{0}$ matrix is

$$
\mathbf{Z}_{0}^{(1)}=\left(\begin{array}{ccc}
\mathbf{1}_{12}^{\top} & \mathbf{0}_{8}^{\top} & \mathbf{0}_{13}^{\top} \\
\mathbf{0}_{12}^{\top} & \mathbf{1}_{8}^{\top} & \mathbf{0}_{13}^{\top} \\
\mathbf{0}_{12}^{\top} & \mathbf{0}_{8}^{\top} & \mathbf{1}_{13}^{\top}
\end{array}\right) .
$$

The second cluster is selected 8 times. A $\mathbf{Z}_{0}$ matrix for choosing the periods, $\mathrm{AM}, \mathrm{PM}$ or EV , on the days that cluster two is selected is

$$
\mathbf{Z}_{0,2}^{(2)}=\left(\begin{array}{llllllll}
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}\right),
$$

where the exponent gives the sampling level and the index accompanying 0 is the cluster to which this matrix $\mathbf{Z}_{0}$ applies. The row totals for $\mathbf{Z}_{0,2}^{(2)}, 5,5$, and 6 are half the AM, PM, EV totals for cluster 2 in

Table 5.2. We now consider period AM for cluster 2. Considering row 1 of $\mathbf{Z}_{0,2}^{(2)}$, AM is selected 5 times. The third level matrix has five columns and 6 rows, corresponding to the time-point-site (1-4, 1-5, 1-6, 2-4, $2-5,2-6)$ with row totals $(1,2,2,2,1,2)$ as given in the AM line for cluster 2, see Table 5.2. Here one has to stratify by time point: one site need to be selected at each time point. A candidate $\mathbf{Z}_{0}$ where the first three rows are for time point 1 and the last 3 for time point 2 is

$$
\mathbf{Z}_{0,2, \mathrm{AM}}^{(3)}=\left(\begin{array}{lllll}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 \\
\hline 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1
\end{array}\right) .
$$

Thus the set of possible level 3 matrices $\mathbf{Z}_{2, A M}^{(3)}$ is obtained by permuting the columns of the first three and of the last three rows of $\mathbf{Z}_{0,2, A M}^{(3)}$ independently. The third level of sampling involves 9 matrices similar to $\mathbf{Z}_{0,2, \mathrm{AM}}^{(3)}$.

## References

Barvinok, A. (2010). On the number of matrices and a random matrix with prescribed row and column sums and 0-1 entries. Advances in Mathematics, 224(1), 316-339.

Chen, X.-H., and Dempster, A.P. (1994). Weighted finite population sampling to maximize entropy. Biometrika, 81(3), 457-469.

Deville, J.-C., and Maumy-Bertrand, M. (2006). Extension of the indirect sampling method and its application to tourism. Survey Methodology, 32, 2, 177-185. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9552-eng.pdf.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: the cube method. Biometrika, 91(4), 893912.

Grafström, A., and Matei, A. (2015). Coordination of conditional Poisson samples. Journal of Official Statistics, 31(4), 649-672.

Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. Computational Statistics \& Data Analysis, 74, 81-94.

Ida, I.O., Rivest, L.-P. and Daigle, G. (2018). Using balanced sampling in creel surveys. Survey Methodology, 44, 2, 239-252. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54954-eng.pdf.

Juillard, H., Chauvet, G. and Ruiz-Gazen, A. (2017). Estimation under cross-classified sampling with application to a childhood survey. Journal of the American Statistical Association, 112, 850-858.

Kozfkay, J.R., and Dillon, J.C. (2010). Creel survey methods to assess catch, loss, and capture frequency of white sturgeon in the Snake River, Idaho. North American Journal of Fisheries Management, 30(1), 221229.

Matei, A., and Tillé, Y. (2005). Maximal and minimal sample co-ordination. Sankhyā: The Indian Journal of Statistics, 590-612.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. and Wagner, H. (2020). vegan: Community Ecology Package.

Rivest, L.-P. (2021). Limiting properties of an equiprobable sampling scheme for 0-1 matrices. Statistics \& Probability Letters, 172, 109047.

Rivest, L.-P., and Ebouele, S.E. (2020). Sampling a two dimensional matrix. Computational Statistics \& Data Analysis, 149, 106971.

Tillé, Y. (2006). Sampling Algorithms. New York: Springer Science \& Business Media.
Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P. and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. Environmetrics, 26(8), 557-570.

# Targetted double control of burden in multiple surveys 

Alina Matei, Paul A. Smith, Marc J.E. Smeets and Jonas Klingwort ${ }^{1}$


#### Abstract

Sample coordination methods aim to increase (in positive coordination) or decrease (in negative coordination) the size of the overlap between samples. The samples considered can be from different occasions of a repeated survey and/or from different surveys covering a common population. Negative coordination is used to control the response burden in a given period, because some units do not respond to survey questionnaires if they are selected in many samples. Usually, methods for sample coordination do not take into account any measure of the response burden that a unit has already expended in responding to previous surveys. We introduce such a measure into a new method by adapting a spatially balanced sampling scheme, based on a generalization of Poisson sampling, together with a negative coordination method. The goal is to create a double control of the burden for these units: once by using a measure of burden during the sampling process and once by using a negative coordination method. We evaluate the approach using Monte-Carlo simulation and investigate its use for controlling for selection "hot-spots" in business surveys in Statistics Netherlands.


Key Words: Coordinated sampling; Negative coordination; Survey burden; Burden "hot-spots".

## 1. Introduction

Sample coordination methods seek to alter the size of the overlap(s) between two or more samples relative to the case where all the samples are selected independently. Positive coordination refers to the case where the overlap is larger than under independent sampling, and is generally used to reduce the variance of measures of change between successive periods of repeating surveys, though it can also be used to link together information from two separate surveys. Negative coordination is when the overlap is smaller than under independent sampling, and is used particularly to reduce the number of surveys in which a particular unit is selected in a given period, and therefore to control the perceived burden of responding (Bradburn, 1978). Bottone, Modugno and Neri (2021) have shown that increased perceived burden is associated with higher attrition and partial response, and also with lower data quality.

In fact the effect of negative coordination is to spread out a fixed overall burden across either or both of more units and more time, so the total burden is the same, but the risk of any particular unit having a large burden within a short period is reduced - ideally to zero, though this is not always possible in practice, because detailed stratification can result in some parts of the population being relatively heavily sampled. Negative coordination has been widely applied in business surveys where sampling fractions tend to be large, and we focus on examples from surveys of businesses and institutions, though the approach can be applied in various situations.

There are many available methods for sample coordination, and Matei and Smith (2023) give an overview of them. An important approach in sample coordination is the use of permanent random numbers (PRNs). PRNs were introduced by Brewer, Early and Joyce (1972) to coordinate Poisson samples, and have been widely used since as the basis of other, different methods (see Ohlsson, 1995, for an overview). They

[^3]have also formed the basis of a number of sample coordination systems (which are also reviewed by Matei and Smith, 2023), which generally have to deal with a series of practical issues as well as following the theory of a particular sample coordination method. Coordination systems operate most flexibly where there are many units, and so benefit small units particularly. But they also operate to spread the response burden evenly for units of all sizes.

The consequence is that although a sample coordination system reduces the overlaps between samples and therefore the current burden, there are still some units which appear more frequently than others across the range of samples being coordinated. This is at least partly driven by the use of stratified designs with many small strata, some of which have large inclusion probabilities. Statistics Netherlands operates a sample coordination system called the Survey Burden System (Smeets and Boonstra, 2018) which uses a PRN coordination approach which takes account of the accumulated burden of the units. This system does indeed result in relatively only a few units being included in multiple samples, but these "hot-spots" are a challenge because of the burden they represent for particular units and the consequent effects on response and relations with respondents.

A related situation is identified by Landry (2011) in Statistics Canada's Survey of Employment, Payroll and Hours (SEPH). Landry suggests the use of a take-none (cut-off) stratum, mainly to control the response burden on the smallest businesses. Although cut-off sampling is widely used, there is a risk of bias in estimates from such samples, so a better approach to burden control allowing unbiased estimation would be preferable.

In this paper we therefore seek a method which follows the requirements of negative sample coordination, but allows for some additional control for units which have particular characteristics. We adapt the approach of spatially correlated Poisson (SCP) sampling (Grafström, 2012) to this problem, by introducing a measure of the response burden in the sampling process.

In the remainder of the paper we introduce the framework and notation for sample coordination in Section 2, outline the procedure for spatially correlated Poisson sampling in Section 2.4 and develop the methodology for sample coordination with targetted double control in Section 3. We evaluate our proposal in Section 3.2; we provide Monte-Carlo simulation studies using the MU284 population from Särndal, Swensson and Wretman (1992) in Section 3.3, and real data on the business population in the Netherlands to assess its ability to deal with hot-spots in Section 3.4. Section 4 concludes with a discussion.

## 2. Sample coordination and SCP sampling

### 2.1 Framework and notation

We consider the framework of two overlapping finite populations of units, denoted by $U_{1}$ and $U_{2}$. One selects samples $s_{1}$ from $U_{1}$ and $s_{2}$ from $U_{2}$, using the sampling designs $p_{1}$ and $p_{2}$, respectively. The set formed by the two samples can be seen as a bivariate sample $s=\left(s_{1}, s_{2}\right) \subseteq U_{1} \times U_{2}$, having a joint sampling design $p$, with the marginals $p_{1}$ and $p_{2}$. The samples $s_{1}$ and $s_{2}$ are drawn dependently to alter the size of the overlap(s) between them relative to the case where the samples are selected independently. Thus, the samples $s_{1}$ and $s_{2}$ are said to be coordinated if

$$
p\left(s_{1}, s_{2}\right) \neq p_{1}\left(s_{1}\right) p_{2}\left(s_{2}\right)
$$

(see Cotton and Hesse, 1992; Mach, Reiss and Şchiopu-Kratina, 2006).
The size of the overlap between $s_{1}$ and $s_{2}$, denoted by $c$, represents the number of units common to $s_{1}$ and $s_{2}$. It is in general a random variable having expectation

$$
E(c)=\sum_{k \in U} \pi_{k, 12}
$$

with

$$
\pi_{k, 12}=P\left(k \in s_{1}, k \in s_{2}\right)=\sum_{s_{1}, k \in s_{1}} \sum_{s_{2}, k \in s_{2}} p\left(s_{1}, s_{2}\right), \forall k \in U
$$

and $U=U_{1} \cup U_{2}$ is the so-called "overall population". Maximizing/minimizing $E(c)$ in positive/negative sample coordination represents an overall standard to evaluate a coordination method.

Let $\pi_{k 1}=P\left(k \in s_{1}\right)$ and $\pi_{k 2}=P\left(k \in s_{2}\right)$ be the first-order inclusion probabilities of unit $k \in U$ in the first and second sample, respectively. We consider that $\pi_{k 1}=0$ if $k \notin U_{1}$ (if $s_{1}$ and $s_{2}$ are samples for two periods of the same survey, these new units represent "births") and $\pi_{k 2}=0$ if $k \notin U_{2}$ ("deaths").

Based on probability theory, the following bounds are available for the joint probability $\pi_{k, 12}$, for any $k \in U$

$$
\begin{equation*}
\operatorname{ALB}_{k}=\max \left(0, \pi_{k 1}+\pi_{k 2}-1\right) \leq \pi_{k, 12} \leq \min \left(\pi_{k 1}, \pi_{k 2}\right)=\mathrm{AUB}_{k} \tag{2.1}
\end{equation*}
$$

where ALB stands for "absolute lower bound" and AUB for "absolute upper bound". One obtains the lower and upper bounds for $E(c)$ by applying the sum over all units $k \in U$ on the left and right side of Expression (2.1) (see Matei and Tillé, 2005):

$$
\begin{equation*}
\mathrm{ALB}=\sum_{k \in U} \mathrm{ALB}_{k} \leq E(c) \leq \sum_{k \in U} \mathrm{AUB}_{k}=\mathrm{AUB} . \tag{2.2}
\end{equation*}
$$

If $s_{1}$ is drawn before $s_{2}$, and $s_{2}$ is selected conditionally on $s_{1}$ using a probability $P\left(s_{2} \mid s_{1}\right)$, one can obtain any value of $\pi_{k, 12} \in\left[\mathrm{ALB}_{k}, \mathrm{AUB}_{k}\right]$ using the conditional probabilities (Cotton and Hesse, 1992)

$$
\begin{aligned}
& P\left(k \in S_{2} \mid k \in S_{1}\right)=\pi_{k, 12} / \pi_{k 1} \\
& P\left(k \in S_{2} \mid k \notin S_{1}\right)=\left(\pi_{k 2}-\pi_{k, 12}\right) /\left(1-\pi_{k 1}\right)
\end{aligned}
$$

We focus below on the negative coordination of two samples. Reaching the lower bound in Expression (2.1) for all units $k \in U$ is equivalent to creating the "best" possible degree of negative coordination between samples based on their overlap minimization.

### 2.2 Sample coordination with PRNs

As underlined in Section 1, sample coordination methods with PRNs are commonly used in practice. They are based on the following basic idea: one associates a uniform random number drawn independently
from the Unif $(0,1)$ distribution with each unit $k \in U$. These numbers are called "permanent" since they are used in the selection process over time and over surveys for units which persist in the population. For a "birth" (a new unit which appears in the population), a new PRN is assigned; for a "death" (a unit which disappears from the population), the unit and its associated PRN are deleted from the corresponding survey frame.

Introduced by Brewer et al. (1972), Poisson sampling with PRNs is widely used in sample coordination, especially as a base for coordination systems (see Qualité, 2019, for an example applied in Switzerland). In negative coordination it is implemented as follows: first, one generates the PRNs $u_{1}, u_{2}, \ldots, u_{N}$ independently from the $\operatorname{Unif}(0,1)$ distribution. Next, if $u_{k}<\pi_{k 1}, k \in U$, unit $k$ is included in $s_{1}$. The sample $s_{2}$ is selected in a similar manner, but using the numbers $1-u_{k}$ instead of $u_{k}$ : if $1-u_{k}<\pi_{k 2}$ then unit $k$ is included in $s_{2}$. Using Poisson sampling with PRNs to negatively coordinate $s_{1}$ and $s_{2}$ allows the bound $\mathrm{ALB}_{k}$ in Expression (2.1) to be reached for any unit $k \in U$. This can be shown as follows: unit $k$ is selected in both $s_{1}$ and $s_{2}$ if $u_{k}<\pi_{k 1}$ and $1-u_{k}<\pi_{2 k}$. This is equivalent to $1-\pi_{k 2}<u_{k}<\pi_{k 1}$. The probability that this occurs is $\max \left(0, \pi_{k 1}-\left(1-\pi_{k 2}\right)\right)=\max \left(0, \pi_{k 1}+\pi_{k 2}-1\right)=\operatorname{ALB}_{k}$.

While Poisson sampling with PRNs is a very attractive scheme for sample coordination, it has an important drawback: the resulting samples have random sizes, increasing the variance of the estimates. For this reason, fixed-size sampling designs are sometimes preferred; see, for instance, Pareto sampling with PRNs (Rosén, 1997a, b). Nevertheless, the bounds provided in Expression (2.1) are in general not reached using sampling designs with fixed sample size and unequal inclusion probabilities; for some empirical results based on Monte-Carlo simulation, see Grafström and Matei (2018). Compared to Poisson sampling with PRNs which allows independent selection of units, such sampling designs impose more restriction on the unit selection mainly due to the fixed sample size, and fail in general to reach the bounds provided in Expression (2.1). Theoretical conditions to reach the two overall bounds given in Expression (2.2) are provided by Matei and Tillé (2005).

### 2.3 Response burden and sample coordination

Response burden is a difficult concept to define; it may include objective factors such as the time spent to provide questionnaire responses and subjective factors such as what is perceived as burden by the respondents, see for instance Natkowska and Modak (2014); Bottone et al. (2021). From the statistical perspective, we use the following definition provided by Sunter (1977).

Consider several surveys $j=1,2, \ldots, M$ having associated populations of units $U_{1}, U_{2}, \ldots, U_{M}$, with $U=\cup_{j=1}^{M} U_{j}$. The response burden of unit $k$ in $M$ surveys is a random variable

$$
\mathrm{RB}_{k}=\sum_{j=1}^{M} \beta_{j} \times I_{k j},
$$

with $I_{k j}=1$ if $k \in s_{j}$ and 0 otherwise, and $\beta_{j}$ is the response load imposed by the $j^{\text {th }}$ survey for all units selected to participate in this survey. The expected value of $\mathrm{RB}_{k}$ is given by $E\left(\mathrm{RB}_{k}\right)=\sum_{j=1}^{M} \beta_{j} \pi_{k j}, k \in U$, where $\pi_{k j}=P\left(k \in s_{j}\right)$, and $s_{j}$ is the $j^{\text {th }}$ survey sample, $j=1, \ldots, M$. When $\beta_{j}=1$, for all $j=1, \ldots, M$, we simply obtain that

$$
\begin{equation*}
\mathrm{RB}_{k}=\sum_{j=1}^{M} I_{k j}, \text { and } E\left(\mathrm{RB}_{k}\right)=\sum_{j=1}^{M} \pi_{k j}, k \in U . \tag{2.3}
\end{equation*}
$$

The response burden is commonly associated with a negative coordination of samples. Usually, a coordination method does not use any measure of response burden in the sampling process. If a negative coordination method is applied, the value of the response burden (seen as the realization of a random variable) diminishes for the units not included in the overlap: for instance, for two selected samples $s_{1}, s_{2}$, and considering the first part of Expression (2.3), the value of $\mathrm{RB}_{k}$ is 1 if $k \in s_{1} \backslash s_{2}$ or $k \in s_{2} \backslash s_{1}$ while it is 2 if $k \in s_{1} \cap s_{2}$. Minimizing the sample overlap size implies having fewer units with the value of $\mathrm{RB}_{k}=2$. However, at the overall level, sample coordination methods do not affect $E\left(\mathrm{RB}_{k}\right)$, but try to control for excessive burdens and to allocate burdens in a fair way, as well as to reduce the variance of $\mathrm{RB}_{k}$.

In what follows, we use the expression cumulated response burden to denote the realized value of the $\mathrm{RB}_{k}$ given in the first part of Expression (2.3).

### 2.4 Spatially correlated Poisson sampling

Spatially correlated Poisson (SCP) sampling is a particular case of correlated Poisson sampling, a family of sampling designs introduced by Bondesson and Thorburn (2008). First, we review this method for a generic sample with given first-order inclusion probabilities; next, we give the modification provided by Grafström (2012) which produces SCP samples.

Correlated Poisson sampling is a list sequential method used to draw a random sample $\tilde{s}$ from $U$, with prescribed inclusion probabilities $\pi_{k}, k \in U$. Consider the selection probability $\pi_{k}^{(k-1)}$, that allows unit $k$ to be selected in $\tilde{s}$ at the $k^{\text {th }}$ iteration of the following algorithm (see Step 3b):

Step 1: $\quad$ set $\pi_{\ell}^{(0)}=\pi_{\ell}$, for all $\ell \in U$,
Step 2: $\quad$ set $k=1$,
Step 3: while $k \leq N$ do
3a: generate $\tilde{u}_{k}$ independently from the $\operatorname{Unif}(0,1)$ distribution,
3b: if $\tilde{u}_{k}<\pi_{k}^{(k-1)}$ then $I_{k}=1$ ( $k$ is selected in $\tilde{s}$ ) else $I_{k}=0(k$ is not selected in $\tilde{s})$, where $I_{k}$ is the indicator variable associated to unit $k \in U$,

3c: update the selection probabilities for the remaining units $i=k+1, \ldots, N$ according to

$$
\pi_{i}^{(k)}=\pi_{i}^{(k-1)}-\left(I_{k}-\pi_{k}^{(k-1)}\right) w_{k}^{(i)},
$$

where $w_{k}^{(i)}$ are some weights given by unit $k$ to other units $i=k+1, \ldots, N$,
3d: increment $k$.
The weights can be chosen freely in the following range

$$
\begin{equation*}
-\min \left(\frac{1-\pi_{i}^{(k-1)}}{1-\pi_{k}^{(k-1)}}, \frac{\pi_{i}^{(k-1)}}{\pi_{k}^{(k-1)}}\right) \leq w_{k}^{(i)} \leq \min \left(\frac{\pi_{i}^{(k-1)}}{1-\pi_{k}^{(k-1)}}, \frac{1-\pi_{i}^{(k-1)}}{\pi_{k}^{(k-1)}}\right), \tag{2.4}
\end{equation*}
$$

in order to assure that $0 \leq \pi_{i}^{(k-1)} \leq 1, i=k+1, \ldots, N$. The weights $w_{k}^{(i)}$ may depend on $I_{1}, I_{2}, \ldots, I_{k-1}$ but not on $I_{k}, I_{k+1}, \ldots, I_{N}$. The choice of $w_{k}^{(i)}$ provides different sampling designs; for instance, Poisson sampling is obtained if all $w_{k}^{(i)}$ are zero, $k \in U$. Bondesson and Thorburn (2008) showed that a fixed size sampling is obtained if for each $k \in U: \sum_{i=k+1}^{N} w_{k}^{(i)}=1$ and $\sum_{k \in U} \pi_{k}=n, n$ being the sample size. The prescribed inclusion probabilities are respected since $P(k \in \tilde{s})=\pi_{k}, k \in U$, regardless of the choice of $w_{k}^{(i)}$ (see Remark 1 in Bondesson and Thorburn, 2008). This is due to the fact that

$$
E\left(\pi_{k}^{(k-1)}\right)=E\left(E\left(\pi_{k}^{(k-1)} \mid \pi_{k}^{(k-2)}\right)\right)=E\left(\pi_{k}^{(k-2)}\right)=\ldots=\pi_{k}, \text { for all } k=1,2, \ldots, N,
$$

as underlined by Grafström (2012).
Grafström (2012) applied this method in spatial sampling, where the units in the population have associated geographical coordinates which can be used to compute distances between them. His goal was to draw a balanced sample, so that the selected units are spread over the space under study. To avoid clustering of similar units and to obtain well-spread samples, Grafström (2012) used positive weights $w_{k}^{(i)}$ in Expression (2.4) chosen such that unit $k$ gives maximal weight to the unit closest to $k$ in (Euclidean) distance, among the units $i=k+1, \ldots, N$, then as much weight as possible to the second closest unit, etc. with the restriction that $\sum_{i=k+1}^{N} w_{k}^{(i)}=1$ is fulfilled, for any $k \in U$, and respecting the upper bound for each weight $w_{k}^{(i)}$ in Expression (2.4). This method, called the maximal weight strategy, provides spatially correlated Poisson sampling, which is a balanced spatial sampling design of fixed sample size, assuming that $\sum_{k \in U} \pi_{k}=n$. Moreover, if $\pi_{k}, k \in U$ are proportional to a size measure, SCP sampling becomes a $\pi \mathrm{ps}$ sampling design of fixed sample size.

Grafström and Matei (2018) employed SCP sampling with PRNs to coordinate samples in a manner similar to Poisson sampling with PRNs (Brewer et al., 1972). Thus, for negative coordination, Step 3a of the previous algorithm is executed only once for $s_{1}$, in order to associate a PRN $u_{k}$ with each unit $k \in U$, and $\tilde{u}_{k}$ is replaced by $u_{k}$. Next, to select $s_{1}$ one uses $u_{k}$ instead of $\tilde{u}_{k}$ and the corresponding selection probabilities in Step 3a; to select $s_{2}$, one uses $1-u_{k}$ instead of $\tilde{u}_{k}$ and the corresponding selection probabilities in Step 3a. SCP sampling with PRNs is implemented in the function "scps_coord" of the R package "BalancedSampling" (Grafström, Lisic and Prentius, 2022).

## 3. Targetted double control strategy

### 3.1 Description of the strategy

Due to the cumulated response burden, some units do not answer the survey questionnaires if they are selected in many samples (Lorenc, Kloek, Abrahamsson and Eckman, 2013). We assume that, over time, some such units with a large cumulated burden become "notorious" non-respondents. The same occurs with "hot-spots" in business surveys in Statistics Netherlands (see Section 3.4). We denote these units as nondesired units, and we want to exclude them as much as possible from future selections, while respecting their prescribed inclusion probabilities. It is possible to classify the units of $U$ into two categories: desired (usually with a low cumulated response burden) and non-desired units (usually with a large cumulated response burden).

Our goal is to produce a double control of the response burden for the non-desired units: first by using a measure of the response burden in the sampling process, and second by using a method for negative sample coordination.

To create a targetted double control of non-desired units, we modify spatially correlated Poisson sampling (Grafström, 2012), and use a negatively coordinated sampling scheme. We describe below how to adapt SCP sampling for the coordination framework with targetted double control.

As explained in Section 2.4, in spatial sampling, the units in $U$ have associated geographical coordinates. It is thus possible to compute distances between units, usually using Euclidean distance. Units close in distance provide, in general, similar information. Spatially balanced sampling allows the selection of units that are spread over the space, and thus avoids collecting similar information.

We propose to replace the matrix of geographical coordinates in SCP sampling by the vector formed by a measure based on the response burden of each unit. Other information, such as the inclusion probabilities, can also be included, and the vector becomes a matrix. This could lead to an extra spread with respect to the inclusion probabilities. Similar units usually have inclusion probabilities close to each other and non-desired units often have larger inclusion probabilities.

The Euclidean distances between units are next computed using this new vector or matrix. If a nondesired unit is selected in the current sample, SCP sampling avoids selecting a similar unit. Next, we use a negative coordination method for samples. We call this method the targetted double control strategy, while a SCP sample used in this strategy is called an adapted SCP sample (ASCP sample). For adapted SCP sampling, a measure of the response burden for a unit can simply be defined as 1 if the unit is a non-desired unit and 0 , otherwise; this represents a proxy for large and small cumulated response burdens respectively, and defines a measure of the unit status. Other measures of response burden can be used. For instance, we employ the cumulated response burden in Section 3.4.

These two options (unit status and cumulated response burden respectively) are used in the algorithm given below which describes the adapted SCP sampling for two negatively coordinated samples $s_{1}$ and $s_{2}$ :

Step a: based on previous information (previous surveys), create the vector or matrix which includes the information about the unit status (desired or non-desired) or the cumulated response burdens of the units in $U$;

Step b: select $s_{1}$ using the corresponding inclusion probabilities, by applying SCP sampling with the maximum weight strategy. The Euclidean distances between units are computed using the vector or matrix created in Step a. If the cumulated response burdens of the units are used in the vector or matrix, update them, as well as the vector or matrix, after the selection of $s_{1}$, according to the definition given in Expression (2.3);

Step c: select $s_{2}$, negatively coordinated with $s_{1}$, using the corresponding inclusion probabilities, by applying SCP sampling with the maximum weight strategy. The Euclidean distances between units are computed using the same vector or matrix as for selection of $s_{1}$ if the unit status is used, or the updated ones if the cumulated response burdens are used.

If $s_{1}$ and $s_{2}$ are negatively coordinated using PRNs, $s_{1}$ is drawn using $u_{k}, k \in U$, while $s_{2}$ using $1-u_{k}, k \in U$ ( $u_{k}$ replaces $\tilde{u}_{k}$ in Step 3b of the algorithm given in Section 2.4).

Remark 1 As indicated in Section 2.3, the response burden is a random variable. The Euclidean distances in the previous algorithm are computed conditionally on the realized value of $\mathrm{RB}_{k}, k \in U$.

### 3.2 Effectiveness of the targetted double control strategy

We provide in the next two sections the results of a Monte-Carlo simulations to show the effectiveness of the proposed strategy, and use two methods to test its performance:

- Method 1: two samples are negatively coordinated using PRNs;
- Method 2: two samples are negatively coordinated but without using PRNs. One new sample $s_{2}$ is drawn and negatively coordinated with an existing sample $s_{1}$ ( $s_{1}$ is fixed and the inclusion probabilities are known). It is possible that $s_{1}$ was selected using PRNs, but these numbers are not available for the second selection. Thus, conditional on $s_{1}$, a new random number $u_{k 2}$ is associated with unit $k \in U$ : if $k \in s_{1}$, one generates $u_{k 2}$ independently from the $\operatorname{Unif}\left(0, \pi_{k 1}\right)$ distribution; otherwise one generates $u_{k 2}$ independently from the $\operatorname{Unif}\left(\pi_{k 1}, 1\right)$ distribution. Next, $k$ is selected in $s_{2}$ using the number $1-u_{k 2}$, and the probability $\pi_{k 2}, k \in U$.

In practice, Method 1 is applied in the general case when it is possible to draw new samples for both survey 1 and 2 . Method 2 is applied if it is not possible to draw a new sample $s_{1}$ for practical reasons. For example, if the businesses have already received a questionnaire for survey 1 , or when a new survey is added to a coordination system and the PRNs of the businesses cannot be extracted from the system.

Five measures provided below are used to quantify the performance of the proposed strategy. Measures 1 and 2 focus on the selection of non-desired units. Measure 1 quantifies the number of pairs ( $s_{1}, s_{2}$ ) with given numbers of non-desired units in common between $s_{1}$ and $s_{2}$ (that is, the number of non-desired units in the overlap), and it is the most important for the study of the proposed strategy application. We expect that the proposed strategy will reduce the variance of the number of non-desired units in the overlap compared to its competitors and independent sampling. Measure 2 is related to $\mathrm{ALB}_{k}$, with $k$ being a nondesired unit. Ideally, one wants to reach the lower bound $\mathrm{ALB}_{k}$ given in Expression (2.1) for $P\left(k \in s_{1}\right.$, $\left.k \in s_{2}\right)=\pi_{k, 12}$, for any $k \in U$. As underlined in Section 2.2, we are able to reach it when both $s_{1}$ and $s_{2}$ are selected using Poisson sampling with PRNs (so when Method 1 is applied). Method 2 does not always allow it to be reached. We hope that the proposed strategy provides values of the estimated $P\left(k \in s_{1}, k \in s_{2}\right)$ of the non-desired units close to this bound, showing that such units have small chances to be selected in the two samples. In some cases, however, a direct comparison with $\mathrm{ALB}_{k}$ is not possible (see for example Method 2, framework 2 in Section 3.3), but we expect that the proposed strategy provides values of the estimated $P\left(k \in s_{1}, k \in s_{2}\right)$ lower than its competitors, or at least similar.

Measures 3, 4 and 5 concern the overall performance of a negative coordination method. A value of Measure 3 close to ALB indicates an important degree of negative coordination of the two samples.

Measures 4 and 5 are measures of the overlap (relative) variance. As before, we expect that our strategy is able to provide lower values than its competitors for Measures 3, 4, and 5, or at least similar.

The five measures are:

- Measure 1: number of pairs of samples ( $s_{1}, s_{2}$ ), with $s_{1}$ and $s_{2}$ containing in common a number of given non-desired units through simulations;
- Measure 2: values of the estimated $P\left(k \in s_{1}, k \in s_{2}\right)$ of the non-desired units through simulations;
- Measure 3: the Monte-Carlo expected overlap

$$
E_{\mathrm{sim}}(c)=\frac{1}{m} \sum_{\ell=1}^{m} c_{\ell}^{1,2},
$$

where $c_{\ell}^{1,2}=\left|s_{1 \ell} \cap s_{2 \ell}\right|$, and $s_{1 \ell}, s_{2 \ell}$, are the samples drawn in the $\ell^{\text {th }}$ run, and $\left|s_{1 \ell} \cap s_{2 \ell}\right|$ represents the number of common units of $s_{1 \ell}$ and $s_{2 \ell}, m$ is the number of runs; for Method 2, when $s_{1}$ is fixed (as in some simulations below), $c_{\ell}^{1,2}=\left|s_{1} \cap s_{2 \ell}\right|$, where $s_{2 \ell}$, is the sample drawn in the $\ell^{\text {th }}$ run, where $\left|s_{1} \cap s_{2 \ell}\right|$ represents the number of common units of $s_{1}$ and $s_{2 \ell}$;

- Measure 4: the Monte-Carlo variance of the overlap

$$
\operatorname{Var}_{\mathrm{sim}}(c)=\frac{1}{m-1} \sum_{\ell=1}^{m}\left(c_{\ell}^{1,2}-E_{\mathrm{sim}}(c)\right)^{2} ;
$$

- Measure 5: the Monte-Carlo coefficient of variation of the overlap

$$
\mathrm{CV}_{\mathrm{sim}}(c)=\frac{\sqrt{V_{\mathrm{sim}}(c)}}{E_{\mathrm{sim}}(c)} .
$$

### 3.3 Simulation with MU284 population

We consider as $U$ region 2 of the well-known MU284 data (see Appendix B in Särndal et al., 1992). Samples $s_{1}$ and $s_{2}$ with expected sizes $n_{1}=10$ and $n_{2}=6$ are selected respectively from this region which contains in total $N=48$ units. No births or deaths are used. The inclusion probabilities $\pi_{k 1}$ and $\pi_{k 2}, k \in U$ are respectively proportional to variables P75 and P85, the population size of Swedish municipalities in 1975 and 1985. We call this setting the "MU284 population". Six units from $U$ (with labels 4, 12, 21, 22, 32 and 44) are declared non-desired units for both $s_{1}$ and $s_{2}$. As a measure of the response burden of a unit $k$ we use a binary variable to compute the Euclidean distances in adaptive SCPS: 1 , if unit $k$ is a nondesired unit, and 0 otherwise (the unit status).

We provide below the results of the Monte-Carlo simulations using Methods 1 and 2 and different sampling schemes. In the simulations 100,000 runs are used.

In Method 1 , both $s_{1}$ and $s_{2}$ are random in each run and have the same type (the same sampling scheme is used), but different inclusion probabilities. They are negatively coordinated with PRNs in the first 5 cases below, while the $6^{\text {th }}$ case refers to independent sample selections. In each run of the Monte-Carlo simulation, we draw:

1. two Poisson samples;
2. two Pareto samples;
3. two adapted SCP samples; for $s_{1}, s_{2}$, only $\pi_{1}$ and respectively $\pi_{2}$ are used to compute the Euclidean distances between the units (this case is indicated ASCP $\quad \pi$ in the following tables; ASCP stands for adapted SCP);
4. two adapted SCP samples; for $s_{1}, \pi_{1}$ and the values of the unit status are used to compute the Euclidean distances; for $s_{2}, \pi_{2}$ and the values of the unit status are used to compute the Euclidean distances (this case is indicated by ASCP ${ }_{-} \pi$ _inf in the following tables);
5. two adapted SCP samples; for $s_{1}, s_{2}$, the values of the unit status are used to compute the Euclidean distances (this case is indicated by ASCP_inf in the following tables);
6. two independent adapted SCP samples (without negative coordination); for $s_{1}, s_{2}$, the values of the unit status are used to compute the Euclidean distances (this case is indicated by "ASCP_inf without coordination" in the following tables).

Pareto sampling is introduced in the Monte-Carlo simulation because it provides fixed sample sizes, as well as the adaptive SCPS. The inclusion probabilities are used in the adaptive SCPS in cases 3 and 4 above to compare with case 5 , which only uses the binary information to compute the Euclidean distances.

Method 2 is applied in two different frameworks in order to make the connection with the application given in Section 3.4 which uses real data:

1. framework 1: both $s_{1}$ and $s_{2}$ are random in each run; $s_{1}$ is a Poisson sample, while $s_{2}$ is a sample of the type enumerated for Method 1 (Poisson, Pareto, etc.);
2. framework 2: the first sample $s_{1}$ is fixed $\left(s_{1}=\{4,6,12,18,22,35,44\}\right)$ and is a Poisson sample, and only $s_{2}$ is random in each run; $s_{2}$ is a sample of the type enumerated for Method 1 (Poisson, Pareto, etc.).

Note that Measure $2\left(\hat{P}\left(k \in s_{1}, k \in s_{2}\right)\right)$ can be compared to $\mathrm{ALB}_{k}, k \in U$ for Method 1 and Method 2 (but only in framework 1). In Method 2, framework 2, $s_{1}$ is fixed and we are not able to reconstruct by simulation $\quad \sum_{s_{1}, k \in s_{1}} \sum_{s_{2}, k \in s_{2}} p\left(s_{1}, s_{2}\right)=P\left(k \in s_{1}, k \in s_{2}\right)$. Thus, we only estimate $\sum_{s_{2}, k s_{2}} p\left(s_{1}, s_{2}\right)$. A conditional $\mathrm{ALB}_{k}$ cannot be used in this case. The corresponding tables below do not include values of $\mathrm{ALB}_{k}$.

The tables in Section 3.3.1 present the results for Method 1, while for Method 2, they are given in Section 3.3.2 (framework 1) and Section 3.3.3 (framework 2). Using Measure 1 (see Tables 3.1, 3.4 and 3.7), the possible number of non-desired units common to $s_{1}$ and $s_{2}$ is between 0 and 6 for Method 1 and Method 2, framework 1, and between 0 and 4 for Method 2, framework 2. Both Poisson and Pareto sampling reach the maxima of the ranges. In contrast, adaptive SCPS (even without PRNs) shrinks the distribution of the possible number of non-desired units in common, and avoids the selection of a large number of nondesired units, resulting in a decreased variance of their number, compared with its competitors.

The use of the binary variable (without any supplementary information, such as the inclusion probabilities) in the adaptive SCPS seems to be the best choice to compute the Euclidean distances between
units (the case ASCP_inf). For this setting, in general, ASCP_inf performs the best, and selects mostly 1 (see Table 3.7) or 2 non-desired units in common (see Tables 3.1 and 3.4), that is, fewer than the other two methods. However, no pairs $\left(s_{1}, s_{2}\right)$ provide 0 non-desired units in common in Method 1 or Method 2, framework 2 (see Tables 3.1 and 3.7), as shown by Poisson and Pareto sampling.

For Measure 2, in Method 1 and Method 2 (framework 1), Poisson sampling is able to reach the ALB $_{k}$ for any non-desired unit $k$, as expected (Tables 3.2 and 3.5). ASCP_inf and Pareto sampling perform similarly in the case of these two methods, and provide values of $\hat{P}\left(k \in s_{1}, k \in s_{2}\right)$ equal to $\mathrm{ALB}_{k}$ or slightly larger. In Method 2, framework 2, $\hat{P}\left(k \in s_{1}, k \in s_{2}\right)$ cannot be compared to ALB $_{k}$, because of the way the samples are simulated. For this method, ASCP_inf presents values of Measure 2 in agreement with Poisson and Pareto sampling, and no sampling method is the best.

In Tables 3.3 and 3.6, Pareto sampling shows lower values for the expected overlap (Measure 3; excepting Poisson sampling which reaches ALB as expected), indicating a very good overall degree of negative sample coordination. However, it displays a large estimated variance of the overlap (Measure 4), comparable to that of Poisson sampling. Compared to Pareto sampling, ASCP_inf shows a larger value for Measure 3 in Method 1 and Method 2, framework 1, but substantially reduces the values of Measures 4 and 5, indicating a better precision in estimating the overlap between $s_{1}$ and $s_{2}$. In Method 2, framework 2, ASCP_inf again performs the best for all Measures 3, 4 and 5 (see Table 3.9).

### 3.3.1 Method 1: both $s_{1}$ and $s_{2}$ are random in each run; $s_{1}, s_{2}$ are samples of the same type, but with different inclusion probabilities

Table 3.1
MU284 population, Method 1, Measure 1: number of pairs ( $s_{1}, s_{2}$ ) by possible number of non-desired units in common over 100,000 runs (so the row sums are equal to $\mathbf{1 0 0 , 0 0 0 ) .}$

|  | Possible number of non-desired units in common |  |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Design | $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ |
| Poisson | 7,469 | 27,170 | 36,007 | 21,879 | 6,536 | 919 | 20 |
| Pareto | 5,225 | 25,946 | 38,965 | 23,534 | 5,784 | 541 | 5 |
| ASCP_ $\pi$ | 49 | 8,646 | 69,411 | 21,145 | 749 | 0 | 0 |
| ASCP- $\pi$ inf | 0 | 3,614 | 93,807 | 2,575 | 4 | 0 | 0 |
| ASCP_inf | 0 | 3,443 | 93,406 | 3,140 | 11 | 0 | 0 |
| ASCP_inf without coordination | 0 | 831 | 45,796 | 53,127 | 246 | 0 | 0 |

Table 3.2
MU284 population, Method 1, Measure 2: $\hat{P}\left(k \in s_{1}, k \in s_{2}\right)$, with $k$ being a non-desired unit, 100,000 runs.

|  |  |  | $\hat{\boldsymbol{P}}\left(\boldsymbol{k} \in \boldsymbol{s}_{\mathbf{1}}, \boldsymbol{k} \in \boldsymbol{s} \mathbf{2}\right)$ |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Non-desired unit $\boldsymbol{k}$ | $\mathbf{4}$ | $\mathbf{1 2}$ | $\mathbf{2 1}$ | $\mathbf{2 2}$ | $\mathbf{3 2}$ |
| Poisson | 0.63 | 0.03 | 0.26 | 0.35 | 0.34 |
| Pareto | 0.63 | 0.07 | 0.27 | 0.35 | 0.34 |
| ASCP_ $\pi$ | 0.63 | 0.12 | 0.30 | 0.35 | 0.36 |
| ASCP_ $\pi$ inf | 0.63 | 0.03 | 0.27 | 0.35 | 0.35 |
| ASCP_inf | 0.63 | 0.06 | 0.26 | 0.35 | 0.34 |
| ASCP_inf without coordination $^{\text {ALB }_{k}}$ | 0.64 | 0.25 | 0.38 | 0.35 | 0.35 |

Table 3.3
MU284 population, $\operatorname{ALB}=1.96$, Method 1, Measures 3, 4, 5: $E_{\text {sim }}(c), \operatorname{Var}_{\text {sim }}(c), \mathrm{CV}_{\text {sim }}(c), 100,000$ runs.

| Design | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0} \times \mathbf{C V}_{\text {sim }}(\boldsymbol{c})$ |
| :--- | :---: | :---: | :---: |
| Poisson | 1.96 | 1.13 | 54.44 |
| Pareto | 2.00 | 0.99 | 49.53 |
| ASCP_ $\pi$ | 2.18 | 0.34 | 26.53 |
| ASCP_ $\pi$ inf | 2.03 | 0.09 | 15.08 |
| ASCP_inf | 2.01 | 0.08 | 13.70 |
| ASCP_inf without coordination | 3.06 | 0.68 | 26.91 |

### 3.3.2 Method 2, framework 1: both $s_{1}$ and $s_{2}$ are random in each run; $s_{1}$ is a Poisson sample

Table 3.4
MU284 population, Method 2, framework 1, Measure 1: number of pairs ( $s_{1}, s_{2}$ ) by possible number of nondesired units in common, both $s_{1}$ and $s_{2}$ are random in each run over 100,000 runs (so the row sums are equal to $\mathbf{1 0 0 , 0 0 0}$ ).

|  | Possible number of non-desired units in common |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| Design | $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ |
| Poisson | 114,060 | 51,724 | 125,551 | 157,299 | 108,236 | 37,832 |
| Pareto | 5,995 | 26,389 | 37,893 | 23,216 | 5,931 | 569 |
| ASCP_ $\pi$ | 1,053 | 16,383 | 45,920 | 34,313 | 2,331 | 0 |
| ASCP_ $\pi$ inf | 1,025 | 16,950 | 48,830 | 33,175 | 20 | 0 |
| ASCP_inf | 1,663 | 20,987 | 44,084 | 33,266 | 0 | 0 |
| ASCP_inf without coordination | 315 | 6,178 | 34,797 | 58,418 | 292 | 0 |

Table 3.5
MU284 population, Method 2, framework 1, Measure 2: both $s_{1}$ and $s_{2}$ are random in each run; 100,000 runs.

|  | $\hat{\boldsymbol{P}}\left(\boldsymbol{k} \in \boldsymbol{s}_{\mathbf{1}}, \boldsymbol{k} \in \boldsymbol{s}_{\mathbf{2}}\right)$ |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Non-desired unit $\boldsymbol{k}$ | $\mathbf{4}$ | $\mathbf{1 2}$ | $\mathbf{2 1}$ | $\mathbf{2 2}$ | $\mathbf{3 2}$ | $\mathbf{4 4}$ |
| Poisson | 0.63 | 0.03 | 0.26 | 0.35 | 0.34 |  |
| Pareto | 0.63 | 0.06 | 0.26 | 0.35 | 0.34 |  |
| ASCP_ $\pi$ | 0.63 | 0.13 | 0.34 | 0.35 | 0.36 |  |
| ASCP_ $\pi$ inf | 0.63 | 0.03 | 0.35 | 0.34 | 0.37 |  |
| ASCP_inf | 0.63 | 0.06 | 0.26 | 0.35 | 0.37 | 0.41 |
| ASCP_inf without coordination $_{\text {ALB }_{k}}$ | 0.64 | 0.25 | 0.38 | 0.43 | 0.42 | 0.42 |

Table 3.6
MU284 population, $A L B=1.96$, Method 2, framework 1, Measures 3, 4, 5: $E_{\text {sim }}(c), \operatorname{Var}_{\text {sim }}(c), C_{\text {sim }}(c)$, both $s_{1}$ and $s_{2}$ are random in each run; 100,000 runs.

| Design | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0} \times \mathbf{C V}_{\text {sim }}(\boldsymbol{c})$ |
| :--- | :---: | :---: | :---: |
| Poisson | 1.96 | 1.12 | 54.41 |
| Pareto | 1.98 | 1.03 | 51.03 |
| ASCP_ $\boldsymbol{\pi}$ | 2.24 | 0.63 | 35.38 |
| ASCP_ $\pi$ inf | 2.18 | 0.56 | 34.28 |
| ASCP_inf | 2.10 | 0.61 | 37.20 |
| ASCP_inf without coordination | 3.05 | 0.82 | 29.78 |

### 3.3.3 Method 2, framework 2: $s_{1}=\{4,6,12,18,22,35,44\}$ is fixed and is a Poisson sample, while $s_{2}$ is random in each run

Table 3.7
MU284 population, Method 2, framework 2, Measure 1: number of pairs ( $s_{1}, s_{2}$ ) by possible number of nondesired units in common, $s_{1}=\{4,6,12,18,22,35,44\}$ is fixed in each run over 100,000 runs (so the row sums are equal to 100,000 ).

|  | Possible number of non-desired units in common |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Design | $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ |
| Poisson | 11,970 | 38,609 | 36,849 | 12,033 | 539 |
| Pareto | 14,497 | 47,328 | 33,006 | 5,085 | 84 |
| ASCP_ $\pi$ | 2,449 | 69,047 | 24,291 | 4,183 | 30 |
| ASCP_ $\pi$ inf | 0 | 81,609 | 17,689 | 702 | 0 |
| ASCP_inf | 0 | 84,227 | 15,773 | 0 | 0 |
| ASCP_inf without coordination | 0 | 18,869 | 60,320 | 20,802 | 9 |

Table 3.8
MU284 population, Method 2, framework 2, Measure 2: $s_{1}=\{4,6,12,18,22,35,44\}$ is fixed in each run; 100,000 runs.

|  |  | $\hat{\boldsymbol{P}}\left(\boldsymbol{k} \in \boldsymbol{s}_{\mathbf{1}}, \boldsymbol{k} \in \boldsymbol{s}_{\mathbf{2}}\right)$ | $\mathbf{2 2}$ | $\mathbf{3 2}$ | $\mathbf{4 4}$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Non-desired unit $\boldsymbol{k}$ | $\mathbf{4}$ | $\mathbf{1 2}$ | $\mathbf{2 1}$ | 0.40 |  |
| Poisson | 0.64 | 0.05 | 0 | 0.41 | 0 |
| Pareto | 0.59 | 0.04 | 0 | 0.33 | 0 |
| ASCP_ $\pi$ | 0.64 | 0.13 | 0 | 0.32 | 0 |
| ASCP_ $\pi$ inf | 0.64 | 0.05 | 0 | 0.36 | 0 |
| ASCP_inf | 0.64 | 0.09 | 0 | 0.32 | 0.14 |
| ASCP_inf without coordination | 0.65 | 0.37 | 0 | 0.50 | 0.11 |

Table 3.9
MU284 population, Method 2, framework 2, Measures 3, 4, 5: $E_{\text {sim }}(c), \operatorname{Var}_{\text {sim }}(c), \mathrm{CV}_{\text {sim }}(c), s_{1}=\{4,6,12,18,22$, $35,44\}$ is fixed in each run; 100,000 runs.

| Design | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0 \times ( \mathbf { V } _ { \text { sim } } ( \boldsymbol { c } )}$ |
| :--- | :---: | :---: | :---: |
| Poisson | 1.51 | 0.76 | 57.99 |
| Pareto | 1.29 | 0.60 | 60.19 |
| ASCP_ $\pi$ | 1.30 | 0.35 | 45.12 |
| ASCP_ $\pi$ inf | 1.19 | 0.17 | 34.47 |
| ASCP_inf | 1.16 | 0.13 | 31.48 |
| ASCP_inf without coordination | 2.31 | 0.64 | 34.66 |

### 3.4 Simulation with business surveys: Application to "hot-spot units" at Statistics Netherlands

Statistics Netherlands (CBS) operates a sample coordination system for business surveys. Despite the sample coordination, each year several businesses are still heavily sampled, mainly because of the number of drawn samples, different stratification schemes and large sampling fractions. This results in a large cumulated response burden for these specific businesses. Therefore, CBS started monitoring the number of surveys for which a business was sampled within the last twelve months to identify so-called "hot-spot
units" (equivalent to non-desired units, see Section 3.1). CBS classifies businesses with 0-9 employees (size classes 0-3) as hot-spot units if they are sampled $\geq 3$ times within the last twelve months. Businesses with 10-19 employees (size class 4) are classified as "hot-spot units" if they are sampled $\geq 4$ times within the last twelve months. For larger businesses, no hot-spot units were defined despite the large sample fractions that are required for these businesses. Businesses with more than 50 employees are usually sampled with inclusion probability 1 . These large businesses usually have dedicated staff to fill in the questionnaires. The impact on daily business is therefore lower for these businesses. Moreover, sample coordination is generally not suitable for businesses with a sampling fraction of 1 .

For the application we consider the population of Dutch businesses in 2021 with $0-19$ employees ( $N=$ $1,810,581$ ), the Structural Business Survey (SBS) with sample size $n_{\text {SBS }}=54,491$, the Investment Survey (INV) with $n_{\mathrm{INV}}=30,090$ and the Finance Monitor (FIN) with $n_{\mathrm{FIN}}=6,977$. In this population there are 1,693 hot-spot units. Table 3.10 shows the distribution of businesses by size class (columns) and number of samples (rows). The hot-spot units are highlighted in italics. In 2021, there were no businesses with 0-19 employees that were selected in more than nine surveys by the Dutch coordination system.

All three surveys are annual surveys with a stratified sample design with equal probability within strata. The strata are defined by a combination of industrial classification according to NACE (Nomenclature statistique des Activités économiques dans la Communauté Européenne, the standard European industrial classification.) and size class. SBS and INV are coordinated by the Dutch coordination system (see Smeets and Boonstra, 2018) and use the same stratification. FIN is independent from SBS and INV, but the system coordinates the samples for FIN from year to year. FIN only selects businesses with $\geq 2$ employees (size class 2 and larger) and uses different combinations of NACE codes than SBS and INV to define the strata.

We consider the following scenarios using Methods 1 and 2 (see Section 3.2) and different combinations of the surveys:

1. Method 1 , where both samples $s_{1}$ and $s_{2}$ are drawn for SBS. Both samples use the SBS allocation of 2021 (scenario 1);
2. Method 1 , where $s_{1}$ is drawn for SBS and $s_{2}$ is drawn for INV. Both samples use the allocation of the corresponding survey of 2021 (scenario 2);
3. Method 2, where $s_{2}$ is drawn for SBS, conditional on the existing SBS sample $s_{1}$ of 2021, i.e., $s_{1}$ is fixed. Sample $s_{2}$ uses the SBS allocation of 2021 (scenario 3);
4. Method 2, where $s_{2}$ is separately drawn for SBS, INV and FIN, conditional on sample $s_{1}$, that is obtained by combining the existing samples of SBS, INV and FIN of 2021, i.e., $s_{1}$ is fixed. The samples for $s_{2}$ use the allocation of the corresponding survey of 2021. It is possible that a unit is drawn for each of the three separate samples (scenario 4).

Table 3.10
Distribution of all Dutch businesses with 0-19 employees in 2021 by size class and number of surveys; hot-spot units are highlighted in italics.

| Number of surveys | Size class |  |  |  |  |  |  |  |  |  |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathbf{0}$ | $\mathbf{1}$ |  |  |  |  |  | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | Total |
| 0 | 381,591 | $1,051,551$ | 202,137 | 39,544 | 8,047 | $1,682,870$ |  |  |  |  |  |
| 1 | 4,832 | 40,580 | 35,529 | 16,640 | 11,732 | 109,313 |  |  |  |  |  |
| 2 | 85 | 985 | 2,509 | 3,196 | 7,194 | 13,969 |  |  |  |  |  |
| 3 | 8 | 34 | 125 | 423 | 2,736 | 3,326 |  |  |  |  |  |
| 4 | 0 | 4 | 6 | 46 | 785 | 841 |  |  |  |  |  |
| 5 | 0 | 1 | 1 | 3 | 167 | 172 |  |  |  |  |  |
| 6 | 0 | 0 | 1 | 1 | 56 | 58 |  |  |  |  |  |
| 7 | 0 | 0 | 0 | 1 | 22 | 23 |  |  |  |  |  |
| 8 | 0 | 0 | 0 | 0 | 7 | 7 |  |  |  |  |  |
| 9 | 0 | 0 | 0 | 0 | 2 | 2 |  |  |  |  |  |
| Total | 386,516 | $1,093,155$ | 240,308 | 59,854 | 30,748 | $1,810,581$ |  |  |  |  |  |

In this simulation, method 2 (in scenarios 3 and 4) is only considered under framework 2 (see Section 3.3). Scenarios 1 and 3 represent coordination over time for one survey. Scenario 2 represents coordination over time and over two surveys with common stratification. Scenario 4 represents coordination over time and over two surveys with different stratifications. Because of the large population sizes and since the majority of the strata do not contain any hot-spots at all, in all scenarios a selection of strata is used for the simulation. First, in scenarios 1 and 3 the take-all strata $\left(\pi_{k 1}=\pi_{k 2}=1\right)$ are excluded and in scenarios 2 and 4 both the take-all $\left(\pi_{k 2}=1\right)$ and take-none strata $\left(\pi_{k 2}=0\right)$ are excluded. In scenario 4 , a stratum is only excluded if $\pi_{k 2}=0$ or 1 in this stratum for all three surveys. Second, strata are selected based on the population size and the expected number of sampled hot-spot units, such that the total population size $N$ is around 1,000 . In scenarios 1 and 3, strata with less than 300 businesses in the population and at least 10 expected hot-spots in the sample are selected. In scenario 2, strata with less than 500 businesses and at least six expected hot-spots are selected. In scenario 4, all strata with at least one expected hot-spot in the sample for all three surveys are selected. Table 3.11 gives the population size and sample information for the considered scenarios. The populations of businesses and hot-spot units are a subset of the populations shown in Table 3.10. The sample size of the combined samples in scenario 4 is denoted by $n_{\text {comb }}$.

Both the information of whether a particular business is considered a hot-spot (binary information as used in Section 3.3) and the number of surveys for which the business was sampled within the last 12 months are measures of its cumulated response burden. In the simulation with CBS business surveys, we use both measures to compute the Euclidean distance between the units. However, because of the stratified sampling with equal inclusion probabilities within the strata, the inclusion probabilities of the surveys are not used to compute the Euclidean distances between the units in this simulation, as in Section 3.3.

Table 3.11
Population and sample information for the scenarios.

| scenario | method | $N$ | strata | hot-spots | sample $s_{1}$ |  | sample $s_{2}$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  | $n_{\text {SBS }}$ | $\boldsymbol{n}_{\text {comb }}$ | $\boldsymbol{n}_{\text {SBS }}$ | $n_{\text {INV }}$ | $n_{\text {FIN }}$ |
| 1 | 1 | 894 | 5 | 97 | 617 |  | 617 |  |  |
| 2 | 1 | 1,053 | 8 | 161 | 659 |  |  | 487 |  |
| 3 | 2 | 894 | 5 | 97 | 617 |  | 617 |  |  |
| 4 | 2 | 418 | 34 | 62 |  | 418 | 384 | 278 | 37 |

We provide the results of a Monte-Carlo simulation for the considered scenarios and different sampling schemes. The sampling schemes are applied per stratum. Due to the substantial computational burden, 5,000 runs are used. In each run of the Monte-Carlo simulation we draw:

- two Poisson samples for Method 1 and one Poisson sample for Method 2;
- two Pareto samples for Method 1 and one Pareto sample for Method 2;
- two adapted SCP samples for Method 1 and one adapted SCP sample for Method 2; for $s_{1}, s_{2}$ the measure of cumulated response burden based on hot-spot status is used to define the Euclidean distances between the units (indicated by ASCP_inf);
- two adapted SCP samples for Method 1 and one adapted SCP sample for Method 2; for $s_{1}, s_{2}$ the measure of cumulated response burden given by the number of surveys is used to compute the Euclidean distances between the units (indicated by ASCP_inf_svy);
- two independent adapted SCP samples (without negative coordination) for Method 1 and one independent adapted SCP sample for Method 2; for $s_{1}, s_{2}$ the measure of cumulated response burden based on the hot-spot status is used to compute the Euclidean distance between the units (indicated by ASCP_inf without coordination).

In the tables and figures below, we use the following notation: ASCP_inf indicates results based on adapted SCP sampling with hot-spot status used to define the Euclidean distances between the units and negative coordination, ASCP_inf_svy for adapted SCP sampling with cumulated response burden and negative coordination (svy stands for survey), while "ASCP_inf without coordination" for adapted SCP sampling and independent sample selection. The results of scenarios 1, 2, and 3 are shown in Figure 3.1 and Table 3.12. The results of scenario 4 are shown in Figure 3.2 and Table 3.13 (for ASCP_inf hot-spot status is used as the burden measure and for ASCP_inf_svy the number of surveys is used). In scenarios 1,2 and 3 all measures give similar results to the simulation with the MU284 population. The results of Measure 2 are in line with the results presented for the MU284 population and are not shown here to save space. In these scenarios, adapted SCP sampling with the measure of cumulated response burden based on hot-spot status is the best sampling strategy. This is because ASCP_inf leads to the smallest variation of the overlaps, not only for all businesses but also for the hot-spot units. This implies that the overall response burden is most evenly spread by ASCP_inf. In scenario 4 the differences between the sampling schemes are smaller. This is caused by FIN using different strata than SBS and INV, which leads to small strata when adapted SCP sampling is applied as a coordination method for the three surveys together. When $s_{2}$ is drawn for SBS or INV, the ASCP_inf_svy sampling scheme is slightly better than ASCP_inf. When in scenario 4 sample $s_{2}$ is drawn for the SBS, the adapted SCP sampling and independent sampling perform similarly. This has to do with the selection of the strata in this scenario. The inclusion probabilities of SBS and INV are large in these strata, while the inclusion probabilities of FIN are small. Moreover, 28 strata are take-all strata for SBS and 17 strata are take-all for INV, while FIN has no take-all strata in this selection. When Method 2 is applied as the coordination method and if $k \in s_{1}$, the random number $u_{k 2}$ is generated from Unif $\left(0, \pi_{k 1}\right)$. If $\pi_{k 1}$ is close to 1 then $u_{k 2}$ is generated from a distribution that is approximately equal to $\operatorname{Unif}(0,1)$. Generating $u_{k 2}$ from $\operatorname{Unif}(0,1)$ implies independent sampling.

Figure 3.1 Scenarios 1, 2, 3; Measure 1: number of pairs of samples ( $s_{1}, s_{2}$ ) (bullets) by possible number of hot-spot units in common ( $\mathbf{y}$-axis), $\mathbf{5 , 0 0 0}$ runs. The size of the bullets is an indication of the number of sampled pairs.


Figure 3.2 Scenario 4; Measure 1: number of pairs of samples ( $s_{1}, s_{2}$ ) (bullets) by possible number of hot-spot units in common ( $\mathbf{y}$-axis), 5,000 runs. The size of the bullets is an indication of the number of sampled pairs.


Table 3.12
Scenarios 1, 2 and 3; Measures 3, 4, 5: $E_{\text {sim }}(c), \operatorname{Var}_{\text {sim }}(c), \mathrm{CV}_{\text {sim }}(c), 5,000$ runs.

| Design - Scenario 1 | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0} \times \mathbf{C V}_{\text {sim }}(\boldsymbol{c})$ |
| :--- | :---: | :---: | :---: |
| Poisson | 352 | 109.0 | 2.97 |
| Pareto | 352 | 20.3 | 1.28 |
| ASCP_inf | 357 | 0.7 | 0.23 |
| ASCP_inf_svy | 365 | 2.5 | 0.45 |
| ASCP_inf without coordination | 455 | 35.2 | 1.31 |
| Design - Scenario 2 | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0} \times \mathbf{C V}_{\text {sim }}(\boldsymbol{c})$ |
| Poisson | 211 | 136.0 | 5.53 |
| Pareto | 211 | 49.9 | 3.35 |
| ASCP_inf | 212 | 0.6 | 0.38 |
| ASCP_inf_svy | 214 | 2.4 | 0.73 |
| ASCP_inf without coordination | 300 | 35.4 | 1.98 |
| Design - Scenario $\mathbf{3}$ | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0 \times \mathbf { C V } _ { \text { sim } } ( \boldsymbol { c } )}$ |
| Poisson | 352 | 68.7 | 2.36 |
| Pareto | 352 | 14.3 | 1.07 |
| ASCP_inf | 357 | 1.4 | 0.34 |
| ASCP_inf_svy | 365 | 3.6 | 0.52 |
| ASCP_inf without coordination | 455 | 35.5 | 1.31 |

Table 3.13
Scenario 4; Measures 3, 4, 5: $E_{\text {sim }}(c), \operatorname{Var}_{\text {sim }}(c), \mathrm{CV}_{\text {sim }}(c), 5,000$ runs.

| Design - Scenario 4 SBS | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0} \times \mathbf{C V}_{\text {sim }}(\boldsymbol{c})$ |
| :--- | :---: | :---: | :---: |
| Poisson | 375 | 11.10 | 0.890 |
| Pareto | 378 | 0.00 | 0.000 |
| ASCP_inf | 379 | 0.91 | 0.252 |
| ASCP_inf_svy | 379 | 0.87 | 0.246 |
| ASCP_inf without coordination | 379 | 0.92 | 0.254 |
| Design - Scenario 4 INV | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0 \times \mathbf { C V } _ { \text { sim } } ( \boldsymbol { c } )}$ |
| Poisson | 271 | 36.00 | 2.210 |
| Pareto | 274 | 0.00 | 0.000 |
| ASCP_inf | 273 | 1.93 | 0.508 |
| ASCP_inf_svy | 273 | 1.99 | 0.516 |
| ASCP_inf without coordination | 274 | 1.57 | 0.584 |
| Design - Scenario 4 FIN | $\boldsymbol{E}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{V a r}_{\text {sim }}(\boldsymbol{c})$ | $\mathbf{1 0 0 \times \mathbf { C V } _ { \text { sim } } ( \boldsymbol { c } )}$ |
| Poisson | 21 | 12.20 | 17.00 |
| Pareto | 24 | 0.00 | 0.00 |
| ASCP_inf | 21 | 3.65 | 8.95 |
| ASCP_inf_svy | 21 | 3.71 | 9.13 |
| ASCP_inf without coordination | 24 | 5.11 | 9.40 |

## 4. Discussion

The strategy developed in Section 3.1 provides an approach to negative coordination of samples which fulfils the requirement to reduce the overlap size between two or more samples, and additionally reduces the variance of the number of nondesirable units (units with particular characteristics) in the overlap compared to its competitors. Thus, a double control of the response burden is targetted at this specific set of units. The coordination strategy results in a more even spread of the response burden of these units. The
targetting can be achieved through an indicator variable, or through a continuous variable demonstrating the size of a unit for the characteristic of interest (such as the cumulative response burden).

We consider several variants of the approach, depending on the kind of information used to designate non-desired units. In general ASCP_inf performs the best because it uses only binary information and nondesired units are therefore as similar as possible to each other on this characteristic, and as different as possible from other units. This approach is therefore better at avoiding samples with clusters of non-desired units. Other variants may however be better in situations where there is a gradation of non-desirability. In our simulations, the proposed strategy shows a smaller variance of the number of non-desired units in the overlap (especially for ASCP_inf) compared to Poisson and Pareto sampling. This is due to the spread of the units in the space generated by the measure of response burden used (spread obtained by using the algorithm given in Section 2.4), as indicated in the tables and figures related to Measure 1. On the other hand, similar results to the competitor methods were obtained for the expected overlap size, while the variance of the overlap size was smaller than for Poisson and Pareto sampling with PRNs in most cases. A single exception concerning this variance was provided by Pareto sampling with PRNs in Table 3.12.

Targetted double control is an effective strategy for managing situations where some businesses are selected for relatively many surveys in a short period, as demonstrated by the application to hot-spot units in Statistics Netherlands. The problem of hot-spots is not eliminated, but it is reduced because the response burden is more controlled within the constraints of the survey designs.

The targetted double control strategy can therefore be used to formalise an approach to dealing with businesses that complain that they have been selected in too many surveys. Without such a system, these are sometimes dealt with in an ad hoc way by moving them (explicitly or implicitly) to a take-none stratum, to relieve the burden. But this approach is not fair in that it can be different for businesses with the same characteristics depending on whether they complain or not. For a single sample selected at any given moment, targetted double control is "fair" in that it minimises the number of such units included in the sample by spreading the selected units through the space generated by the measure of the response burden used (and thus avoids the clustering of non-desired units) and because it respects the inclusion probabilities so that unbiased estimates can be obtained. Statistics Canada have considered extending a take-none stratum to deal with small units which may receive a disproportionate burden (Landry, 2011), but we consider that using targetted double control would be a better solution in this case too. Targetted double control therefore addresses an important practical problem in the coordination of multiple samples in a finite population.

Applications of targetted double control are not restricted to selection hot-spots, however, and any kind of undesirable unit could in principle be the target of the method, as long as the undesirability property is observable or predictable from available data sources.

## Acknowledgements

We are grateful to the referees for their comments.

## References

Bondesson, L., and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. Scandinavian Journal of Statistics, 35, 466-483.

Bottone, M., Modugno, L. and Neri, A. (2021). Response burden and data quality in business surveys. Journal of Official Statistics, 37(4), 811-836.

Bradburn, N. (1978). Respondent burden. In Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, 35, 35-40.

Brewer, K., Early, L. and Joyce, S. (1972). Selecting several samples from a single population. Australian Journal of Statistics, 3, 231-239.

Cotton, F., and Hesse, C. (1992). Tirages coordonnés d'échantillons. Technical Report E9206, Direction des Statistiques Économiques, INSEE, Paris, France.

Grafström, A. (2012). Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference, 142(1), 139-147.

Grafström, A., Lisic, J. and Prentius, W. (2022). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.6.3, https://CRAN.R-project.org/package=BalancedSampling.

Grafström, A., and Matei, A. (2018). Coordination of spatially balanced samples. Survey Methodology, 44, 2, 215-238. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54953eng.pdf.

Landry, S. (2011). Managing response burden by controlling sample selection and survey coverage. Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, 2011, 2276-2284.

Lorenc, B., Kloek, W., Abrahamsson, L. and Eckman, S. (2013). Description of actual response burden and response behaviour with Swedish register. In Comparative Report on Integration of Case Study Results Related to Reduction of Response Burden and Motivation of Businesses for Accurate Reporting, (Eds., D. Giesen, M. Bavdaž and I. Bolko), BLUE-ETS Deliverable 8.1, 11-27. doi: 10.13140/RG.2.2. 24741.63207.

Mach, L., Reiss, P.T. and Şchiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the northwest corner rule. Journal of the American Statistical Association, 101(476), 1671-1679.

Matei, A., and Smith, P.A. (2023). Sample coordination methods and systems for establishment surveys. In Advances in Business Statistics, Methods and Data Collection, (Eds., G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J. Sakshaug, K. Thompson and A. van Delden), Chapter 27. Hoboken: Wiley.

Matei, A., and Tillé, Y. (2005). Maximal and minimal sample co-ordination. Sankhyā, 67(3), 590-612.

Natkowska, M., and Modak, A. (2014). Theme: Sample Co-ordination. European Commission, Eurostat, Directors of Methodology.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In Business Survey Methods (Eds., B.G. Cox, D.A. Binder, B.N. Chinnapa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley \& Sons, Inc., USA, Chapter 9, 153-169.

Qualité, L. (2019). Coordinated sampling: Theory, method and application at SFSO. Presentation, ENBES workshop on Coordinated Sampling for Business Surveys. https://statswiki.unece.org/display/ENBES/ ENBES+Workshop+2019+\%28The+Hague\%29+-+Workshop+on+Coordinated+Sampling?preview=/ 234225830/244092344/Slides_Qualite_2019.pdf.

Rosén, B. (1997a). Asymptotic theory for order sampling. Journal of Statistical Planning and Inference, 62, 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. Journal of Statistical Planning and Inference, 62, 159-191.

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

Smeets, M., and Boonstra, H.J. (2018). Sampling coordination of business surveys at Statistics Netherlands. In The Unit Problem and Other Current Topics in Business Survey Methodology, (Eds., B. Lorenc, P.A. Smith, M. Bavdaž, G. Haraldsen, D. Nedyalkova, L.-C. Zhang and T. Zimmermann), Cambridge Scholars, Newcastle-upon-Tyne, 127-137.

Sunter, A.B. (1977). Response burden, sample rotation, and classification renewal in economic surveys. International Statistical Review, 45, 209-222.

# QR prediction for statistical data integration 

Estelle Medous, Camelia Goga, Anne Ruiz-Gazen, Jean-François Beaumont, Alain Dessertaine and Pauline Puech ${ }^{1}$


#### Abstract

In this paper, we investigate how a big non-probability database can be used to improve estimates of finite population totals from a small probability sample through data integration techniques. In the situation where the study variable is observed in both data sources, Kim and Tam (2021) proposed two design-consistent estimators that can be justified through dual frame survey theory. First, we provide conditions ensuring that these estimators are more efficient than the Horvitz-Thompson estimator when the probability sample is selected using either Poisson sampling or simple random sampling without replacement. Then, we study the class of QR predictors, introduced by Särndal and Wright (1984), to handle the less common case where the non-probability database contains no study variable but auxiliary variables. We also require that the non-probability database is large and can be linked to the probability sample. We provide conditions ensuring that the QR predictor is asymptotically design-unbiased. We derive its asymptotic design variance and provide a consistent design-based variance estimator. We compare the design properties of different predictors, in the class of QR predictors, through a simulation study. This class includes a model-based predictor, a model-assisted estimator and a cosmetic estimator. In our simulation setups, the cosmetic estimator performed slightly better than the model-assisted estimator. These findings are confirmed by an application to La Poste data, which also illustrates that the properties of the cosmetic estimator are preserved irrespective of the observed non-probability sample.


Key Words: Cosmetic estimator; Dual frame; GREG estimator; Non-probability sample; Probability sample; Variance estimator.

## 1. Introduction

In the field of economics and social sciences, surveys are usually based on probability sampling methods. At the French postal service (La Poste) for example, the postal traffic is estimated through quarterly probability surveys. Controlling the sampling design allows for design-based inference without resorting to modeling of the study variables; this feature is attractive to many survey statisticians. Neyman (1934) is usually known as the founding paper of probability sampling theory. Since then, the literature on this topic has grown rapidly with an interplay between theory and practice (see Rao (2005) for the most important contributions).

Recently, survey statisticians have observed a decline in response rates together with an increase of the survey costs, which make probability sampling more challenging. In addition, large non-probability samples, such as administrative data or web-based surveys, become available often at low cost (see, e.g., Beaumont (2020) and Rao (2021) for more details). These observations are also true at La Poste where, for cost reasons, the size of probability samples is bound to decrease while a big database containing the automatically processed postal mail is available. Even if non-probability samples are associated with

[^4]unknown selection mechanisms and may suffer from selection bias and measurement errors, they provide timely information on the population of interest. This context leads survey statisticians to study the integration or combination of data from probability and non-probability samples.

The literature on data integration in survey sampling has grown rapidly recently, and the reader may refer to several reviews on the subject (see Beaumont (2020), Yang and Kim (2020), Rao (2021), Kim (2022) and $\mathrm{Wu}(2022)$ ). If we focus on the problem of combining probability and non-probability samples, the different data integration methods can be divided into three groups depending on whether the study variable is observed in the probability sample only, in the non-probability sample only, or in both samples (see e.g., Rao (2021)). Most methods tackle the problem of the study variable observed in the nonprobability sample only, e.g., Kim (2022). In this context, the objective is to address the selection bias by combining data from the non-probability sample with auxiliary data available in a probability sample.

At La Poste, the problem is rather that the study variables (such as the different types of mails sent) are only available in the probability sample whereas auxiliary information is only available in the nonprobability database. Such a context is rather rare in practice and has therefore not been studied in detail so far. The aim of the present paper is to study this particular context thoroughly. The method we recommend and study in detail is applicable nicely if (i) the overlap between the probability and the non-probability samples is ideally large but at least non-empty, and (ii) it is possible to accurately match observations from both samples. At La Poste, the non-probability sample represents more than $80 \%$ of the population; as a result, its intersection with the probability sample is large. The matching, however, is a difficult task that La Poste is still investigating.

In the situation where the study variables are measured in both samples, Kim and Tam (2021) propose a design-based dual frame approach to improve the efficiency of the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), which uses the probability sample only. The total of the study variable over the whole population is estimated by summing the true total over the non-probability sample and an estimator of the total over the complementary of the non-probability sample. Kim and Tam (2021) propose several estimators that can be deduced from a calibration perspective.

In Section 2, we revisit the approach of Kim and Tam (2021) and derive general results on the efficiency of their proposed dual frame estimators. In the situation where the study variable is not measured in the nonprobability sample, we propose to replace the true unknown total over the non-probability sample by some prediction. In Section 3, we adapt the general class of QR predictors, introduced in Wright (1983), to data integration. This class of estimators includes the well-known model-assisted (GREG) and model-based estimators, but also the cosmetic estimator (Särndal and Wright, 1984). We first exhibit a condition under which the QR predictors can be written in a projection form. We then derive a QR condition such that these predictors are identical to model-assisted predictors. In Section 4, we look at the asymptotic properties of the QR estimators. We show that they are asymptotically unbiased under the model and the sampling design. We also prove that, under the QR condition, the predictors are asymptotically design-unbiased. We derive their asymptotic design variance and provide a design-consistent variance estimator. In Section 5, we use

Monte Carlo simulations to compare several QR predictors and show that the cosmetic estimator is a good compromise for several setups. In Section 6, we consider an application to La Poste data and illustrate the impact of the non-probability sample on the estimators. Finally we conclude and give perspectives in Section 7.

## 2. Study variable observed in both samples

We are interested in estimating the population total $T=\sum_{k \in U} y_{k}$, where $y_{k}$ is the value of the variable of interest $Y$ for unit $k$ of the population $U$. A probability sample $s_{P}$ is drawn from $U$ using a sampling design $p\left(s_{P} \mid \mathbf{Z}\right)$, where the population matrix $\mathbf{Z}$ contains design information such as strata identifiers. The sample inclusion indicator, $I_{k}, k \in U$, takes the value 1 if unit $k$ is selected in $s_{P}$, and 0 otherwise. The probability that a given population unit $k$ is selected in $s_{P}$ is $\pi_{k}=\mathrm{E}_{p}\left(I_{k} \mid \mathbf{Z}\right)$. We assume in the present section that the variable of interest $Y$ is observed for each unit of the probability sample but also for each unit in the non-probability sample $s_{\mathrm{NP}} \subset U$. The inclusion indicator in $s_{\mathrm{NP}}$ for the population unit $k \in U$ is denoted as $\delta_{k}$ (i.e., $\delta_{k}=1$, if $k \in s_{\mathrm{NP}}$, and $\delta_{k}=0$, otherwise). We assume that $\delta_{k}$ is available for each unit of the probability sample $s_{P}$. Denote by $N$ (resp. $N_{\mathrm{NP}}$ ) the size of $U$ (resp. $s_{\mathrm{NP}}$ ) and by $n$ the expected size of $s_{P}$. Let $\hat{T}_{\mathrm{HT}}=\sum_{k \in s_{p}} d_{k} y_{k}$ be the well-known expansion or Horvitz-Thompson estimator of $T$ with the sampling weights $d_{k}=1 / \pi_{k}$. If $\pi_{k}>0$, for all $k \in U, \hat{T}_{\mathrm{HT}}$ is a design-unbiased estimator of $T$.

The non-probability sample $s_{\mathrm{NP}}$ is usually a cheap and large source of data. Its selection mechanism is unknown, and its selection bias cannot be ignored when making inference. On the other hand, the probability sample $s_{P}$ is assumed representative (without selection bias), yet often expensive and of relatively small size. By combining information from the two samples, we can expect to find an estimator more precise than the expansion estimator obtained using $s_{p}$.

Kim and Tam (2021) propose two estimators using combined data from $s_{P}$ and $s_{\mathrm{NP}}$ and we propose to revisit the properties of these estimators. The total can be decomposed as:

$$
T=T_{\mathrm{NP}}+T_{C},
$$

where $T_{\mathrm{NP}}=\sum_{k \in s_{\mathrm{NP}}} y_{k}=\sum_{k \in U} \delta_{k} y_{k}$ and $T_{C}=\sum_{k \in U-s_{\mathrm{NP}}} y_{k}=\sum_{k \in U}\left(1-\delta_{k}\right) y_{k}$. Since $y_{k}$ is measured for all units of $s_{\mathrm{NP}}, T_{\mathrm{NP}}$ is known, and we only have to estimate $T_{C}$. Kim and Tam (2021) propose the following estimator:

$$
\begin{equation*}
\hat{T}_{\mathrm{DI}}=T_{\mathrm{NP}}+\sum_{k \in S_{P}} d_{k}\left(1-\delta_{k}\right) y_{k}, \tag{2.1}
\end{equation*}
$$

where $T_{C}$ is estimated using the expansion estimator. As pointed out by Beaumont (2020), this can be viewed as a dual frame problem, with frames $U$ and $s_{\mathrm{NP}}$, where the sample $s_{P}$ is randomly selected from $U$ and a census is taken from $s_{\mathrm{NP}}$. In this context of two sampling frames, $\hat{T}_{\mathrm{DI}}$ is an estimator that results from a direct application of the method proposed by Bankier (1986). One may think that $\hat{T}_{\mathrm{DI}}$ is more efficient
than $\hat{T}_{\mathrm{HT}}$, especially if the size of the non-probability sample is large, but this is not true in general. The following proposition shows that, while the variance of $\hat{T}_{\mathrm{DI}}$ is always smaller than the variance of $\hat{T}_{\mathrm{HT}}$ for Poisson sampling, the property is only true under a condition on the study variable for simple random sampling without replacement.

## Proposition 2.1

(i) For Poisson sampling, the variance of $\hat{T}_{\mathrm{DI}}$ is less than or equal to the variance of $\hat{T}_{\mathrm{HT}}$.
(ii) For simple random sampling without replacement, the variance of $\hat{T}_{\mathrm{DI}}$ is less than or equal to the variance of $\hat{T}_{\mathrm{HT}}$ if and only if

$$
\mathrm{CV}_{\mathrm{NP}}^{2} \geq-\frac{N_{\mathrm{NP}}}{N_{\mathrm{NP}}-1}\left(1+\frac{N_{\mathrm{NP}}}{N}-2 \frac{\bar{Y}_{U}}{\bar{Y}_{\mathrm{NP}}}\right),
$$

where $\bar{Y}_{U}=\frac{1}{N} \sum_{k \in U} y_{k}$ is the mean of $Y$ over $U, \bar{Y}_{\mathrm{NP}}=\frac{1}{N_{\mathrm{NP}}} \sum_{k \in U} \delta_{k} y_{k}$ is the mean of $Y$ over $s_{\mathrm{NP}}$, and $\mathrm{CV}_{\mathrm{NP}}=\sqrt{S_{Y, \mathrm{NP}}^{2}} / \bar{Y}_{\mathrm{NP}}$ is the coefficient of variation of $Y$ in $s_{\mathrm{NP}}$, with $S_{Y, \mathrm{NP}}^{2}=$ $\frac{1}{N_{\mathrm{NP}}-1} \sum_{k \in U} \delta_{k}\left(y_{k}-\bar{Y}_{\mathrm{NP}}\right)^{2}$.

The proof of Proposition 2.1 is given in the appendix. Intuitively, the result (ii) of Proposition 2.1 can be explained by the fact that the size of $s_{P}$ is fixed for simple random sampling without replacement in the expression of $\hat{T}_{\mathrm{HT}}$ while the size of $s_{P} \cap U-s_{\mathrm{NP}}$ is random for $\hat{T}_{\mathrm{DI}}$. In other words, the estimator $\hat{T}_{\mathrm{DI}}$ is calibrated on $N_{\mathrm{NP}}$ and $T_{\mathrm{NP}}$, but not on $N$ while $\hat{T}_{\mathrm{HT}}$ is calibrated on $N$.

If the size of the population $U$ is known, Kim and Tam (2021) propose to improve $\hat{T}_{\mathrm{DI}}$ by using the following estimator:

$$
\hat{T}_{\mathrm{PDI}}=T_{\mathrm{NP}}+\hat{T}_{C}^{(\mathrm{Ha})},
$$

where

$$
\hat{T}_{C}^{(\mathrm{Ha})}=\left(N-N_{\mathrm{NP}}\right) \frac{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right) y_{k}}{\sum_{k \in s_{P}} d_{k}\left(1-\delta_{k}\right)}
$$

is a Hájek-type estimator of the total $T_{C}$. Kim and Tam (2021) proved that $\hat{T}_{\mathrm{PDI}}$ is a Generalized Regression (GREG) estimator calibrated on $N, N_{\mathrm{NP}}$ and $T_{\mathrm{NP}}$. Its expression can be further generalized by including additional calibration variables.

Following Kim and Tam (2021), it is possible to use the linearization approach and derive the approximate variance of $\hat{T}_{\mathrm{PDI}}$, denoted as $\mathrm{AVar}\left(\hat{T}_{\mathrm{PDI}}\right)$. For Poisson sampling, the independence of the inclusion indicators reduces the comparison of $\hat{T}_{\mathrm{PDI}}$ and $\hat{T}_{\mathrm{DI}}$ to the comparison of Horvitz-Thompson and Hájek estimators of the total $T_{C}=\sum_{k \in U}\left(1-\delta_{k}\right) y_{k}$. The Hájek estimator can be significantly more efficient than the Horvitz-Thompson estimator but it is not true in general (see, e.g., Särndal, Swensson and Wretman (1992)).

For Poisson sampling, the estimator $\hat{T}_{\text {PDI }}$ can be substantially more efficient than the Horvitz-Thompson estimator $\hat{T}_{\mathrm{HT}}$, as illustrated in our simulation study in Section 5. For simple random sampling without replacement, the approximate variance of $\hat{T}_{\mathrm{PDI}}$ can be compared to the variance of $\hat{T}_{\mathrm{HT}}$ in more general conditions than in Kim and Tam (2021). Proposition 2.2 below shows that the approximate variance of $\hat{T}_{\mathrm{PDI}}$ is smaller than the variance of $\hat{T}_{\mathrm{HT}}$ for simple random sampling without replacement, and gives the expression of the difference between the variances.

Proposition 2.2 For simple random sampling without replacement with sampling fraction $f=n / N$,

$$
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)-\mathrm{A} \operatorname{Var}\left(\hat{T}_{\mathrm{PDI}}\right)=\frac{N^{2}(1-f)}{(N-1) n}\left(\sum_{k \in U} \delta_{k}\left(y_{k}-\bar{Y}_{U}\right)^{2}+\sum_{k \in U}\left(1-\delta_{k}\right)\left(\bar{Y}_{C}-\bar{Y}_{U}\right)^{2}\right),
$$

where $\bar{Y}_{U}=\frac{1}{N} \sum_{k \in U} y_{k}$ is the mean of $Y$ over $U$, and $\bar{Y}_{C}=\frac{1}{N-N_{N P}} \sum_{k \in U}\left(1-\delta_{k}\right) y_{k}$ is the mean of $Y$ over $U-s_{\mathrm{NP}}$.

In the present section, the study variable $Y$ is assumed to be measured in both samples, $s_{P}$ and $s_{\mathrm{NP}}$. In the next section, we alleviate this assumption by considering that the study variable is not known in the nonprobability sample. This situation is the one encountered at La Poste where not all variables of interest are measured in the automatically processed postal mail. The big non-probability database is based on an image recognition process and covers around $80 \%$ of the postal mails. This database contains some relevant auxiliary information such as the departure dates from the sending post office. However, such data are subject to selection bias (e.g., mails with atypical shape are not automatically processed), and measurement errors (e.g., errors in barcode scanning during the image recognition process). In such a situation, we propose to use the intersection between the big database and the probability sample, where the auxiliary variables together with the study variable are available, and predict the unknown $y_{k}$ for $k \in s_{\mathrm{NP}}-s_{P}$.

## 3. Prediction estimators for study variable unobserved in the nonprobability sample

Recall that the finite population total of $Y$ can be decomposed as $T=T_{\mathrm{NP}}+T_{C}$. The total $T_{C}$ is estimated as in Section 2 by the Hájek-type estimator $\hat{T}_{C}^{(\mathrm{Ha})}$. In the present section, $y_{k}$ is unknown for $k \in s_{\mathrm{NP}}$, and contrarily to Section 2, the total $T_{\mathrm{NP}}$ has to be estimated. In order to do so, we introduce a working model for $Y$ and the general QR class of predictors of $T_{\mathrm{NP}}$ that does not require $y_{k}$ to be known for units in $s_{\mathrm{NP}}$. We assume that a vector of auxiliary variables $\mathbf{x}_{k}=\left(X_{k 1}, \ldots, X_{k p}\right)^{\top}$ is available for each unit $k$ of a nonprobability sample $s_{\mathrm{NP}} \subset U$. We also assume that $\delta_{k}$ and $\delta_{k} \mathbf{x}_{k}$ are available for each unit $k$ of the probability sample $s_{p}$. Table 3.1 gives a summary of the characteristics of the data we consider in the remainder of this paper.

Table 3.1
Data characteristics in the data integration context of Section 3.

| Sample | $\boldsymbol{y}_{\boldsymbol{k}}$ measured | $\boldsymbol{\delta}_{\boldsymbol{k}}$ available | known selection <br> mechanism | Auxiliary variables <br> available |
| :---: | :---: | :---: | :---: | :---: |
| $s_{P}$ | Yes | Yes | Yes | No |
| $s_{\mathrm{NP}}$ | No | Yes | No | Yes |

The variable $Y$ is not available in $s_{\mathrm{NP}}$ and we cannot use anymore $\hat{T}_{\mathrm{PDI}}$ since the total $T_{\mathrm{NP}}=\sum_{k \in U} \delta_{k} y_{k}$ is unknown. The idea behind the class of estimators introduced in this section is to predict $y_{k}$ for $k \in s_{\mathrm{NP}}$ by using regression modelling between $Y$ and the auxiliary variables, and then predict $T_{\mathrm{NP}}$. We assume the following working model between the study variable $Y$ and the vector of auxiliary variables $\mathbf{x}_{k}$ :

$$
\begin{equation*}
y_{k}=\mathbf{x}_{k}^{\top} \boldsymbol{\beta}+\varepsilon_{k}, \quad k \in s_{\mathrm{NP}}, \tag{3.1}
\end{equation*}
$$

where the errors $\varepsilon_{k}$ are independent with expectation $\mathrm{E}_{m}\left(\varepsilon_{k}\right)=0$ and variance $\operatorname{Var}_{m}\left(\varepsilon_{k}\right)$ proportional to $v\left(\mathbf{x}_{k}\right)=v_{k}$ for some known positive constants $v_{k}$. The subscript $m$ indicates that the expectation and variance are taken with respect to model (3.1) conditionally on observed auxiliary variables $\mathbf{x}_{k}, k \in s_{\mathrm{NP}}$. Note that model (3.1) only needs to hold for units in the non-probability sample. A model for $Y$ does not need to be explicitly specified for units $k \in U-s_{\mathrm{NP}}$ as we always make inferences conditional on $y_{k}$, $k \in U-s_{\mathrm{NP}}$.

We define a predictor $\hat{y}_{k}$ of $y_{k}$ for $k \in s_{\mathrm{NP}}$ by $\hat{y}_{k}=\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}$ with

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}=\left(\sum_{k \in s_{p}} q_{k} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}\right)^{-1}\left(\sum_{k \in s_{p}} q_{k} \delta_{k} \mathbf{x}_{k} y_{k}\right), \tag{3.2}
\end{equation*}
$$

where $q_{k}$ are known positive constants for $k \in s_{\mathrm{NP}}$. We assume that the $p \times p$ dimensional matrix $\sum_{k \in s_{P}} q_{k} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}$ is nonsingular for all possible samples $s_{P}$.

We propose to estimate $T_{\mathrm{NP}}=\sum_{k \in U} \delta_{k} y_{k}$ by a $Q R$ predictor as suggested in Wright (1983):

$$
\begin{align*}
\hat{T}_{\mathrm{NP}}^{(\mathrm{OR})} & =\sum_{k \in U} \delta_{k} \hat{y}_{k}+\sum_{k \in s_{p}} r_{k} \delta_{k}\left(y_{k}-\hat{y}_{k}\right) \\
& =\sum_{k \in U} \delta_{k} \mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}+\sum_{k \in s_{p}} r_{k} \delta_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right), \tag{3.3}
\end{align*}
$$

where $r_{k} \geq 0$ are predefined constants. The initials Q and R refer to the constants $q_{k}$ and $r_{k}$. The final estimator of $T$ is then given by

$$
\begin{equation*}
\hat{T}^{(\mathrm{QR})}=\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}+\hat{T}_{C}^{(\mathrm{Ha})} . \tag{3.4}
\end{equation*}
$$

Various choices of $q_{k}$ and $r_{k}$ yield predictors $\hat{\mathrm{N}}_{\mathrm{N}}^{(\mathrm{QR})}$ with familiar forms as detailed below.

1. For $q_{k}=d_{k} v_{k}^{-1}$ and $r_{k}=d_{k}$, we obtain the model-assisted or GREG-type estimator:

$$
\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{MA})}+\sum_{k \in s_{p}} \delta_{k} d_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{MA})}\right),
$$

where $\hat{y}_{k}^{(\mathrm{MA})}=\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}^{(\mathrm{MA})}$ with $\hat{\boldsymbol{\beta}}^{(\mathrm{MA})}=\left(\sum_{k \in s_{p}} d_{k} v_{k}^{-1} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}\right)^{-1}\left(\sum_{k \in s_{p}} d_{k} v_{k}^{-1} \delta_{k} \mathbf{x}_{k} y_{k}\right)$.
2. For $q_{k}=v_{k}^{-1}$ and $r_{k}=1$, we obtain the model-based type estimator:

$$
\hat{T}_{\mathrm{NP}}^{(\mathrm{MB})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{MB})}+\sum_{k \in s_{P}} \delta_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{MB})}\right),
$$


3. For $q_{k}=\left(d_{k}-1\right) v_{k}^{-1}$ and $r_{k}=1$, we obtain the cosmetic-type estimator (Särndal and Wright, 1984; Brewer, 1999):

$$
\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{Cos})}+\sum_{k \in s_{P}} \delta_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{Cos})}\right),
$$

where $\hat{y}_{k}^{(\text {(os) })}=\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}^{(\text {Cos })}$ with

$$
\hat{\boldsymbol{\beta}}^{(\mathrm{Cos})}=\left(\sum_{k \in s_{p}}\left(d_{k}-1\right) v_{k}^{-1} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}\right)^{-1}\left(\sum_{k \in s_{p}}\left(d_{k}-1\right) v_{k}^{-1} \delta_{k} \mathbf{x}_{k} y_{k}\right) .
$$

Let us derive some properties for this class of QR predictors. Proposition 3.1 gives a general condition on the constants $q_{k}$ and $r_{k}$ such that the QR predictor can be defined as a sum of predictions over the population. Proposition 3.2 gives another general condition on the constants $q_{k}$ and $r_{k}$ such that the QR predictor is a model-assisted type estimator. The proofs are given in the Appendix.

Proposition 3.1 (projection form) Consider the QR predictor $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}$ given by (3.3). Under the condition that there exists a vector $\boldsymbol{\mu} \in \mathbf{R}^{p}$ such that

$$
\begin{equation*}
\text { (Proj): } \boldsymbol{\mu}^{\top} \mathbf{x}_{k} q_{k}=r_{k} \text { for all } k \in s_{\mathrm{NP}}, \tag{3.5}
\end{equation*}
$$

we have $\sum_{k \in s_{p}} r_{k} \delta_{k}\left(y_{k}-\hat{y}_{k}\right)=0$. In this case, $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}$ can be written in the projection form:

$$
\hat{T}_{\mathrm{NP}}^{(\mathrm{OR})}=\sum_{k \in U} \delta_{k} \hat{y}_{k} .
$$

The model-assisted estimator $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$ and model-based estimator $\hat{T}_{\mathrm{NP}}^{(\mathrm{MB})}$ satisfy Condition (Proj) if there exists a vector $\boldsymbol{\mu} \in \mathbf{R}^{p}$ such that $\boldsymbol{\mu}^{\top} \mathbf{x}_{k}=v_{k}$ for all $k \in s_{\mathrm{NP}}$. This condition is satisfied when $v_{k}$ is one of the auxiliary variables in the model. If $v_{k}=1$, it is satisfied provided that the intercept is included in the model. Condition (Proj) holds for $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ if $\boldsymbol{\mu}^{\top} \mathbf{x}_{k}=v_{k}\left(d_{k}-1\right)^{-1}$ for all $k \in s_{\mathrm{NP}}$. A consequence of Proposition 3.1 is that, for equal probability sampling design such as simple random sampling without replacement, the modelassisted, the model-based and the cosmetic estimators are all equal.

Using Theorem 2 from Wright (1983), we derive the following proposition. For $r_{k}$ satisfying Condition (QR) below and any given $q_{k}$, the QR predictor of $T_{\mathrm{NP}}$ is identical to the model-assisted predictor of $T_{\mathrm{NP}}$ with the same $q_{k}$.

Proposition 3.2 Suppose that the constants $r_{k}$ and $q_{k}$ are such that there exists some vector $\lambda \in \mathbf{R}^{p}$ such that

$$
\begin{equation*}
(Q R): 1-\pi_{k} r_{k}=\pi_{k} q_{k} \mathbf{x}_{k}^{\top} \lambda \text { for all } k \in s_{\mathrm{NP}} \text {. } \tag{3.6}
\end{equation*}
$$

Then:

$$
\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}=\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)},
$$

where

$$
\begin{equation*}
\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)}=\sum_{k \in U} \delta_{k} \mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}+\sum_{k \in s_{p}} d_{k} \delta_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right) \tag{3.7}
\end{equation*}
$$

is the model-assisted type predictor of $T_{\mathrm{NP}}$ with $\hat{\boldsymbol{\beta}}$ given by (3.2).
Following Wright (1983), we note that the (QR) condition always holds for $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$. This condition also holds for the model-based estimator $\hat{T}_{\mathrm{NP}}^{(\mathrm{MB})}$ if and only if there exists a vector $\lambda \in \mathbf{R}^{p}$ such that $v_{k}\left(d_{k}-1\right)=\mathbf{x}_{k}^{\top} \lambda$, for all $k \in s_{\mathrm{NP}}$. This condition is true if we take $v_{k}\left(d_{k}-1\right)$ among the auxiliary variables $\mathbf{x}_{k}$. Condition (QR) holds for the cosmetic estimator $\hat{T}_{\mathrm{NP}}^{(\text {(cos })}$ if and only if there exists a vector $\lambda \in \mathbf{R}^{p}$ such that $v_{k}=\mathbf{x}_{k}^{\top} \lambda$, for all $k \in s_{\mathrm{NP}}$. This condition is true if $v_{k}$ is included in the vector of auxiliary variables.

## 4. Asymptotic properties and variance estimation of QR predictors

Let us consider the class of QR predictors $\hat{T}^{(\mathrm{QR})}$ given in (3.4) and start by studying the prediction error $\hat{T}^{(\mathrm{QR})}-T$ under the model (3.1) and the sampling design $p(\cdot)$. We have

$$
\hat{T}^{(\mathrm{QR})}-T=\left(\hat{T}_{\mathrm{NP}}^{(\mathrm{OR})}-T_{\mathrm{NP}}\right)+\left(\hat{T}_{C}^{(\mathrm{Ha})}-T_{C}\right) .
$$

The first right-hand term depends on the model and the sampling design, while the second right-hand term only depends on the sampling design. Assuming that the sampling design is not informative with respect to the model (3.1), we can prove that the expectation of $\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}-T_{\mathrm{NP}}$ computed with respect to the model is equal to 0 . Indeed, the model bias of $\hat{T}_{\mathrm{NP}}^{(\mathrm{(QR})}$ is given by:

$$
\mathrm{E}_{m}\left(\hat{T}_{\mathrm{NP}}^{(\mathrm{OR})}-T_{\mathrm{NP}}\right)=\sum_{k \in U} \delta_{k} \mathrm{E}_{m}\left(\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}-y_{k}\right)+\sum_{k \in s_{P}} r_{k} \delta_{k} \mathrm{E}_{m}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right),
$$

with $\hat{\boldsymbol{\beta}}=\left(\sum_{k \in s_{p}} q_{k} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}\right)^{-1}\left(\sum_{k \in s_{p}} q_{k} \delta_{k} \mathbf{x}_{k} y_{k}\right)$. Under the model (3.1), $\mathrm{E}_{m}\left(y_{k}\right)=\mathbf{x}_{k}^{\top} \boldsymbol{\beta}$ for all $k \in s_{\mathrm{NP}}$, $\mathrm{E}_{m}(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$ and $\mathrm{E}_{m}\left(\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}-y_{k}\right)=0$, implying that

$$
\begin{equation*}
\mathrm{E}_{m}\left(\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}-T_{\mathrm{NP}}\right)=0 \tag{4.1}
\end{equation*}
$$

The estimator $\hat{T}_{C}^{(\mathrm{Ha})}$ is a Hájek-type estimator of $T_{C}$ and is not design-unbiased for $T_{C}$. Following Särndal (1980), we rather look at its asymptotic design-unbiasedness. An estimator $\hat{T}$ is said to be asymptotically design-unbiased for the finite population total $T$ if $\lim _{N \rightarrow \infty} N^{-1}\left[\mathrm{E}_{p}(\hat{T})-T\right]=0$, where $\mathrm{E}_{p}(\cdot)$ is the expectation computed with respect to the sampling design. The asymptotic framework from Isaki and Fuller (1982) can be considered to allow for the population and sample sizes to grow to infinity. Assuming that the probability of observing an empty intersection set $s_{P} \cap s_{\mathrm{NP}}$ is negligible, then $\hat{T}_{C}^{(\mathrm{Ha})}$ is asymptotically design-unbiased for $T_{C}$. Combining this property with relation (4.1), and considering a non-informative sampling design, we get that $\hat{T}^{(\mathrm{QR})}$ is asymptotically $m p$-unbiased for $T$ under model (3.1).

### 4.1 Bias properties of $\hat{\boldsymbol{T}}^{(\mathrm{Q} \pi)}$

Let us consider now the QR class of predictors that satisfy the $(\mathrm{QR})$ condition given by (3.6). For this class of predictors, the final estimator of $T$ is

$$
\hat{T}^{(\mathrm{Q} \pi)}=\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)}+\hat{T}_{C}^{(\mathrm{Ha})}
$$

and the standardized total error is given by:

$$
\frac{1}{N}\left(\hat{T}^{(\mathrm{Q} \pi)}-T\right)=\frac{1}{N}\left(\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)}-T_{\mathrm{NP}}\right)+\frac{1}{N}\left(\hat{T}_{C}^{(\mathrm{Ha})}-T_{C}\right)
$$

The estimator $\hat{T}^{(\mathrm{Q} \pi)}$ is not exactly design-unbiased because of the nonlinearity of $\hat{\boldsymbol{\beta}}$ and of the Hájek estimator $\hat{T}_{C}^{(\mathrm{Ha})}$. Wright (1983) proved that the $(\mathrm{QR})$ condition given in Proposition 3.2 is a sufficient condition for $\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)}$ to be asymptotically design-unbiased for $T_{\mathrm{NP}}$, provided that

$$
\begin{equation*}
\lim _{N \rightarrow \infty} \frac{1}{N} \mathrm{E}_{p}\left[\left(\sum_{k \in U} \delta_{k} \mathbf{x}_{k}-\sum_{k \in s_{P}} d_{k} \delta_{k} \mathbf{x}_{k}\right)^{\top}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right]=0 \tag{4.2}
\end{equation*}
$$

where $\tilde{\boldsymbol{\beta}}=\left(\sum_{k \in U} \pi_{k} q_{k} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}\right)^{-1} \sum_{k \in U} \pi_{k} q_{k} \delta_{k} \mathbf{x}_{k} y_{k}$, and assuming that $\sum_{k \in U} \pi_{k} q_{k} \delta_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\top}$ is nonsingular. Following Breidt and Opsomer (2000), if the sampling fraction $n / N$ converges to a constant different from 0 , assuming mild conditions on the first and second-order inclusion probabilities of the sampling design, and on the auxiliary information vectors $\mathbf{x}_{k}$ for all $k \in s_{\mathrm{NP}}$, it can be shown that:

$$
\lim _{N \rightarrow \infty} \mathrm{E}_{p}\left\|N^{-1}\left(\sum_{k \in U} \delta_{k} \mathbf{x}_{k}-\sum_{k \in s_{P}} d_{k} \delta_{k} \mathbf{x}_{k}\right)\right\|^{2}=0
$$

where $\|\cdot\|$ is the usual Euclidian norm. Equation (4.2) follows by assuming that the regression coefficient estimator satisfies $\lim _{N \rightarrow \infty} \mathrm{E}_{p}\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|^{2}=0$ (see Cardot, Goga and Lardin (2013) for more details). The estimator $\hat{T}_{C}^{(\mathrm{Ha)}}$ is a Hájek-type estimator which can be shown to be asymptotically design-unbiased for $T_{C}$
if the probability to observe the empty set for $s_{P} \cap s_{\mathrm{NP}}$ is negligible. We conclude that the QR predictor $\hat{T}^{(\mathrm{Q} \pi)}$ is asymptotically design-unbiased for $T$.

### 4.2 Asymptotic variance and variance estimation of $\hat{\boldsymbol{T}}^{(\mathrm{Q} \pi)}$

Because the QR estimator $\hat{T}^{(\mathrm{Q} \pi)}$ is asymptotically design-unbiased, we estimate its asymptotic design variance rather than its design mean square error. We can write the standardized total error as

$$
\frac{1}{N}\left(\hat{T}^{(\mathrm{Q} \pi)}-T\right)=\frac{1}{N}\left(\sum_{k \in s_{p}} d_{k}\left(E_{k}+e_{k}\right)-\sum_{k \in U}\left(E_{k}+e_{k}\right)\right)+R_{1}+R_{2},
$$

where

$$
\begin{gathered}
E_{k}=\delta_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \tilde{\boldsymbol{\beta}}\right), \quad e_{k}=\left(1-\delta_{k}\right)\left(y_{k}-\frac{\sum_{k^{\prime} \in U}\left(1-\delta_{k^{\prime}}\right) y_{k^{\prime}}}{N-N_{\mathrm{NP}}}\right), \\
R_{1}=-\frac{1}{N}\left(\sum_{k \in s_{p}} d_{k} \delta_{k} \mathbf{x}_{k}-\sum_{k \in U} \delta_{k} \mathbf{x}_{k}\right)^{\top}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})
\end{gathered}
$$

and

$$
R_{2}=\left(\frac{1}{N} \sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right)\right)^{-1} \frac{1}{N}\left(N-N_{\mathrm{NP}}-\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right)\right) \frac{1}{N}\left(\sum_{k \in s_{p}} d_{k} e_{k}-\sum_{k \in U} e_{k}\right) .
$$

As in Section 4.1, we assume the usual conditions on the sampling fraction $n / N$, on the first and secondorder inclusion probabilities, on the variable of interest $y_{k}$ and on the auxiliary information vector $\mathbf{x}_{k}$. We use the Landau notations big $O_{p}$ and little $o_{p}$. If $\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|^{2}=o_{p}(1)$ and $\left(\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right) / N\right)^{-1}=O_{p}(1)$, then $R_{1}=o_{p}\left(n^{-1 / 2}\right), R_{2}=O_{p}\left(n^{-1}\right)$ and the standardized total error can be approximated by

$$
\frac{1}{N}\left(\hat{T}^{(\mathrm{Q} \pi)}-T\right) \approx \frac{1}{N}\left(\sum_{k \in s_{p}} d_{k}\left(E_{k}+e_{k}\right)-\sum_{k \in U}\left(E_{k}+e_{k}\right)\right),
$$

where the right-hand side of the above expression is of order $O_{p}\left(n^{-1 / 2}\right)$. Since $\sum_{k \in s_{p}} d_{k}\left(E_{k}+e_{k}\right)$ is the Horvitz-Thompson estimator of the total $\sum_{k \in U}\left(E_{k}+e_{k}\right)$, the asymptotic variance of the QR estimator $\hat{T}^{(\mathrm{Q} \pi)}$ is given by:

$$
\operatorname{AVar}\left(\hat{T}^{(Q \pi)}\right)=\sum_{k \in U} \sum_{l \in U} \Delta_{k l} d_{k} d_{l}\left(E_{k}+e_{k}\right)\left(E_{l}+e_{l}\right) .
$$

Assuming that $\pi_{k l}>0$ for all $k, l \in U$, an estimator of the asymptotic variance is given by

$$
\hat{V}\left(\hat{T}^{(Q \pi)}\right)=\sum_{k \in s_{p}} \sum_{l \in s_{p}} \frac{\Delta_{k l}}{\pi_{k l}} d_{k} d_{l}\left(\hat{E}_{k}+\hat{e}_{k}\right)\left(\hat{E}_{l}+\hat{e}_{l}\right),
$$

where $\hat{E}_{k}=\delta_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right)$ and $\hat{e}_{k}=\left(1-\delta_{k}\right)\left(y_{k}-\sum_{k^{\prime} \in s_{p}} d_{k^{\prime}}\left(1-\delta_{k^{\prime}}\right) y_{k^{\prime}} /\left(N-N_{\mathrm{NP}}\right)\right)$. We can show, under assumptions detailed in Breidt and Opsomer (2000) and Goga, Deville and Ruiz-Gazen (2009), that this estimator is design-consistent for $\mathrm{A} \operatorname{Var}\left(\hat{T}^{(\mathrm{Q} \pi)}\right)$ in the sense that $N^{-2} n\left(\hat{V}\left(\hat{T}^{(\mathrm{Q} \pi)}\right)-\mathrm{A} \operatorname{Var}\left(\hat{T}^{(\mathrm{Q} \pi)}\right)\right)=o_{p}(1)$.

## 5. Simulations

In this section, we show the results of a Monte-Carlo study that compares the efficiency of three special cases of the QR predictor, $\hat{T}^{(\mathrm{QR})}=\hat{T}_{\mathrm{NP}}^{(\mathrm{QR})}+\hat{T}_{C}^{(\mathrm{Ha})}$, given in Section 3, namely the model-assisted, the modelbased and the cosmetic estimators, assuming that $v_{k}=1 \mathrm{in}$ model (3.1). We also compare these estimators with the expansion estimator and the PDI estimator defined in Section 2. To illustrate that the relative superiority of estimators depends on the data structure, we define three different setups based on different artificial populations. As mentioned in Section 3, if the probability samples are drawn using simple random sampling without replacement, the three QR estimators are all equal. Therefore, we focus on Poisson sampling with inclusion probabilities proportional to an auxiliary variable.

### 5.1 Populations and setups

The variables are generated using Gamma distributions to ensure their positiveness. Similar simulation results were obtained with Gaussian distributions but are not reported below. All populations have a size $N=1,000$. We generate two auxiliary variables $X_{1}$ and $X_{2}$, where $X_{1}$ (resp. $X_{2}$ ) follows a Gamma distribution with mean 20 (resp. 30) and standard deviation (Std) 15 (resp. 20). We use different models to generate the variable $Y$ for all population units. For each model, $Y \mid X_{1}, X_{2}$ follows a Gamma distribution with constant variance $\sigma_{Y \mid X_{1}, X_{2}}^{2}$ and mean $\mu_{Y \mid X_{1}, X_{2}}$ which depends on the model.

1. For Model 1, $\mu_{Y \mid X_{1}, X_{2}}$ is a linear function of $X_{1}$ and $X_{2}$ :

$$
\mu_{Y \mid X_{1}, X_{2}}=a_{0}+a_{1} X_{1}+a_{2} X_{2} .
$$

2. For Model 2, $\mu_{Y \mid X_{1}, X_{2}}$ is a quadratic function of $X_{1}$ and a linear function of $X_{2}$ :

$$
\mu_{Y \mid X_{1}, X_{2}}=b_{0}+b_{1}\left(X_{1}-\bar{X}_{1}\right)^{2}+b_{2} X_{2} \text { with } \bar{X}_{1} \text { the mean of } X_{1} \text { over } U \text {. }
$$

3. For Model 3, $\mu_{Y \mid X_{1}, X_{2}}$ is a linear function of $X_{2}$ :

$$
\mu_{Y \mid X_{1}, X_{2}}=c_{0}+c_{2} X_{2} .
$$

To make the results comparable between the three models, we determine the constants $a_{0}, a_{1}, a_{2}, b_{0}$, $b_{1}, b_{2}, c_{0}, c_{2}$, and $\sigma_{Y \mid X_{1}, X_{2}}^{2}$ in such a way that the following characteristics are the same:

- the unconditional mean $\mu$ and variance $\sigma^{2}$ of the variable $Y$,
- the coefficient of determination of the model, denoted as $R^{2}$,
- the ratio of variances for the explanatory variables:

$$
\gamma=\operatorname{Var}\left(a_{1} X_{1}\right) / \operatorname{Var}\left(a_{2} X_{2}\right)=\operatorname{Var}\left(b_{1}\left(X_{1}-\bar{X}_{1}\right)^{2}\right) / \operatorname{Var}\left(b_{2} X_{2}\right)
$$

This ratio is only relevant for models 1 and 2 since $X_{1}$ is not included in Model 3 .
In the following, we set $\mu=100, \sigma^{2}=100$, and $\gamma=0.5$. In Section 5.2 , the $R^{2}$ value is either fixed to 0.8 or varies between 0.1 and 0.96 . The main characteristics of the three population models are summarized in Table 5.1.

Table 5.1
Population models with $\mu=100, \sigma^{2}=100$, and $\gamma=0.5$.

| Model | Mean of $\left(\boldsymbol{X}_{\mathbf{1}}, \boldsymbol{X}_{\mathbf{2}}\right)$ | Std of $\left(\boldsymbol{X}_{\mathbf{1}}, \boldsymbol{X}_{\mathbf{2}}\right)$ | Mean of $\boldsymbol{Y} \mid \boldsymbol{X}_{\mathbf{1}}, \boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{R}^{\mathbf{2}}$ |
| :---: | :---: | :---: | :--- | :--- |
| 1 | $(20,30)$ |  | $\mu_{Y}=a_{0}+a_{1} X_{1}+a_{2} X_{2}$ | equal |
| 2 |  |  | $\mu_{Y}=b_{0}+b_{1}\left(X_{1}-\bar{X}_{1}\right)^{2}+b_{2} X_{2}$ | between |
| 3 |  | $\mu_{Y}=c_{0}+c_{2} X_{2}$ | populations |  |

A non-probability sample of size 900 is drawn using simple random sampling without replacement and is the same for all populations. The probability samples are drawn using Poisson sampling with expected size 200 or 50 and probabilities proportional to $X_{1}$. We consider three setups. In each setup, we generate $Y \mid X_{1}, X_{2}$ using one of the three different population models, and we compute $\hat{y}_{k}, k \in s_{\mathrm{NP}}$ for different QR predictors. The variables used as explanatory variables in the prediction models differ between setups as follows:

1. Setup 1: Informative case. Population Model 1 is used to generate population $Y$ values and only $X_{2}$ is used as explanatory variable in the prediction model along with the intercept.
2. Setup 2: Quadratic case. Population Model 2 is used to generate population $Y$ values and both auxiliary variables $X_{1}$ and $X_{2}$ are used as explanatory variables in the prediction model along with the intercept.
3. Setup 3: Non-informative case. Population Model 3 is used to generate population $Y$ values and only $X_{2}$ is used as explanatory variable in the prediction model along with the intercept.

For the informative and quadratic setups, the prediction model differs from the population model for $Y$, while the correct model is used in the non-informative setup. Table 5.2 gives a summary of the three setups.

Table 5.2
Studied setups.

| Setup | Population | Variables used in prediction | Model correctly specified |
| :--- | :--- | :--- | :---: |
| Informative | $\mu_{Y}=a_{0}+a_{1} X_{1}+a_{2} X_{2}$ | $\mathbf{x}_{k}^{\top}=\left(1, x_{2 k}\right)$ | No |
| Quadratic | $\mu_{Y}=b_{0}+b_{1}\left(X_{1}-\bar{X}_{1}\right)^{2}+b_{2} X_{2}$ | $\mathbf{x}_{k}^{\top}=\left(1, x_{1 k}, x_{2 k}\right)$ | No |
| Non-informative | $\mu_{Y}=c_{0}+c_{2} X_{2}$ | $\mathbf{x}_{k}^{\top}=\left(1, x_{2 k}\right)$ | Yes |

### 5.2 Results

Let us consider the three setups defined above and compare the following estimators:

- $\hat{T}_{\mathrm{HT}}=\sum_{k \in S_{p}} d_{k} y_{k}$,
- $\hat{T}_{\mathrm{PDI}}=T_{\mathrm{NP}}+\left(N-N_{\mathrm{NP}}\right) \frac{\sum_{k \in s_{P}} d_{k}\left(1-\delta_{k}\right) y_{k}}{\sum_{k \in s_{P}} d_{k}\left(1-\delta_{k}\right)}$,
- $\hat{T}^{(\mathrm{MB})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{MB})}+\sum_{k \in s_{p}} \delta_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{MB})}\right)+\left(N-N_{\mathrm{NP}}\right) \frac{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right) y_{k}}{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right)}$,
- $\hat{T}^{(\mathrm{MA})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{MA})}+\sum_{k \in s_{p}} \delta_{k} d_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{MA})}\right)+\left(N-N_{\mathrm{NP}}\right) \frac{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right) y_{k}}{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right)}$,
- $\hat{T}^{(\mathrm{Cos})}=\sum_{k \in U} \delta_{k} \hat{y}_{k}^{(\mathrm{Cos})}+\sum_{k \in s_{p}} \delta_{k}\left(y_{k}-\hat{y}_{k}^{(\mathrm{Cos})}\right)+\left(N-N_{\mathrm{NP}}\right) \frac{\sum_{k \in s_{p}} d_{k}\left(1-\delta_{k}\right) y_{k}}{\sum_{k \in s_{P}} d_{k}\left(1-\delta_{k}\right)}$.

For each setup, $L=10,000$ probability samples $s_{P}$ are drawn according to Poisson sampling as detailed above and several Monte Carlo measures are computed. We compute the Monte Carlo relative bias of a given estimator $\hat{T}$ (either $\hat{T}_{\mathrm{HT}}, \hat{T}^{(\mathrm{MB})}, \hat{T}^{(\mathrm{MA})}, \hat{T}^{(\text {Cos })}$ or $\hat{T}_{\mathrm{PDI}}$ ) as

$$
\operatorname{RB}_{\mathrm{MC}}(\hat{T})=100 \times L^{-1} \sum_{l=1}^{L} \frac{\hat{T}^{(l)}-T}{T},
$$

where $\hat{T}^{(l)}$ is an estimate of $T$ computed for the $l^{\text {th }}$ sample, $l=1, \ldots, L$.
As a measure of efficiency, we compute the Monte Carlo relative mean square error (RMSE) of an estimator $\hat{T}$ (relative to $\hat{T}^{(\text {Cos) })}$ ):

$$
\operatorname{RMSE}_{\mathrm{MC}}(\hat{T})=100 \times \frac{\operatorname{MSE}_{\mathrm{MC}}(\hat{T})}{\operatorname{MSE}_{\mathrm{MC}}\left(\hat{T}^{(\text {Cos })}\right)},
$$

where

$$
\operatorname{MSE}_{\mathrm{MC}}(\hat{T})=L^{-1} \sum_{l=1}^{L}\left(\hat{T}^{(l)}-T\right)^{2}
$$

We also compute the Monte Carlo relative variance (RVar) of an estimator $\hat{T}$ (relative to $\hat{T}^{(\operatorname{Cos})}$ ):

$$
\operatorname{RVar}_{\mathrm{MC}}(\hat{T})=100 \times \frac{\operatorname{Var}_{\mathrm{MC}}(\hat{T})}{\operatorname{Var}_{\mathrm{MC}}\left(\hat{T}^{(\operatorname{Cos})}\right)}
$$

where

$$
\operatorname{Var}_{\mathrm{MC}}(\hat{T})=L^{-1} \sum_{l=1}^{L}\left(\hat{T}^{(l)}\right)^{2}-\left(L^{-1} \sum_{l=1}^{L} \hat{T}^{(l)}\right)^{2}
$$

Table 5.3 contains the simulation results for the three setups when $R^{2}=0.8$. In all setups, we confirm that both $\hat{T}_{\mathrm{PDI}}$ and $\hat{T}_{\mathrm{HT}}$ have a small Monte Carlo bias, as expected. In terms of MSE, $\hat{T}_{\mathrm{PDI}}$ is the most precise estimator, while $\hat{T}_{\mathrm{HT}}$ is the least precise estimator among all estimators. This result is expected since the expansion estimator does not make use of any auxiliary information, while $\hat{T}_{\text {PDI }}$ takes into account the true values of the study variable $y_{k}$ for $k \in s_{\mathrm{NP}}$; i.e., it takes into account the true values of $Y$ for 900 units out of the 1,000 population units. In our context, where the study variable is not observed in $s_{\mathrm{NP}}$, the estimator $\hat{T}_{\text {PDI }}$ is however not computable and serves more as a gold standard.

Table 5.3
Relative bias (in \% of the true value), percent relative (to $\hat{\boldsymbol{T}}^{(\mathrm{Cos})}$ ) variance and MSE of the different estimators for the three different setups; the expected size of the probability sample is 200 and the size of the nonprobability sample is 900 .

| Population parameters | Setup | Monte Carlo measures | $\hat{T}_{\mathrm{HT}}$ | $\hat{\boldsymbol{T}}^{(\mathrm{MB})}$ | $\hat{\boldsymbol{T}}^{(\mathrm{MA})}$ | $\hat{\boldsymbol{T}}^{(\mathrm{Cos})}$ | $\hat{T}_{\text {PDI }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\begin{aligned} & \mu=100 \\ & \sigma^{2}=100 \\ & R^{2}=0.8 \\ & \gamma=0.5 \end{aligned}$ | Setup 1 | $\begin{aligned} & \mathrm{RB} \mathrm{mc} \\ & \mathrm{RVar}_{\mathrm{Mc}} \\ & \mathrm{RMSE}_{\mathrm{Mc}} \end{aligned}$ | $\begin{gathered} -0.13 \\ 23,566.93 \\ 22,897.58 \\ \hline \end{gathered}$ | $\begin{gathered} \hline 3.34 \\ 55.62 \\ 2,715.21 \\ \hline \end{gathered}$ | $\begin{gathered} \hline 0.11 \\ 114.06 \\ 113.91 \end{gathered}$ | $\begin{gathered} \hline 0.11 \\ 100.00 \\ 100.00 \end{gathered}$ | $\begin{gathered} 0.03 \\ 20.97 \\ 20.65 \end{gathered}$ |
|  | Setup 2 | $\begin{aligned} & \mathrm{RB}_{\mathrm{mc}} \\ & \mathrm{RVar}_{\mathrm{Mc}} \\ & \mathrm{RMSE}_{\mathrm{Mc}} \end{aligned}$ | $\begin{gathered} \hline-0.07 \\ 36,947.99 \\ 36,638.27 \end{gathered}$ | $\begin{gathered} \hline-1.65 \\ 84.94 \\ 1,056.44 \end{gathered}$ | $\begin{gathered} -0.06 \\ 118.21 \\ 118.42 \end{gathered}$ | $\begin{gathered} -0.05 \\ 100.00 \\ 100.00 \end{gathered}$ | $\begin{gathered} 0.02 \\ 23.17 \\ 23.15 \end{gathered}$ |
|  | Setup 3 | $\begin{aligned} & \mathrm{RB}_{\mathrm{Mc}} \\ & \mathrm{RVar}_{\mathrm{Mc}} \\ & \mathrm{RMSE}_{\mathrm{Mc}} \end{aligned}$ | $\begin{gathered} 0.03 \\ 41,088.93 \\ 41,080.51 \end{gathered}$ | $\begin{gathered} -0.01 \\ 58.38 \\ 58.39 \end{gathered}$ | $\begin{gathered} 0.01 \\ 100.49 \\ 100.48 \end{gathered}$ | $\begin{gathered} 0.01 \\ 100.00 \\ 100.00 \end{gathered}$ | $\begin{gathered} 0.01 \\ 33.47 \\ 33.51 \end{gathered}$ |

Notes: $\mathrm{RB}=$ relative bias; $\mathrm{RVar}=$ relative variance; $\mathrm{RMSE}=$ relative mean square error; $\mathrm{MC}=$ Monte Carlo.

The Monte Carlo bias of $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ is negligible in the three setups whereas $\hat{T}^{(\mathrm{MB})}$ is biased in the informative and quadratic setups. In these two setups, the prediction model differs from the population model used to generate $Y$ values. In the non-informative setup, where the prediction model is correctly specified, the bias of $\hat{T}^{(\mathrm{MB})}$ is also negligible. The estimator $\hat{T}^{(\mathrm{MA})}$ has the largest variance of the QR predictors in the informative and quadratic setups, while $\hat{T}^{(\mathrm{MB})}$ has the smallest variance in all setups. In the quadratic setup, the variance of $\hat{T}^{(\mathrm{MB})}$ is similar to the variance of $\hat{T}^{(\mathrm{Cos})}$ but $\hat{T}^{(\mathrm{MB})}$ has the highest MSE amongst the QR predictors in both informative and quadratic setups. This means that the bias of $\hat{T}^{(\mathrm{MB})}$ degrades its MSE a lot despite its small variance. In the non-informative setup, $\hat{T}^{(\mathrm{MB})}$ has the lowest MSE amongst the QR predictors. We can see in Table 5.3 that this comes from the absence of bias for $\hat{T}^{(\mathrm{MB})}$ in this setup together with its small variance.

In the informative and quadratic setups, $\hat{T}^{(\operatorname{Cos})}$ is more precise in term of variance than $\hat{T}^{(\mathrm{MA})}$. The estimators $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ are similar in the non-informative setup. Both estimators use weighted regression with slightly different weights $\left(d_{k}\right.$ for $\hat{T}^{(\mathrm{MA})}$ and $d_{k}-1$ for $\hat{T}^{(\mathrm{Cos})}$ ).

To summarize, when the prediction model is incorrectly specified, as in the informative and quadratic setups, both $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ are significantly more efficient than $\hat{T}^{(\mathrm{MB})}$ because of the bias of $\hat{T}^{(\mathrm{MB})}$, even though the bias is not large. On the opposite, if the model is correctly specified but the design weights and $Y$ are uncorrelated, as in the non-informative setup, $\hat{T}^{(\mathrm{MB})}$ is better than $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ in terms of MSE. In all setups, $\hat{T}^{(\mathrm{Cos})}$ is more efficient or similar to $\hat{T}^{(\mathrm{MA})}$.

To better understand the impact of the $R^{2}$ on the results, we also plot, on the $y$-axis of Figures 5.1, 5.2 and 5.3, the RMSE $_{\mathrm{Mc}}$ for 10 different values of $R^{2}$ on the $x$-axis: $0.1,0.2, \ldots, 0.9,0.96$. In order to do that, we generate for each setup ten populations, one for each $R^{2}$ value. Figure 5.1 (resp. Figure 5.2 and 5.3) gives the results for Setup 1 (resp. 2 and 3) with the sample size equal to 200 (resp. 50) on the left (resp. right) column plots. On all plots, the curves correspond to the different estimators with a red curve at 100 for $\hat{T}^{(\text {Cos })}$ (since the RMSE are relative to $\hat{T}^{(\mathrm{Cos})}$ ) and different colors for $\hat{T}_{\mathrm{HT}}, \hat{T}^{(\mathrm{MA})}, \hat{T}^{(\mathrm{MB})}$ and $\hat{T}_{\mathrm{PDI}}$. The plots on the top row of the figures include all the estimators while for the second row (and third row for Figures 5.1 and 5.2), $\hat{T}_{\mathrm{HT}}$ (and $\hat{T}^{(\mathrm{MB})}$ for Figures 5.1 and 5.2) is removed in order to zoom in and ease the comparison between $\hat{T}^{(\mathrm{Cos})}, \hat{T}^{(\mathrm{MA})}, \hat{T}^{(\mathrm{MB})}$ and $\hat{T}_{\mathrm{PDI}}$. The scale on the $y$-axis is kept fixed for the two columns (sample sizes).

As expected, $\hat{T}_{\text {PDI }}$ is by far the best estimator with the smallest MSE in all setups. In all figures, $\hat{T}_{\mathrm{HT}}$ has a very bad relative MSE, especially when $R^{2}$ is high. Note that in fact the absolute MSE of $\hat{T}_{\mathrm{HT}}$ remains stable when $R^{2}$ increases (results not reported), while the MSE of the other estimators improves. This result is expected because $\hat{T}_{\mathrm{HT}}$ does not depend on the distribution of $Y \mid X_{1}, X_{2}$, but depends on $\mu$ and $\sigma^{2}$ which are constant across the populations.

Figure 5.1 (resp. 2) shows the evolution of $\mathrm{RMSE}_{\mathrm{MC}}$ with respect to $R^{2}$ in the informative setup (resp. quadratic setup) for sample $s_{P}$ of expected size 200 (left column) and 50 (right column). In these two setups, not only is $\hat{T}^{(\mathrm{Cos})}$ better than $\hat{T}^{(\mathrm{MB})}$ or $\hat{T}^{(\mathrm{MA})}$, as seen in Table 5.3 , but its gain compared to its competitors increases the most with $R^{2}$. The precision of $\hat{T}^{(\mathrm{MA})}$ also increases, but at a slightly slower pace. The MSE of $\hat{T}^{(\mathrm{MB})}$ worsens with $R^{2}$ because the prediction model differs too much from the population model in these setups. This fact implies a larger bias of $\hat{T}^{(\mathrm{MB})}$ when $R^{2}$ increases. For informative and quadratic setups, a smaller size reduces the difference between $\mathrm{RMSE}_{\mathrm{MC}}(\hat{T})$ of QR predictors.

Figure 5.3 shows the evolution of the RMSE $_{M C}$ with respect to the $R^{2}$ in the non-informative setup. This time, $\hat{T}^{(\mathrm{MB})}$ does not lose precision when $R^{2}$ increases because the prediction model is the same as the population model. All QR predictors show an increase in precision with $R^{2}$, with $\hat{T}^{(\mathrm{Cos})}$ and $\hat{T}^{(\mathrm{MA})}$ having similar precision for all values of $R^{2}$. In this setup, the plots are comparable for the two sample sizes, because the model is correctly specified for all prediction models.

Figure 5.1 Relative MSE (in \%), with the MSE of the cosmetic estimator as the baseline, versus $\boldsymbol{R}^{2}$ in the informative setup.


Note: RMSE = relative mean square error.

Figure 5.2 Relative MSE (in \%), with the MSE of the cosmetic estimator as the baseline, versus $\boldsymbol{R}^{2}$ in the quadratic setup.


Note: RMSE = relative mean square error.

Figure 5.3 Relative MSE (in \%), with the MSE of the cosmetic estimator as the baseline, versus $\boldsymbol{R}^{\mathbf{2}}$ in the noninformative setup.


Note: $\mathrm{RMSE}=$ relative mean square error.

To sum up, if the prediction model is misspecified, the cosmetic estimator is the best choice in our setups. It has the smallest MSE amongst all QR predictors, and its precision increases faster with $R^{2}$ than the other estimators. The advantage of $\hat{T}^{(\mathrm{Cos})}$ over $\hat{T}^{(\mathrm{MA})}$ might disappear in a scenario where the probability sample size would be a smaller fraction of the population size. The estimator $\hat{T}^{(\mathrm{MB})}$ is biased and has the largest MSE, even for smaller values of $R^{2}$. If the model is correctly specified, and $Y$ is not correlated to $X_{1}$ while the first-order inclusion probabilities are proportional to $X_{1}, \hat{T}^{(\mathrm{MB})}$ is the best choice in terms of MSE. However, the efficiency gain achieved by choosing $\hat{T}^{(\mathrm{MB})}$ over $\hat{T}^{(\text {(Cos })}$ in this third setup is significantly smaller than the efficiency loss observed when choosing $\hat{T}^{(\mathrm{MB})}$ over $\hat{T}^{(\text {Cos })}$ in the first two setups. We thus recommend the choice of the Cosmetic estimator as a good compromise in all setups, followed closely by the model-assisted estimator. Similar observations can be made from real data, as shown in the next section.

## 6. Application to La Poste data

### 6.1 Data presentation

In France, more than $90 \%$ of letters sent are sorted out using automatic sorting machines. The information collected by the machines is based on pictures of the letters and form the non-probability database. La Poste has also access to a probability sample and wants to use both databases to estimate monthly totals of different types of letters using the data integration methods proposed above. Some letters such as "lettres vertes" (letters with a green stamp and ecologically transported) are not recognized by the sorting machines, which only take black and white pictures, meaning that the variable of interest is not available in the nonprobability database. The auxiliary information associated with the letters automatically sorted is not easily linked to the probability sample. This issue is currently being investigated at La Poste. Thus, the illustration below is based on data from previous surveys made over the years at La Poste.

Data from these surveys are collected from postmen rounds and contain in particular the number of lettres vertes, the number of letters named "produits $1 b$ " (which include different types of letters, including lettres vertes) and the total number of letters in the round. The idea is to mimic the situation at La Poste where the number of lettres vertes is available in the probability sample but not on the non-probability database whereas the number of produits $l b$ and the total number of letters are available in the nonprobability database but not in the probability sample.

The goal of this section is on one hand to see if the conclusion drawn with simulated populations holds with real data and on the other hand to see if the selection method for the non-probability sample has an impact on the estimators. The population of interest consists in 11,906 rounds from historical data. In the following, we assume that the variable of interest is the number of lettres vertes and the explanatory variable $X_{1}$ (resp. $X_{2}$ ) is the total number of letters (resp. the number of produits 1 lb ).

We consider three configurations. For each configuration, a non-probability sample of size 9,524 is drawn from the population; the non-probability sampling fraction is thus $80 \%$. In the first configuration, the non-probability sample is drawn using SRSWOR. For the second (resp. third) configuration, the nonprobability sample contains the 9,524 rounds with highest (resp. lowest) values of $Y$. Only $X_{2}$ is used as explanatory variable in the prediction model along with the intercept.

We compare the same estimators as in Section 5.2. For each of the 3 non-probability samples, we draw $L=1,000$ probability samples $s_{P}$ of expected size 2,000 using Poisson sampling with probabilities proportional to $X_{1}$. The same Monte-Carlo measures as those in Section 5.2 are computed.

### 6.2 Results

Table 6.1 contains the simulation results for La Poste data for the three non-probability samples. For all configurations, the prediction model appears to be significantly misspecified, which causes $\hat{T}^{(\mathrm{MB})}$ to be less efficient, both in terms of variance and MSE, than in the simulations of Section 5.2. Different diagnostic
statistics could be computed beforehand to highlight the misspecification problem and alternative prediction models could be proposed. The assessment of such models is not in the scope of this paper.

Table 6.1
Relative bias (in \% of the true value), percent relative (to $\hat{\boldsymbol{T}}^{(\mathrm{Cos})}$ ) variance and MSE of the different estimators for the three different non-probability samples; the expected size of the probability sample is 2,000 and the size of the non-probability sample is $\mathbf{9 , 5 2 4}$.

| $S_{N P}$ | Monte Carlo measures | $\hat{T}_{\text {HT }}$ | $\hat{\boldsymbol{T}}^{(\text {(MB) }}$ | $\hat{\boldsymbol{T}}^{(\mathrm{MA})}$ | $\hat{\boldsymbol{T}}^{(\mathrm{Cos})}$ | $\hat{T}_{\text {PDI }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| SRSWOR | $\mathrm{RB}_{\text {мс }}$ | 0.08 | 8.30 | 0.08 | 0.08 | 0.06 |
|  | $\mathrm{RVar}_{\text {mc }}$ | 253.67 | 122.66 | 102.21 | 100.00 | 87.41 |
|  | $\mathrm{RMSE}_{\text {мс }}$ | 252.95 | 5,489.24 | 102.18 | 100.00 | 87.31 |
| Highest $Y$ values | $\mathrm{RB}_{\text {мс }}$ | 0.03 | 5.51 | 0.03 | 0.03 | 0.01 |
|  | $\mathrm{RVar}_{\text {Mc }}$ | 1,528.57 | 197.54 | 101.82 | 100.00 | 3.78 |
|  | $\mathrm{RMSE}_{\text {мс }}$ | 1,528.57 | 13,676.65 | 101.87 | 100.00 | 3.82 |
| Lowest $Y$ values | $\mathrm{RB}_{\text {мс }}$ | -0.02 | 1.23 | 0.13 | 0.13 | 0.41 |
|  | $\mathrm{RVar}_{\text {Mc }}$ | 197.09 | 100.67 | 100.26 | 100.00 | 93.17 |
|  | $\mathrm{RMSE}_{\text {мс }}$ | 195.36 | 184.83 | 100.24 | 100.00 | 92.95 |

Notes: $\mathrm{RB}=$ relative bias; $\mathrm{RVar}=$ relative variance; $\mathrm{RMSE}=$ relative mean square error; $\mathrm{MC}=$ Monte Carlo; SRSWOR $=$ simple random sampling without replacement.

For all configurations, the results are similar to those obtained in the informative and quadratic setups of Section 5.2 with $\hat{T}^{(\mathrm{MA})}$ and $\hat{T}^{(\mathrm{Cos})}$ having similar efficiency and being both more efficient than $\hat{T}^{(\mathrm{MB})}$, which is the only biased estimator. It can be noted that, although the choice of $s_{\mathrm{NP}}$ does not impact the relative efficiency of $\hat{T}^{(\mathrm{MA})}$, it impacts the relative precision of the HT and PDI estimators.

In all configurations, $s_{P}$ is drawn using Poisson sampling and the variance of the data integration estimators can be simplified as follows:

$$
\operatorname{Var}(\hat{T})=\operatorname{Var}\left(\hat{T}_{\mathrm{NP}}\right)+\operatorname{Var}\left(\hat{T}_{C}^{(\mathrm{Ha})}\right)
$$

with $\hat{T}_{\mathrm{NP}}$ the predictor of the total $T_{\mathrm{NP}}$ and $\hat{T}_{C}^{(\mathrm{Ha)})}$ the Hájek estimator of the total $T_{C}$. To further understand the impact of the selection of $s_{\mathrm{NP}}$ on the estimators, we study the variance of $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}, \hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ and $\hat{T}_{C}^{(\mathrm{Haz})}$, and the variance of the Hájek estimator $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}=N_{\mathrm{NP}} \sum_{k \in s_{p}} d_{k} \delta_{k} y_{k} / \sum_{k \in s_{p}} d_{k} \delta_{k}$ of the total $T_{\mathrm{NP}}$.

Table 6.2 contains the relative variance of $\hat{T}_{\mathrm{HT}}, \hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}, \hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}, \hat{T}_{C}^{(\mathrm{Ha})}$ and $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}$ for the second and third configurations, when $s_{\mathrm{NP}}$ contains the highest or lowest values of $Y$, relative to their variance in the first configuration (SRSWOR). As expected, the precision of the HT estimator does not depend on $s_{\mathrm{NP}}$ and is fixed for the three configurations. In the second configuration, $\hat{T}_{C}^{(\mathrm{Ha)}}$ has a much smaller variance than in the other configurations, since only the smallest values of $Y$ are left in $U-s_{\mathrm{NP}}$. Similarly, its variance is greater in the third configuration, when only the largest values of $Y$ are left in $U-s_{\mathrm{NP}}$. A similar reasoning explains the ratio of variances between configurations for $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}, \hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ and $\hat{T}_{\mathrm{NP}}^{(\mathrm{Ha})}$. The important point is that both $\hat{T}_{\mathrm{NP}}^{(\mathrm{MA})}$ and $\hat{T}_{\mathrm{NP}}^{(\mathrm{Cos})}$ are not very sensitive to $s_{\mathrm{NP}}$ in terms of variance.

Table 6.2
Percent relative variance of the estimators when $s_{\mathrm{NP}}$ contains the highest or lowest $\boldsymbol{Y}$ values, relative to their variance when $s_{\mathrm{NP}}$ is drawn with SRSWOR.

| $\boldsymbol{S}_{\mathrm{NP}}$ | $\hat{\boldsymbol{T}}_{\mathrm{HT}}$ | $\hat{\boldsymbol{T}}_{\mathrm{NP}}^{(\mathrm{NA})}$ | $\hat{\boldsymbol{T}}_{\mathrm{NP}}^{\text {(Cos) }}$ | $\hat{\boldsymbol{T}}_{\mathrm{NP}}^{(\text {(Ha) }}$ | $\hat{\boldsymbol{T}}_{C}^{(\mathrm{Ha)}}$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Highest $Y$ values | 101.97 | 111.39 | 110.38 | 101.61 | 0.92 |
| Lowest $Y$ values | 104.61 | 73.07 | 69.03 | 15.43 | 181.29 |

Note: SRSWOR = simple random sampling without replacement.

We could argue that the non-probability samples selected in the second and third configurations are not realistic. Yet, the results observed for more realistic non-probability samples, selected using Poisson sampling proportional to $X_{1}$ or $Y$, were not significantly different from the results observed when $s_{\mathrm{NP}}$ is drawn using SRSWOR and are not reported. Similarly, using both auxiliary variables in the prediction model only impacts the bias of $\hat{T}^{(\mathrm{MB})}$, and the results are again not reported.

To further improve the precision, other estimators of $T_{C}$ may be used. As mentioned before, the data integration methods proposed in this paper can be used when the overlap between the probability and the non-probability samples is non-empty and, ideally, large. In this context, we recommend, based on our empirical results, the choice of the cosmetic or the model-assisted estimators.

## 7. Conclusion

Most of the literature on data integration in finite population tackles the problem of unobserved study variable in the probability sample. In this paper, we have proposed to fill the gap and considered the problem of unobserved study variable in the non-probability sample, assuming that it is observed in the probability sample and that auxiliary information is available in both samples. We have defined a general class of prediction estimators, based on the already known QR class, which includes the model-assisted, modelbased and cosmetic estimators, and studied theoretically their bias and variance properties. We have also derived a variance estimator and compared the three types of estimators with the usual Horvitz-Thompson estimator in different simulation setups, both in terms of bias and MSE, and concluded that the cosmetic estimator is a good compromise in general.

The main conclusion of our experiments is that significant efficiency gains can be achieved by leveraging a big non-probability database that contains auxiliary information associated with the main study variables. For large domains, the efficiency gains obtained from using model-assisted estimators, including the cosmetic estimator, may be sufficient to obtain high-quality estimates of the population parameters of interest. For smaller domains, these estimators may not achieve precision targets. However, they could be used as direct estimates in a small area estimation model, such as the well-known Fay-Herriot area level model. This model requires area level auxiliary information. The big non-probability database would be a natural candidate for providing the auxiliary information required for producing small area estimates. Small
area estimation methods often yield significant precision gains over direct estimators at the expense of introducing model assumptions.

## Acknowledgements

We would like to thank the Associate Editor and two reviewers for their constructive suggestions and comments that helped us to significantly improve our paper. This work has been partly supported by the French Agence Nationale de la Recherche through CIFRE contract 2019/1966 and through the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

## Appendix

## Proof of Proposition 2.1

We recall that $\hat{T}_{\mathrm{DI}}=\sum_{k \in U} \delta_{k} y_{k}+\sum_{k \in s_{p}}\left(1-\delta_{k}\right) d_{k} y_{k} \quad$ and $\quad \hat{T}_{\mathrm{HT}}=\sum_{k \in s_{p}} d_{k} y_{k}=\sum_{k \in s_{p}} \delta_{k} d_{k} y_{k}+$ $\sum_{k \in s_{p}}\left(1-\delta_{k}\right) d_{k} y_{k}$. Thus, we have:

$$
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)-\operatorname{Var}\left(\hat{T}_{\mathrm{DI}}\right)=\operatorname{Var}\left(\sum_{k \in s_{p}} \delta_{k} d_{k} y_{k}\right)+2 \operatorname{Cov}\left(\sum_{k \in s_{p}} \delta_{k} d_{k} y_{k}, \sum_{k \in s_{p}}\left(1-\delta_{k}\right) d_{k} y_{k}\right) .
$$

(i) For Poisson sampling, we have:

$$
\operatorname{Cov}\left(\sum_{k \in s_{p}} \delta_{k} d_{k} y_{k}, \sum_{k \in s_{p}}\left(1-\delta_{k}\right) d_{k} y_{k}\right)=\sum_{k \in U} \delta_{k}\left(1-\delta_{k}\right)\left(d_{k}-1\right) y_{k}^{2}=0
$$

and

$$
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)-\operatorname{Var}\left(\hat{T}_{\mathrm{DI}}\right)=\operatorname{Var}\left(\sum_{k \in s_{p}} \delta_{k} d_{k} y_{k}\right)=\sum_{k \in U} \delta_{k}\left(d_{k}-1\right) y_{k}^{2} \geq 0,
$$

which proves the first part of the proposition.
(ii) For simple random sampling without replacement, let $\bar{Y}_{U}=\sum_{k \in U} y_{k} / N, \bar{Y}_{\mathrm{NP}}=\sum_{k \in U} \delta_{k} y_{k} / N_{\mathrm{NP}}$, $S_{Y, \mathrm{NP}}^{2}=\sum_{k \in U} \delta_{k}\left(y_{k}-\bar{Y}_{\mathrm{NP}}\right)^{2} /\left(N_{\mathrm{NP}}-1\right)$ and $\mathrm{CV}_{\mathrm{NP}}^{2}=S_{Y, \mathrm{NP}}^{2} / \bar{Y}_{\mathrm{NP}}^{2}$. Using some simple calculus, we have:

$$
\begin{aligned}
& \operatorname{Var}\left(\sum_{k \in s_{P}} \delta_{k} d_{k} y_{k}\right)=\frac{N}{n} \frac{N-n}{N(N-1)}\left(N\left(N_{\mathrm{NP}}-1\right) S_{Y, \mathrm{NP}}^{2}+N_{\mathrm{NP}} \bar{Y}_{\mathrm{NP}}^{2}\left(N-N_{\mathrm{NP}}\right)\right), \\
& \operatorname{Cov}\left(\sum_{k \in s_{P}} \delta_{k} d_{k} y_{k}, \sum_{k \in s_{p}}\left(1-\delta_{k}\right) d_{k} y_{k}\right)=-\frac{N}{n} \frac{N-n}{N(N-1)} N_{\mathrm{NP}} \bar{Y}_{\mathrm{NP}}\left(N \bar{Y}_{U}-N_{\mathrm{NP}} \bar{Y}_{\mathrm{NP}}\right),
\end{aligned}
$$

and thus

$$
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)-\operatorname{Var}\left(\hat{T}_{\mathrm{DI}}\right)=\frac{N}{n} \frac{N-n}{N(N-1)}\left(N\left(N_{\mathrm{NP}}-1\right) S_{Y, \mathrm{NP}}^{2}+N_{\mathrm{NP}} \bar{Y}_{\mathrm{NP}}\left(\left(N+N_{\mathrm{NP}}\right) \bar{Y}_{\mathrm{NP}}-2 N \bar{Y}_{U}\right)\right) .
$$

We conclude that $\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)$ is larger than or equal to $\operatorname{Var}\left(\hat{T}_{\mathrm{DI}}\right)$ if and only if

$$
N\left(N_{\mathrm{NP}}-1\right) S_{Y, \mathrm{NP}}^{2}+N_{\mathrm{NP}} \bar{Y}_{\mathrm{NP}}\left(\left(N+N_{\mathrm{NP}}\right) \bar{Y}_{\mathrm{NP}}-2 N \bar{Y}_{U}\right) \geq 0,
$$

which is equivalent to:

$$
\mathrm{CV}_{\mathrm{NP}}^{2} \geq-\frac{N_{\mathrm{NP}}}{N_{\mathrm{NP}}-1}\left(1+\frac{N_{\mathrm{NP}}}{N}-2 \frac{\bar{Y}_{U}}{\bar{Y}_{\mathrm{NP}}}\right)
$$

and proves the second part of the proposition.

## Proof of Proposition 2.2

We have:

$$
\begin{gathered}
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)=\operatorname{Var}\left(\sum_{k \in s_{P}} d_{k} y_{k}\right)=N^{2}(1-f) \frac{S_{Y, U}^{2}}{n}, \\
\operatorname{AVar}\left(\hat{T}_{\mathrm{PDI}}\right)=\operatorname{Var}\left(\sum_{k \in s_{P}}\left(1-\delta_{k}\right) d_{k}\left(y_{k}-\bar{Y}_{C}\right)\right)=\operatorname{Var}\left(\sum_{k \in s_{P}} d_{k} \tilde{y}_{k}\right)=N^{2}(1-f) \frac{S_{\tilde{Y}, U}^{2}}{n}
\end{gathered}
$$

where

$$
\begin{aligned}
S_{Y, U}^{2} & =\frac{1}{N-1} \sum_{k \in U}\left(y_{k}-\bar{Y}_{U}\right)^{2}, \\
\tilde{y}_{k} & =\left(1-\delta_{k}\right)\left(y_{k}-\bar{Y}_{C}\right), k \in U \\
S_{\tilde{Y}, U}^{2} & =\frac{1}{N-1} \sum_{k \in U}\left(\tilde{y}_{k}-\bar{Y}_{C}\right)^{2}=\frac{1}{N-1} \sum_{k \in U} \tilde{y}_{k}^{2} .
\end{aligned}
$$

Using some basic but tedious calculus, we obtain:

$$
\begin{aligned}
\operatorname{Var}\left(\hat{T}_{\mathrm{HT}}\right)-\operatorname{AVar}\left(\hat{T}_{\mathrm{PDI}}\right) & =N^{2}(1-f) \frac{S_{Y, U}^{2}-S_{\tilde{Y}, U}^{2}}{n} \\
& =N^{2}(1-f) \frac{1}{n} \frac{1}{N-1}\left(\sum_{k \in U} \delta_{k}\left(y_{k}-\bar{Y}_{U}\right)^{2}+\left(N-N_{\mathrm{NP}}\right)\left(\bar{Y}_{C}-\bar{Y}_{U}\right)^{2}\right) \\
& =N^{2}(1-f) \frac{1}{n} \frac{1}{N-1}\left(S_{Y, \mathrm{NP}}^{2}\left(N_{\mathrm{NP}}-1\right)+N_{\mathrm{NP}} \frac{N}{N-N_{\mathrm{NP}}}\left(\bar{Y}_{\mathrm{NP}}-\bar{Y}_{U}\right)^{2}\right) .
\end{aligned}
$$

## Proof of Proposition 3.1

Let $\quad \mathbf{R}_{s_{P}}=\operatorname{diag}\left(r_{k} \delta_{k}\right)_{k \in s_{p}}, \mathbf{X}_{s_{p}}=\left(\mathbf{x}_{k}^{\top}\right)_{k \in s_{p}}, \mathbf{y}_{s_{p}}=\left(y_{k}\right)_{k \in s_{p}} \quad$ and $\quad \mathbf{Q}_{x s_{p}}^{\top}=\mathbf{X}_{s_{p}}^{\top} \operatorname{diag}\left(q_{k} \delta_{k}\right)_{k \in s_{p}}$. Then $\hat{\boldsymbol{\beta}}=$ $\left(\mathbf{Q}_{x s_{p}}^{\top} \mathbf{X}_{s_{p}}\right)^{-1} \mathbf{Q}_{x s_{p}}^{\top} \mathbf{y}_{s_{p}}$. We can write the sum $\sum_{k \in s_{p}} r_{k} \delta_{k}\left(y_{k}-\hat{y}_{k}\right)$ in a matrix form as follows:

$$
\sum_{k \in s_{p}} r_{k} \delta_{k}\left(y_{k}-\hat{y}_{k}\right)=\mathbf{1}_{s_{p}}^{\top} \mathbf{R}_{s_{p}}\left(\mathbf{y}_{s_{p}}-\mathbf{X}_{s_{p}} \hat{\boldsymbol{\beta}}\right),
$$

where $\mathbf{1}_{s_{P}}$ is a vector of ones with dimension the size of $s_{P}$. If the condition $\boldsymbol{\mu}^{\top} \mathbf{x}_{k} q_{k}-r_{k}=0$ is fulfilled for all $k \in s_{\mathrm{NP}}$, then $\delta_{k}\left(\boldsymbol{\mu}^{\top} \mathbf{x}_{k} q_{k}-r_{k}\right)=0$ for all $k \in s_{P}$ and as a consequence, $\boldsymbol{\mu}^{\top} \mathbf{Q}_{x s_{P}}^{\top}=\mathbf{1}_{s_{P}}^{\top} \mathbf{R}_{s_{P}}$. We get then $\mathbf{1}_{s_{P}}^{\top} \mathbf{R}_{s_{P}}\left(\mathbf{y}_{s_{P}}-\mathbf{X}_{s_{p}} \hat{\boldsymbol{\beta}}\right)=0$.

## Proof of Proposition 3.2

We have

$$
\begin{aligned}
\hat{T}_{\mathrm{NP}}^{(\mathrm{OR})}-\hat{T}_{\mathrm{NP}}^{(\mathrm{Q} \pi)} & =\sum_{k \in s_{p}}\left(r_{k}-d_{k}\right) \delta_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right) \\
& =-\lambda^{\top} \sum_{k \in s_{p}} q_{k} \delta_{k} \mathbf{x}_{k}\left(y_{k}-\mathbf{x}_{k}^{\top} \hat{\boldsymbol{\beta}}\right)=0 .
\end{aligned}
$$

## References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. Journal of the American Statistical Association, 81(396), 1074-1079.

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

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. The Annals of Statistics, 28(4), 1026-1053.

Brewer, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. Survey Methodology, 25, 2, 205-212. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4883eng.pdf.

Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. Electronic Journal of Statistics, 7, 562-596.

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample data. Biometrika, 96(3), 691-709.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. Journal American Statistical Association, 77(377), 89-96.

Kim, J.-K. (2022). A gentle introduction to data integration in survey sampling. The Survey Statistician, 85, 19-29.

Kim, J.-K., and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 89(2), 382-401.

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

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. Survey Methodology, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. Sankhy $\bar{a}$ B, 83(1), 242-272.

Särndal, C.-E. (1980). On $\pi$-inverse weighting best linear unbiased weighting in probability sampling. Biometrika, 67(3), 639-650.

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

Särndal, C.-E., and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. Scandinavian Journal of Statistics, 11(3), 146-156.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. Journal of the American Statistical Association, 78(384), 879-884.

Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, 48, 2, 283311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002eng.pdf.

Yang, S., and Kim, J.-K. (2020). Integration of survey data and big observational data for finite population inference using mass imputation. Japanese Journal of Statistics and Data Science, 3, 625-650.

# Constructing all determinantal sampling designs 

Vincent Loonis ${ }^{1}$


#### Abstract

In this article, we use a slightly simplified version of the method by Fickus, Mixon and Poteet (2013) to define a flexible parameterization of the kernels of determinantal sampling designs with fixed first-order inclusion probabilities. For specific values of the multidimensional parameter, we get back to a matrix from the family $P^{\Pi}$ from Loonis and Mary (2019). We speculate that, among the determinantal designs with fixed inclusion probabilities, the minimum variance of the Horvitz and Thompson estimator (1952) of a variable of interest is expressed relative to $P^{\Pi}$. We provide experimental R programs that facilitate the appropriation of various concepts presented in the article, some of which are described as non-trivial by Fickus et al. (2013). A longer version of this article, including proofs and a more detailed presentation of the determinantal designs, is also available.


Key Words: Determinantal process; Balanced sampling; Semidefinite optimization.

## 1. Introduction

In sampling theory, a random sample $\mathbb{S}$ is a random variable whose realizations are elements of the set $2^{U}$ of the parts of a finite population $U$ of size $N$, indexed by $k=1, \ldots, N$. Each part $s$ of $2^{U}$ is called a sample. The probability law of $\mathbb{S}$ is called sampling design (Tillé, 2001). Apart from the terminology, these concepts are the same as those used in point processes on a finite population, in probability or statistics. Among point processes, determinantal processes have been the subject of many studies because they appear in various fields: random matrices, mathematical physics or machine learning. Using these processes in the context of sampling theory and finite populations leads to the concept of determinantal sampling design studied by Loonis and Mary (2019).

Determinantal sampling designs directly inherit properties from determinantal processes established in different frameworks. They are parameterized by Hermitian matrices whose eigenvalues lie between 0 and 1 (Macchi, 1975; Soshnikov, 2000). These are called contracting Hermitian matrices. Such a matrix will then be notated $K$, called the kernel, and the associated determinantal sampling design will be notated $\mathrm{DSD}(K)$. The inclusion probabilities of such designs are known at all orders and are parameterized by $K$. This feature differentiates determinantal designs from most of the usual, somewhat complex, designs giving them a real practical interest, beyond their theoretical curiosity. Usually, if the first-order inclusion probabilities are known, the second-order inclusion probabilities are often approximated and the others are most often unknown. For determinantal designs, the first- and second-order inclusion probabilities are given, respectively, by the diagonal and non-diagonal terms of the kernel. Higher-order inclusion probabilities are also expressed directly in terms of $K$.

Moreover, the distribution of the size of a determinantal random sample is that of a sum of independent Bernoulli variables, where the parameters are the eigenvalues of $K$ (Hough, Krishnapur, Peres and Virág,

[^5]2006). There are algorithms for selecting determinantal random samples (Hough, Krishnapur, Peres and Virág, 2006; Scardicchio, Zachary and Torquato, 2009; Lavancier, Møller and Rubak, 2015). Determinantal designs are negatively associated (Lyons, 2003). In other words, if $A$ and $B$ are two disjointed parts of $U$, then $\operatorname{pr}(A \cup B \subseteq \mathbb{S}) \leq \operatorname{pr}(A \subseteq \mathbb{S}) \operatorname{pr}(B \subseteq \mathbb{S})$. As a result, determinantal designs satisfy the Sen-YatesGrundy conditions (Sen, 1953; Yates and Grundy, 1953), which correspond to the case where $A$ and $B$ are singletons. They also satisfy the strong Rayleigh property (Yuan, Su and $\mathrm{Hu}, 2003$; Brändén and Jonasson, 2012; Pemantle and Peres, 2014). From this technical property, stronger than negative association, a central limit theorem as well as deviation and concentration inequalities (Soshnikov, 2000, 2002; Pemantle et Peres, 2014) directly follow for the Horvitz-Thompson estimator of the total of a variable of interest $y$ (Horvitz and Thompson, 1952).

The specifics of sampling theory lead Loonis and Mary (2019) to focus on novel properties of determinantal designs, such as the necessary and sufficient conditions for the perfect estimation of a total with the Horvitz-Thompson estimator, the search for optimal kernels, or the explicit construction of families of particular contracting Hermitian matrices, such as fixed-diagonal orthogonal projection matrices. These matrices are associated with the important practical case of fixed-size determinantal sampling designs with given first-order inclusion probabilities. More practically, the authors observe that the simplest determinantal designs can be mobilized for populations of several thousand individuals. The size of the population that is compatible with more sophisticated constructions, particularly those associated with optimization issues, is several hundred.

Constructing a fixed-diagonal orthogonal projection matrix is a specific case of the more general problem of constructing Hermitian matrices, whose diagonal and spectrum are fixed. This problem has been the subject of many studies in the literature, some of which are recent. Schur (1911), Horn (1954) and Kadison (2002) studied the necessary and sufficient conditions for the existence of such matrices in complex and real cases. However, these studies do not specify how to construct such matrices. Dhillon, Heath Jr., Sustik and Tropp (2005) propose algorithms for constructing some of these matrices. The algorithm by Fickus et al. (2013) makes it possible to construct them all. In the previous two cases, the results are not known analytically. Loonis and Mary (2019) explicitly show several examples of projection matrices, whose diagonal terms are constant. Building on the work of Kadison (2002), they present formulae for specifically constructing a family $P^{\Pi}$ of real orthogonal projection matrices of any diagonal $\Pi$, provided that $\sum_{k=1}^{N} \Pi_{k}$ is an integer.

In this article, we focus specifically on the method by Fickus et al. (2013), in the context of determinantal designs, and draw connections with the family $P^{\Pi}$. In the first section, we introduce some algebraic notations and concepts that are useful for the overall understanding of the article. In the second section, we reiterate the main properties of determinantal sampling designs. More precisely, we present the properties of the family of designs associated with matrices $P^{\Pi}$. The results of this section are taken directly, without proof, from Loonis and Mary (2019). Readers who wish to understand the foundations of these results or are interested in extensions, such as asymptotic properties, are encouraged to refer to the original article. Algorithm 3.1 and theorem 3.1 are new, and their proofs are provided in the longer version of the article.

In the third section, we present the main principles and parameters of the method by Fickus et al. (2013). The latter are not easy to handle directly. We show that, at the cost of simplification, it is possible to modify them to describe the semidefinite positive Hermitian matrices $K^{\Pi}$ of diagonal $\Pi$ with a triplet $(M, \Omega, \rho)$. $M$ is an integer giving the number of strictly positive eigenvalues of $K^{\Pi} . \Omega$ and $\rho$ are matrices with respective sizes $(M \times N)$ and $(M \times(N-1))$, all of whose coefficients have value in [0,1]. In terms of sampling theory, the columns of matrix $\Omega$ directly influence the variability of the sample size falling into domains of the form $D^{k}=\{1, \ldots, k\}, k=1, \ldots, N$. As for the matrix $\rho$, it determines the value of the nondiagonal coefficients of the constructed matrices and, as such, the second-order inclusion probabilities of the associated determinantal designs.

In the fourth section, for a probability vector $\Pi$ such that $\sum_{k=1}^{N} \Pi_{k}=n \in \mathbb{N}^{*}$, we examine the choices $\left(M=n, \Omega=0^{(n \times N)}, \rho\right)$ and $\left(M=n, \Omega=1^{(n \times N)}, \rho\right)$. We show that the first is directly related to the family $P^{\Pi}$ and explain the coefficients of the matrices obtained for the second choice.

In the final section, we carry out simulations and applications. We limit ourselves to real fixed-diagonal orthogonal projection matrices. In this set, we minimize the variance of the Horvitz-Thompson estimator of a variable's total. Given the specificity of the eligible set, we rewrite the problem in the form of an optimization problem on manifolds and mobilize adapted algorithms (Absil, Mahony and Sepulchre, 2009; Boumal, Mishra, Absil and Sepulchre, 2014; Townsend, Koep and Weichwald, 2016). We find that the result is also associated with the family $P^{\Pi}$. From this we deduce a conjecture defining the lower bound of the variance of the Horvitz-Thompson estimator of the total of variable $y$, among the determinantal designs with fixed first-order inclusion probabilities.

In addition to appropriating and programming the method by Fickus et al. (2013) for use in determinantal sampling designs, the main contributions of this document are algorithm 3.1, theorems 3.1, 5.1 and 5.2, and conjecture 6.1. The points considered important for interpreting certain results are presented in the form of remarks or examples. The proofs, which are long and technical, are provided in a long version of the article. The latter also provides a comparison of the performances, in terms of balancing, of the determinantal designs with those of equivalent designs (Deville and Tillé, 2004; Chauvet and Tillé, 2006; Leuenberger, Eustache, Jauslin and Tillé, 2022).

## 2. Algebraic notations and reminders

In the rest of the article, $K$ is a contracting Hermitian matrix of size $(N \times N)$. A random variable, whose distribution is a determinantal design with kernel $K$, will be notated $\mathbb{S} \sim \operatorname{DSD}(K)$. The coefficients of $K$ are complex. The conjugate of the complex number $z$ is $\bar{z}$ and its modulus is $|z| . K$ is such that $K=\bar{K}^{\top}$ (Hermitian) and its eigenvalues are in the interval [0,1] (contracting). If all eigenvalues are 0 or $1, K$ is an orthogonal projection matrix. The strictly positive eigenvalues of $K$ are $M$ in number. The main submatrix of order $k=1, \ldots, N$ of $K$ is the square matrix made up of the upper left corner of size $k$ of $K$. The submatrix of order $k=N$ is $K$ itself.

The set of eigenvalues, called spectrum, of the main submatrix of order $k$ is represented by the vector $\lambda^{k}$. The spectrum of $K$ is $\lambda^{N}$. Each $\lambda^{k}$ has at most $M$ strictly positive eigenvalues, which are also less than or equal to 1 . Each main submatrix is also Hermitian and contracting. By convention, $\lambda^{k}$ is truncated to its own strictly positive coefficients at first. If applicable, it is supplemented with 0 s to be of constant size $M$. The coefficients of $\lambda^{k}$ are notated $\lambda_{j}^{k}, j=1, \ldots, M, k=1, \ldots, N$ and are, again by convention, sorted from lowest to highest: $0 \leq \lambda_{1}^{k} \leq \ldots \leq \lambda_{j}^{k} \leq \ldots \leq \lambda_{M}^{k} \leq 1$ with $\lambda_{1}^{N}>0$.

The $\lambda^{k}$, as spectra of the main submatrices, satisfy the Cauchy interlacing conditions (Horn and Johnson, 1991). By placing $\lambda_{0}^{k}=0$ for $k=1, \ldots, N$, these conditions are written

$$
\begin{equation*}
\forall j=1, \ldots, M, \forall k=1, \ldots, N-1: \lambda_{j-1}^{k+1} \leq \lambda_{j}^{k} \leq \lambda_{j}^{k+1} . \tag{2.1}
\end{equation*}
$$

The letter $k$ is used to index both the individuals of the population and the steps of the algorithm by Fickus et al. (2013). This choice is justified by the fact that step $k$ determines individual $k$ 's contribution to the inclusion probabilities of all orders involving it. In general, the first-order inclusion probability of individual $k=1, \ldots, N$ is $\pi_{k}$, the coefficient of vector $\pi$. In some cases, which leave no room for doubt, the letter $\pi$ is used based on its conventional usage. The notations $\Pi_{k}$ and $\Pi$ will be used when we want $\pi_{k}$ to take the value $\Pi_{k}$, set in advance. For example, $\Pi_{k}=n / N$ or $\Pi_{k}$ is proportional to a size criterion, like the number of employees for companies. The various inclusion probabilities contained in vector $\Pi$ are not necessarily ordered. This is true for constructing matrices $P^{\Pi}$ from Loonis and Mary (2019). For the method by Fickus et al. (2013), these probabilities will be sorted from highest to lowest, without losing generality. In this case, we will use the notation $\Pi^{\triangleright}$, specifying that $1>\Pi_{1}^{\triangleright} \geq \ldots \geq \Pi_{k}^{\triangleright} \geq \ldots \geq \Pi_{N}^{\triangleright}>0$. The notation $\Sigma^{\Pi^{\top}}$ refers to a permutation matrix such that $\Pi^{\triangleright}=\Sigma^{\Pi^{\top}} \Pi$.

To be consistent with sampling theory, the notations $N$ and $M$ are inverted, compared with those from Fickus et al. (2013). The ordering convention of $\lambda^{k}$ also differs from the one chosen by these authors. This seemed to simplify the appropriation of the methods from a programming perspective. This choice can occasionally make the wording of the Schur (1911) and Horn (1954) theorem, as well as some formulae that derive from it, less intuitive. In this context, this theorem stipulates that there is a Hermitian matrix $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$ with diagonal $\Pi^{\triangleright}$ and spectrum $\lambda^{N}$ if and only if

$$
\left\{\begin{array}{l}
\sum_{k=1}^{N} \Pi_{k}^{\triangleright}=\sum_{j=1}^{M} \lambda_{j}^{N}=\mu  \tag{2.2}\\
\sum_{s=M-j+1}^{N} \Pi_{s}^{\triangleright}=\mu-\sum_{s=1}^{M-j} \Pi_{s}^{\triangleright} \geq \sum_{s=1}^{j} \lambda_{s}^{N}, j=1, \ldots, M-1 .
\end{array}\right.
$$

By adapting the notations, this theorem also applies to all the main submatrices of a Hermitian matrix because they are also Hermitian.

Finally, a matrix $A$ of size $(P \times P)$ is unitary if $\bar{A}^{\top} A=I_{P}$, where $I_{P}$ is the identity matrix. The square matrix of size $(P \times P)$, all terms of which equal 1 , is notated $J_{P}$. The symbol $\odot$ indicates Hadamard's matrix product: $A=B \odot C \Rightarrow A_{k l}=B_{k l} C_{k l}$. The notation $x^{(P \times Q)}$ indicates a matrix of size $(P \times Q)$, all coefficients of which equal $x$. The notation $x^{P}$ is used for vectors of size $P$, all coefficients of which equal $x . D_{x}$ is a diagonal matrix, the diagonal of which is $x$.

## 3. Reminders about determinantal sampling designs

### 3.1 Definition and inclusion probabilities

A random variable $\mathbb{S}$ over $2^{U}$ has as law a determinantal sampling design if there is a contracting Hermitian matrix $K$ such that

$$
\begin{equation*}
\forall s \in 2^{U}, \operatorname{pr}(s \subseteq \mathbb{S})=\operatorname{det}\left(K_{\mid s}\right), \tag{3.1}
\end{equation*}
$$

where $K_{\mid s}$ is the submatrix constructed by extracting the rows and columns of $K$ indexed by the elements of $s$. This definition directly results in the calculation of the inclusion probabilities for all orders, particularly those of orders 1 and 2:

$$
\left\{\begin{array}{l}
\pi_{k}=\operatorname{pr}(\{k\} \subseteq \mathbb{S})=\operatorname{det}\left(K_{\mid k, k}\right)=K_{k k},  \tag{3.2}\\
\pi_{k l}=\operatorname{pr}(\{k, l\} \subseteq \mathbb{S})=\operatorname{det}\left(K_{\mid\{k, l\}}\right)=K_{k k} K_{l l}-K_{k l} K_{l k}=K_{k k} K_{l l}-K_{k l} \bar{K}_{k l}=K_{k k} K_{l l}-\left|K_{k l}\right|^{2}
\end{array}\right.
$$

The diagonal terms $K_{k k}$ correspond to the first-order inclusion probabilities of design $\operatorname{DSD}(K)$, whereas the modulus of the non-diagonal terms is used in the expression of the second-order inclusion probabilities. The properties of the trace matrix application give rise to a relationship between inclusion probabilities and eigenvalues, which frequently appear in the proofs:

$$
\begin{equation*}
\operatorname{Tr}(K)=\sum_{k=1}^{N} K_{k k}=\sum_{k=1}^{N} \pi_{k}=\sum_{j=1}^{M} \lambda_{j}^{N} . \tag{3.3}
\end{equation*}
$$

Matrix $\Delta$, of size $(N \times N)$, whose coefficients are $\Delta_{k k}=\pi_{k}\left(1-\pi_{k}\right)$ and $\Delta_{k l}=\pi_{k l}-\pi_{k} \pi_{l}$, is expressed in terms of $K$ :

$$
\begin{cases}\Delta_{k k} & =K_{k k}\left(1-K_{k k}\right),  \tag{3.4}\\ \Delta_{k l} & =-\left|K_{k l}\right|^{2}, \\ \Delta & =K \odot\left(I_{N}-\bar{K}\right) .\end{cases}
$$

This matrix is used in precision calculations. It also makes it possible to see that the determinantal designs confirm the Sen-Yates-Grundy conditions (Sen, 1953; Yates and Grundy, 1953) because $\Delta_{k l}=\pi_{k l}-$ $\pi_{k} \pi_{l}=-\left|K_{k l}\right|^{2} \leq 0$ and, therefore, $\pi_{k l} \leq \pi_{k} \pi_{l}$. As such, the Sen-Yates-Grundy conditions provide an upper bound for the second-order inclusion probabilities of determinantal designs, function of single-order inclusion probabilities.

### 3.2 Sample size, fixed-size determinantal designs and selection algorithm

Let $\ddagger \mathbb{S}$ be the size of random sample $\mathbb{S} \sim \operatorname{DSD}(K)$. The distribution of $\sharp \mathbb{S}$ is that of a sum of $M$ independent Bernoulli variables, whose parameters are the $M$ strictly positive eigenvalues $\lambda_{j}^{N}, j=$ $1, \ldots, M$ of $K$ (Hough et al., 2006). The moments of order 1 and 2 of $\mathbb{Z}$ are deduced and equal

$$
\left\{\begin{array}{l}
\mathbb{E}(\not \mathbb{\mathbb { S }})=\sum_{j=1}^{M} \lambda_{j}^{N}=\sum_{k=1}^{N} \pi_{k},  \tag{3.5}\\
\operatorname{var}(\sharp \mathbb{S})=\sum_{j=1}^{M} \lambda_{j}^{N}\left(1-\lambda_{j}^{N}\right) .
\end{array}\right.
$$

Remark 3.1 $\mathbb{S} \sim \operatorname{DSD}(K)$ will be of fixed size, $\mathbb{E}(\nmid \mathbb{S})=\sharp \mathbb{S}=n \in \mathbb{N}^{*}$, if and only if $\operatorname{var}(\sharp \mathbb{S})=0$. In other words, if $\lambda_{j}^{N}$ equals 1 for all $j=1, \ldots, M$. Because $K$ is Hermitian, $K$ is an orthogonal projection matrix.

Lavancier et al. (2015) propose an algorithm for selecting a random sample whose distribution is a determinantal design of fixed size. This algorithm requires a spectral decomposition of $K$ as input, which, when $N$ is large, can be computationally intensive. One of the challenges of then constructing kernels $K$ will be to directly provide this decomposition. If $K$ is not a projection matrix, Hough et al. (2006) show that $\operatorname{DSD}(K)$ can be written as a mixture of fixed-size determinantal designs. Thus, it is possible to refer to the case of projection matrices to select a random-sized determinantal sample.

### 3.3 Estimating a total, variance of the estimators, balanced designs

For any sampling design, whose first-order inclusion probabilities are strictly positive, the unknown total $t_{y}=\sum_{k=1}^{N} y_{k}$ of a variable of interest $y$, assimilated to a vector of $\mathbb{R}^{N}$, is estimated without bias by the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) such that $\hat{t}_{y}=\sum_{k \in \mathrm{~S}} y_{k} / \pi_{k}$ and whose variance is

$$
\begin{equation*}
\operatorname{var}\left(\hat{t}_{y}\right)=\sum_{k \in U} \sum_{l \in U} \frac{y_{k}}{\pi_{k}} \frac{y_{l}}{\pi_{l}} \Delta_{k l}=y^{\top} D_{\pi}^{-1} \Delta D_{\pi}^{-1} y . \tag{3.6}
\end{equation*}
$$

If the design is determinantal, a consequence of (3.4) is that the variance of $\hat{t}_{y}$ is a function of $K$ :

$$
\begin{equation*}
\operatorname{var}\left(\hat{t}_{y}\right)=y^{\top}\left(I_{N} \odot K\right)^{-1}\left[K \odot\left(I_{N}-\bar{K}\right)\right]\left(I_{N} \odot K\right)^{-1} y \tag{3.7}
\end{equation*}
$$

For a set of inclusion probabilities fixed a priori, a balanced design is such that

$$
\begin{equation*}
\forall q=1, \ldots, Q, \quad \hat{t}_{x^{q}}=\sum_{k \in \mathbb{S}} \frac{x_{k}^{q}}{\Pi_{k}}=t_{x^{q}} \Leftrightarrow \operatorname{var}\left(\hat{t}_{x^{q}}\right)=0 \tag{3.8}
\end{equation*}
$$

where the $x^{q}, q=1, \ldots, Q$, are a set of auxiliary variables, in other words, variables whose value is known for all individuals in the population, particularly via the sampling frame. Deville and Tillé (2004) propose an efficient approach for approximately resolving this problem. These authors use algebraic and probability methods. By staying within the determinantal family, another possibility is to use optimization techniques and to find $K^{\text {opt }}$ such that

$$
K^{\mathrm{opt}}=\underset{K}{\operatorname{argmin}} \sum_{q=1}^{Q} \alpha_{q} \operatorname{var}\left(\hat{x}_{x^{q}}\right) \text {, s.c. }\left\{\begin{array}{l}
K=\bar{K}^{\top}  \tag{3.9}\\
\operatorname{diag}(K)=\Pi \\
0 \preceq K \preceq I_{N},
\end{array}\right.
$$

where $\operatorname{var}\left(\hat{t}_{x^{q}}\right)$ is given by (3.7), $\Pi$ is fixed a priori, and $0 \preceq K \preceq I_{N}$ indicates that the eigenvalues of $K$ are between 0 and 1 . The coefficients $\alpha_{q}$ make it possible to manage the relative importance of the variables.

The choice $\alpha_{q}=1 / t_{x_{q}}^{2}$ leads to minimizing the sum of the squares of the variation coefficients of the various estimators. Resolving such problems involves non-linear positive semidefinite optimization. Loonis and Mary (2019) propose heuristics to find approximate solutions. We will in the following use optimization methods on manifolds (Absil et al., 2009; Boumal et al., 2014; Townsend et al., 2016).

### 3.4 A few examples of determinantal designs

### 3.4. Constructing one design from another

Below we present some general properties for constructing new determinantal designs from a given determinantal design. Let $\mathbb{S} \sim \operatorname{DSD}(K)$, then

1. the complement $\mathbb{S}^{c}$ of $\mathbb{S}$ in $U$ is determinantal and has as distribution $\operatorname{DSD}\left(I_{N}-K\right)$,
2. the restriction $\mathbb{S} \cap D$ of $\mathbb{S}$ to the domain $D$ included in $U$ is determinantal of distribution $\operatorname{DSD}\left(K_{\mid D}\right)$,
3. $\operatorname{DSD}(K)$ is stratified if and only if $K$ is a block diagonal, up to a permutation of the rows and the columns,
4. if $U_{1}$ is a unitary matrix, the matrix $U_{1} K \bar{U}_{1}^{\top}$ has the same eigenvalues as $K$. Therefore, there is a determinantal design associated with it. Among the unitary transformations, rotations prove useful for defining optimization heuristics (Loonis and Mary, 2019).

### 3.4.2 A family of fixed-size determinantal designs and any inclusion probabilities fixed a priori: The family $\boldsymbol{P}^{\square}$

Loonis and Mary (2019) reiterate that the Poisson design is determinantal, whereas the simple random design is determinantal only in the cases $n=1$ and $n=N-1$. The authors provide examples of determinantal designs with constant inclusion probabilities. They construct a family of fixed-size determinantal designs for any inclusion probabilities fixed a priori. To the extent that it will reappear in various contexts in the following sections, we specifically present the properties of the designs associated with this family in this section.

Let $\Pi$ be a vector of $] 0,1\left[{ }^{N}\right.$, such that $\sum_{k=1}^{N} \Pi_{k}=n, n \in \mathbb{N}^{*}$. Loonis and Mary (2019) mobilize a result from Kadison (2002) to exhibit an explicit formula of the coefficients of a real orthogonal projection matrix $P^{\Pi}$ with diagonal $\Pi$ (Table 3.1). The determinantal design associated with this matrix is of fixed size and has inclusion probabilities $\Pi$. The formula is based on the integers $k_{r}(r=1, \ldots, n-1)$ such that $\sum_{k=1}^{k_{r}-1} \Pi_{k}<r$ and $\sum_{k=1}^{k_{r}} \Pi_{k} \geq r$, the real numbers $\alpha_{k_{r}}=r-\sum_{k=1}^{k_{k}-1} \Pi_{k}$ and $\gamma_{r}^{r^{\prime}}$ such that

$$
\gamma_{r}^{\prime}=\sqrt{\prod_{j=r+1}^{\prime} \frac{\left(\Pi_{k_{j}}-\alpha_{k_{j}}\right) \alpha_{k_{j}}}{\left(1-\alpha_{k_{k}, j}\right)\left(1-\left(\Pi_{k_{j}}-\alpha_{k_{j}}\right)\right)}},
$$

for $r<r^{\prime}, \gamma_{r}^{r^{\prime}}=1$ otherwise.

Table 3.1
Coefficients $\boldsymbol{P}_{k l}^{\Pi}$ of $\boldsymbol{P}^{\Pi}: \boldsymbol{k}<\boldsymbol{l}$.

|  | $l$ |  |
| :---: | :---: | :---: |
| $k$ | $l=k_{r^{\prime}+1}$ | $k_{r^{\prime}}<l<k_{r^{\prime}+1}$ |
| $k_{r}<k<k_{r+1}$ | $\sqrt{\Pi_{k}} \sqrt{\frac{\left(1-\Pi_{l}\right) \alpha_{l}}{\left.1-\alpha_{l}\right)} \gamma_{r}^{r^{\prime}}}$ | $\sqrt{\Pi_{k} \Pi_{l} \gamma_{r}^{r^{\prime}}}$ |
| $k=k_{r}$ | $-\sqrt{\frac{\left(1-\Pi_{l}\right) \alpha_{l}}{1-\alpha_{l}}} \sqrt{\frac{\left(1-\Pi_{k}\right)\left(\Pi_{k}-\alpha_{k}\right)}{1-\left(\Pi_{k}-\alpha_{k}\right)}} \gamma_{r}^{r^{\prime}}$ | $-\sqrt{\Pi_{l}} \sqrt{\frac{\left(1-\Pi_{k}\right)\left(\Pi_{k}-\alpha_{k}\right)}{1-\left(\Pi_{k}-\alpha_{k}\right)}} \gamma_{r}^{r_{r}^{\prime}}$ |

Knowledge of the coefficients $P_{k l}^{\Pi}$ makes it possible to deduce information about the inclusion probabilities of order greater than 1 of design $\operatorname{DSD}\left(P^{\Pi}\right)$ :

1. If $\{k, l\}$ is an element of $] k_{r}, k_{r+1}\left[{ }^{2}\right.$ then $\pi_{k l}=0$ because $P_{k l}^{\Pi}=\sqrt{\Pi_{k} \Pi_{l}}$,
2. If $|k-l|$ is large $P_{k l}^{\Pi} \simeq 0$ and $\pi_{k l} \simeq \Pi_{k} \Pi_{l}$ is maximal under the Sen-Yates-Grundy constraint,
3. If $j \in] k_{r-1}, k_{r}\left[, k=k_{r}, l \in\right] k_{r}, k_{r+1}\left[\right.$, then $\pi_{j k l}=0$,
4. If there are integers $r_{1}, \ldots, r_{H}$ such that $\sum_{k=1}^{r_{r_{h}}} \Pi_{k}=r_{h}, h=1, \ldots, H$, then the design $\operatorname{DSD}\left(P^{\Pi}\right)$ is stratified based on the strata $\left.] k_{r_{h-1}}, k_{r_{h}}\right]$,
5. If $n$ divides $N$ and $\Pi_{k}=n / N$, then $\operatorname{DSD}\left(P^{\Pi}\right)$ is a 1-per-stratum design. It selects an individual from each of the $n$ groups of size $N / n$ taken consecutively within the population: $N / n$ first, and so on.

Example 3.1 illustrates some of these properties and also shows that the construction of $P^{\Pi}$ depends on the order of the individuals.

Example 3.1 Construction of $P^{\Pi}$ and $P^{\Pi^{\triangleright}}$ for $N=7, n=4$ and $\Pi=\left(\frac{1}{2}, \frac{3}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\right)^{\top}$. $\Pi^{\triangleright}$ is the vector of probabilities sorted from highest to lowest: $\Pi^{\triangleright}=\left(\frac{4}{5}, \frac{3}{4}, \frac{3}{4}, \frac{3}{5}, \frac{1}{2}, \frac{2}{5}, \frac{1}{5}\right)^{\top}$.

$$
(a): P^{\Pi}=\left(\begin{array}{ccccccc}
\frac{1}{2} & \frac{1}{2 \sqrt{2}} & \frac{1}{2 \sqrt{2}} & 0 & 0 & 0 & 0 \\
\frac{1}{2 \sqrt{2}} & \frac{3}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\
\frac{1}{2 \sqrt{2}} & -\frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{5} & \frac{\sqrt{2}}{5} & \frac{2}{5 \sqrt{3}} & \frac{\sqrt{2}}{5 \sqrt{3}} \\
0 & 0 & 0 & \frac{\sqrt{2}}{5} & \frac{2}{5} & \frac{2 \sqrt{2}}{5 \sqrt{3}} & \frac{2}{5 \sqrt{3}} \\
0 & 0 & 0 & \frac{2}{5 \sqrt{3}} & \frac{2 \sqrt{2}}{5 \sqrt{3}} & \frac{3}{5} & -\frac{\sqrt{2}}{5} \\
0 & 0 & 0 & \frac{\sqrt{2}}{5 \sqrt{3}} & \frac{2}{5 \sqrt{3}} & -\frac{\sqrt{2}}{5} & \frac{4}{5}
\end{array}\right),
$$

$$
(b): P^{\Pi^{\triangleright}}=\left(\begin{array}{ccccccc}
\frac{4}{5} & \frac{1}{2 \sqrt{5}} & \frac{1}{2 \sqrt{5}} & \frac{3}{5 \sqrt{7}} & \frac{-1}{\sqrt{210}} & \frac{2}{15 \sqrt{7}} & \frac{1}{15} \sqrt{\frac{2}{7}} \\
\frac{1}{2 \sqrt{5}} & \frac{3}{4} & -\frac{1}{4} & \frac{-3}{2 \sqrt{35}} & \frac{-1}{2 \sqrt{42}} & \frac{-1}{3 \sqrt{35}} & \frac{-1}{3 \sqrt{70}} \\
\frac{1}{2 \sqrt{5}} & -\frac{1}{4} & \frac{3}{4} & \frac{-3}{2 \sqrt{35}} & \frac{-1}{2 \sqrt{42}} & \frac{-1}{3 \sqrt{35}} & \frac{-1}{3 \sqrt{70}} \\
\frac{3}{5 \sqrt{7}} & \frac{-3}{2 \sqrt{35}} & \frac{-3}{2 \sqrt{35}} & \frac{3}{5} & \frac{1}{\sqrt{30}} & \frac{2}{15} & \frac{\sqrt{2}}{15} \\
\frac{-1}{\sqrt{210}} & \frac{-1}{2 \sqrt{42}} & \frac{-1}{2 \sqrt{42}} & \frac{1}{\sqrt{30}} & \frac{1}{2} & -\sqrt{\frac{2}{15}} & \frac{-1}{\sqrt{15}} \\
\frac{2}{15 \sqrt{7}} & \frac{-1}{3 \sqrt{35}} & \frac{-1}{3 \sqrt{35}} & \frac{2}{15} & -\sqrt{\frac{2}{15}} & \frac{2}{5} & \frac{\sqrt{2}}{5} \\
\frac{1}{15} \sqrt{\frac{2}{7}} & \frac{-1}{3 \sqrt{70}} & \frac{-1}{3 \sqrt{70}} & \frac{\sqrt{2}}{15} & \frac{-1}{\sqrt{15}} & \frac{\sqrt{2}}{5} & \frac{1}{5}
\end{array}\right) .
$$

If the initial order of the population, $\{1,2,3,4,5,6,7\}$ corresponds to vector $\Pi$, then the determinantal design associated with $P^{\Pi^{\triangleright}}$ applies to the population in the order (7, 2, 3, 6, 1, 5, 4) (or (7, 3, 2, 6, 1, 5, 4)). In connection with the previous properties, we see that $\Pi_{1}+\Pi_{2}+\Pi_{3}=2$. The determinantal design associated with $P^{\Pi}$ is indeed stratified based on the strata $(1,2,3)$ and $(4,5,6,7)$. For example, we have $P_{25}^{\Pi}=0$ and $\pi_{25}=\Pi_{2} \Pi_{5}-P_{25}^{\Pi^{2}}=\Pi_{2} \Pi_{5}$, according to (3.2). The design associated with $P^{\Pi^{\triangleright}}$ is not stratified because it is impossible to rearrange the rows and columns of this matrix to obtain a block diagonal matrix. The designs associated with $P^{\Pi}$ and $P^{\Pi^{\triangleright}}$ are different, even up to a permutation of the rows and the columns.

We conclude this part with two new results pertaining to matrices $P^{\Pi}$. The first provides a basis of eigenvectors of $P^{\Pi}$. This basis can be used as input for the algorithm of Lavancier et al. (2015).

## Algorithm 3.1 (Construction of $\Phi^{\boldsymbol{N}^{\top}}$, orthonormal eigenbasis of $\boldsymbol{P}^{\Pi}$ )

1. Set $k_{0}=0$. Define the integers $k_{r}(r=1, \ldots, n-1)$ such that $\sum_{k=1}^{k_{r}-1} \Pi_{k}<r$ and $\sum_{k=1}^{k_{r}} \Pi_{k} \geq r$ and the real numbers $\alpha_{k_{r}}=r-\sum_{k=1}^{k_{r}-1} \Pi_{k}$,
2. For every $k \in U$, calculate $s_{k}$ and $c_{k}$ such that

- if there is $r$ such that $k=k_{r}, s_{k_{r}}= \pm \sqrt{\frac{1-\Pi_{k_{r}}}{1-\alpha_{k_{r}}}}$
- otherwise for $k_{r}<k<k_{r+1}, s_{k}= \pm \sqrt{\frac{\Pi_{k}}{r+1-\sum_{i=1}^{k-1} \Pi_{i}}}$
- $c_{k}= \pm \sqrt{1-s_{k}^{2}}$, for every $k$,

3. Construct $\Phi^{N}$, matrix of size $(n \times N)$ of which all coefficients are nil except in $\left(r+1, k_{r}+1\right)$ (for $r=0, \ldots, n-1)$ where they equal 1. Let $\varphi_{k}$ be the column $k$ of $\Phi^{N}$,
4. Incrementally update the columns of $\Phi^{N}$ as follows, for $k=1, \ldots, N-1$ :
(a) calculate $C_{1}=s_{k} \varphi_{k}-c_{k} \varphi_{k+1}$
(b) calculate $C_{2}=c_{k} \varphi_{k}+s_{k} \varphi_{k+1}$
(c) respectively replace $\varphi_{k}$ and $\varphi_{k+1}$ with $C_{1}$ and $C_{2}$,
5. $\quad K^{\Pi}=\Phi^{N^{\top}} \Phi^{N}$ is such that $K^{\Pi}=H_{N} \odot P^{\Pi}$, where $H_{N}$ is a real symmetric matrix whose diagonal terms equal 1 and non-diagonal terms equal 1 or $-1 . \Phi^{N^{\top}}$ is an orthonormal eigenbasis of $K^{\Pi}$,
6. The matrix proposed by Loonis and Mary (2019) corresponds to the systematic choice of + in the choices $\pm$ of step 2. Another choice leads to the same coefficients as those in Table 3.1, up to the sign. Regardless of the choices, all matrices $K^{\Pi}$ have the same first- and second-order inclusion probabilities for a given vector $\Pi$.

The new second result, below, shows that the matrix $P^{\Pi}$ guarantees the lowest variability, among the determinantal designs, of the number of units sampled in the domains of the form $D^{k}=\{1, \ldots, k\}$ for $k=1, \ldots, N$. This optimality property of family $P^{\Pi}$ is one of those on which conjecture 6.1 will be based.

Theorem 3.1 Let $\Pi$ be a vector of $] 0,1\left[{ }^{N}\right.$ such that $\sum_{k=1}^{N} \Pi_{k}=n \in \mathbb{N}^{*}$ and $P^{\Pi}$, the Loonis and Mary matrix (2019). Let $P_{k}^{\Pi}$ be the main submatrix of order $k$ of $P^{\Pi}, k=1, \ldots, N$, and $\lambda^{k}=\left\{\lambda_{j}^{k}\right\}_{j=1}^{n}$, the vector of its strictly positive eigenvalues, supplemented where applicable by 0s to be of size $n$, then

1. $\lambda^{k}=\left\{\lambda_{j}^{k}\right\}_{j=1}^{n}$ is composed of the eigenvalue 1 with multiplicity $\left\lfloor\sum_{s=1}^{k} \Pi_{s}\right\rfloor$, the eigenvalue $\left\{\sum_{s=1}^{k} \Pi_{s}\right\}$ with multiplicity 1 and the eigenvalue 0 with multiplicity $n-\left\lfloor\sum_{s=1}^{k} \Pi_{s}\right\rfloor-1$, where $\lfloor x\rfloor$ and $\{x\}$ indicate the whole and decimal parts of $x$.
2. let $\mathbb{S} \sim \operatorname{DSD}(K)$ with inclusion probabilities given by $\Pi, D_{k}$ be the domain $D_{k}=\{1, \ldots, k\}$ and $\sharp \mathbb{S} \cap D_{k}$ be the random number of individuals of $\mathbb{S}$ that are in $D_{k}$, then, for every $k$.

$$
\underset{\text { s.c.diag }(K)=\Pi}{\operatorname{Min}} \operatorname{var}\left(\sharp \mathbb{S} \cap D_{k}\right)=\left\{\sum_{s=1}^{k} \Pi_{s}\right\}\left(1-\left\{\sum_{s=1}^{k} \Pi_{s}\right\}\right) .
$$

For every $k$, the minimum is reached, particularly for $K=P^{\Pi}$.

## 4. Constructing all determinantal sampling designs with fixed firstorder inclusion probabilities

### 4.1 Introduction

The method of Fickus et al. (2013) makes it possible to construct all Hermitian matrices with given diagonal and spectrum. For each of them, it provides an orthonormal eigenbasis. The method is described by the authors as non-trivial. The goal of this section is not to understand the whole process; we attempt to understand its broad strokes and identify its parameters and their constraints. We reformulate the latter to achieve parameterization involving mutually independent parameters. This section is technical and makes
it possible to justify and understand the notations of theorems 5.1 and 5.2 , which introduce a new property of matrices $P^{\Pi}$ and a new family of fixed-diagonal projection matrices, the coefficients of which are explicitly known.

Given our topic, we limit ourselves to the case of contracting Hermitian matrices, whose diagonal is $\Pi$ and spectrum, $\lambda^{N}$, is a known element of $\left.] 0,1\right]^{M}$, with $M \in \mathbb{N}^{*}$. The mobilization of $\lambda^{N}$ can seem less intuitive to statisticians than that of $\Pi$. In practice, provided that $\sum_{k=1}^{N} \Pi_{k}=n \in \mathbb{N}^{*}$, the natural choice will be $M=n$ and $\lambda^{N}=1^{n}$, which leads to orthogonal projection matrices and thus to fixed-size determinantal designs.

Remark 4.1 An important point is that the algorithm directly constructs only matrices whose diagonal terms are ordered from highest to lowest. It will be able to directly construct matrix (b) of example 3.1, but not matrix (a) of the same example. For the latter, the algorithm will provide the matrix of example 4.1. The two matrices are the same up to a permutation $\Sigma^{\top}$ of their rows and columns. The properties of the associated designs are the same, but for a differently sorted population.

Next, we assume that the population is sorted in a way that the inclusion probabilities are given by the vector $\Pi^{\triangleright}$. The matrices that we are attempting to create are notated $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$.

Example 4.1 The algorithm of Fickus et al. (2013) will be able to construct matrix $P^{\Pi}$ of example 3.1 up to a permutation. It will provide matrix $K^{\Pi^{\triangleright}}$ such that

$$
K^{\Pi^{\triangleright}}=\Sigma^{\Pi \top} P^{\Pi} \Sigma^{\Pi}=\left(\begin{array}{ccccccc}
\frac{4}{5} & 0 & 0 & -\frac{\sqrt{2}}{5} & 0 & \frac{2}{5 \sqrt{3}} & \frac{\sqrt{2}}{5 \sqrt{3}} \\
0 & \frac{3}{4} & -\frac{1}{4} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\
0 & -\frac{1}{4} & \frac{3}{4} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\
-\frac{\sqrt{2}}{5} & 0 & 0 & \frac{3}{5} & 0 & \frac{2 \sqrt{2}}{5 \sqrt{3}} & \frac{2}{5 \sqrt{3}} \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 \\
\frac{2}{5 \sqrt{3}} & 0 & 0 & \frac{2 \sqrt{2}}{5 \sqrt{3}} & 0 & \frac{2}{5} & \frac{\sqrt{2}}{5} \\
\frac{\sqrt{2}}{5 \sqrt{3}} & 0 & 0 & \frac{2}{5 \sqrt{3}} & 0 & \frac{\sqrt{2}}{5} & \frac{1}{5}
\end{array}\right),
$$

where $\Sigma^{\Pi^{\top}}$ is a permutation matrix, which transforms $\Pi$ into $\Pi^{\triangleright}$. If $\operatorname{DSD}\left(P^{\Pi}\right)$ applies to the population indexed by $(1,2,3,4,5,6,7)$, then $\operatorname{DSD}\left(K^{\Pi^{\triangleright}}\right)$ applies to the population in the order $(7,2,3,6,1,5,4)$ (or
$(7,3,2,6,1,5,4)$ because the value $3 / 4$ appears twice in the inclusion probabilities). The designs $\operatorname{DSD}\left(P^{\Pi}\right)$ and $\operatorname{DSD}\left(K^{\Pi^{\triangleright}}\right)$, applied to the same population, based on a tailored order, are equivalent.

### 4.2 Sequential construction of main submatrices

Fickus et al. (2013) sequentially create all square matrices $K_{k}^{M, \Pi^{\triangleright}, \lambda^{N}}$ of size $(k \times k), k=1, \ldots, N$, which are main submatrices of order $k$ of at least one matrix of type $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$. The algorithm is put into the form

$$
K_{1}^{M, \Pi^{\triangleright}, \lambda^{N}}=\left(\Pi_{1}^{\triangleright}\right), K_{k}^{M, \Pi^{\triangleright}, \lambda^{N}}=\left(\begin{array}{cc}
K_{k-1}^{M, \Pi^{\triangleright}, \lambda^{N}} & b_{k} \\
\bar{b}_{k}^{\top} & \Pi_{k}^{\triangleright}
\end{array}\right), k=2, \ldots, N,
$$

where $b_{k}$ is a vector of size $k-1$. This vector $b_{k}$ is constructed in such a way that the spectrum of $K_{k}^{M, \Pi^{\triangleright}, \lambda^{N}}$ is equal to $\lambda^{k}$, which is a construction parameter. The value of this parameter, chosen by the statistician, must be compatible with a process that ultimately yields, in $k=N$, a matrix $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$ with the initially desired diagonal $\Pi^{\triangleright}$ and spectrum $\lambda^{N}$.

At chosen $\lambda_{k}$, the authors show that various $b_{k}$ are possible. They therefore introduce a second matrix parameter: $V^{k-1}$, of size $(M \times M)$, which reflects the variability of the $b_{k}$. The index $k-1$ refers to the fact that the structure of matrix $V^{k-1}$ depends on $\lambda^{k-1}$. The nature of the constraints on the parameters is explained later. The way $b_{k}$ is derived from $\lambda_{k}$ and $V^{k-1}$ is detailed in the appendix (Section A.5). Example 4.2 shows how the algorithm works.

Example 4.2 (How the algorithm works) In example 4.1, we have $M=4=n, N=7$. Because matrix $K^{\Pi^{\triangleright}}$ is a projection matrix, it follows that $\lambda^{\top}=1^{4}$. According to Fickus et al. (2013), there are two sequences of multidimensional parameters $\left\{\lambda^{k}\right\}_{k=1}^{6}$ and $\left\{V^{k}\right\}_{k=1}^{6}$ that we do not intend to explain here, and that, in six steps, lead to the matrix renamed $K^{\Pi^{\triangleright}}=K_{7}^{4, \Pi^{\triangleright}, 1^{4}}$. The first steps are

Moving from step $k-1$ to $k$ depends, in practice, on parameters $\lambda^{k}$ and $V^{k-1}$. A different choice at one of the steps would lead to a different final matrix, but it would still be a projection matrix with diagonal $\Pi^{\triangleright}$.

Remark 4.2 In sampling theory, vector $\lambda^{k}$ parameterizes the variance of the sample size in the domain $D_{k}^{\triangleright}=\{1, \ldots, k\}$, based on (3.5). The term $b_{k}$, a function of $\lambda_{k}$ and $V^{k-1}$, is used in the expression of secondorder inclusion probabilities and, ultimately, in that of the variance of the Horvitz-Thompson estimator based on (3.7).

### 4.3 Reformulation of the constraints on vector parameters $\left\{\lambda^{k}\right\}_{k=1}^{N}$

For $1 \leq k \leq N-1$, every spectrum $\lambda^{k}$ is subject to two types of constraints. As a spectrum of a main submatrix of a Hermitian matrix, $\lambda^{k}$ is subject to the Cauchy interlacing constraints (2.1) with vectors $\lambda^{k-1}$ and $\lambda^{k+1}$. As a spectrum of a Hermitian matrix with diagonal $\left(\Pi_{1}^{\triangleright}, \ldots, \Pi_{k}^{\triangleright}\right)^{\top}, \lambda^{k}$ satisfies the Schur-Horn theorem (2.2). Fickus et al. (2013) show that this set of constraints will be respected if and only if, for $k=1, \ldots, N-1, j=1, \ldots, M, \lambda_{j}^{k} \in\left[A_{j}^{k}, B_{j}^{k}\right]$, where the formulae explaining $A_{j}^{k}$ and $B_{j}^{k}$ are provided in the appendix (equations [A.1] and [A.2], Section A.3).

For $k=N$, Fickus et al. (2013) consider that $\lambda^{N}$ is given exogenously. Noting that, for fixed $\Pi^{\triangleright}$, the only constraints imposed on $\lambda^{N}$ are those of the Schur-Horn theorem, we show that $\lambda^{N}$ will be the spectrum of a Hermitian matrix with diagonal $\Pi^{\triangleright}$ if and only if for $j=1, \ldots, M, \lambda_{j}^{N} \in\left[A_{j}^{N}, B_{j}^{N}\right]$, where the formulae explaining $A_{j}^{N}$ and $B_{j}^{N}$ are provided in the appendix (proposition A.1, Section A.4).

A feature of $A_{j}^{k}$ and $B_{j}^{k}$, for $k \leq N-1$ and $j>0$, is that these two bounds are a function only of $\Pi^{\triangleright}$, $\lambda^{k+1}$ and $\left\{\lambda_{s}^{k}\right\}_{s=1}^{j-1}$. Likewise, for $k=N$ and $j>0, A_{j}^{N}$ and $B_{j}^{N}$ depend only on $\Pi^{\triangleright}$ and $\left\{\lambda_{s}^{N}\right\}_{s=1}^{j-1}$, with the convention $\lambda_{0}^{k}=0$, for $k=1, \ldots, N$. These remarks introduce a new parameterization of all the eigenvalues used in the method of Fickus et al. (2013).

## Proposition 4.1

1. By setting $\lambda_{j}^{k}=A_{j}^{k}+\Omega_{j k}\left(B_{j}^{k}-A_{j}^{k}\right)$, with $\Omega_{j k} \in[0,1]$, all $\left\{\left\{\lambda^{k}\right\}_{k=1}^{k=N}\right\}$ of the spectra of the main submatrices of the contracting Hermitian matrices with diagonal $\Pi^{\triangleright}$, having $M$ strictly positive eigenvalues, can be parameterized by all the matrices of size $(M \times N)$, whose coefficients have value in $[0,1]$. Such a matrix is notated $\Omega$.
2. According to this parameterization, the eigenvalue $\lambda_{j}^{k}$ is a function only of $\Pi^{\triangleright}$ and

$$
\left\{\Omega_{1 N}, \ldots, \Omega_{M N}, \Omega_{1(N-1)}, \ldots, \Omega_{M(N-1)}, \ldots, \Omega_{1(k+1)}, \ldots, \Omega_{M(k+1)}, \Omega_{1 k}, \ldots, \Omega_{j k}\right\}
$$

3. The value $\Omega_{j k}=0$ (resp. $\Omega_{j k}=1$ ) leads to the smallest (resp. largest) possible value $\lambda_{j}^{k}$, conditional on

$$
\left\{\lambda_{1 N}, \ldots, \lambda_{M N}, \lambda_{1(N-1)}, \ldots, \lambda_{M(N-1)}, \ldots, \lambda_{1(k+1)}, \ldots, \lambda_{M(k+1)}, \lambda_{1 k}, \ldots, \lambda_{(j-1) k}\right\}
$$

4. The matrix $\Omega=0^{(M \times N)}$ (resp. $\Omega=1^{(M \times N)}$ ) leads to systematically retaining the smallest (resp. largest) possible eigenvalue $\lambda_{j}^{k}$, conditional on

$$
\left\{\lambda_{1 N}, \ldots, \lambda_{M N}, \lambda_{(N-1)}, \ldots, \lambda_{M(N-1)}, \ldots, \lambda_{1(k+1)}, \ldots, \lambda_{M(k+1)}, \lambda_{1 k}, \ldots, \lambda_{(j-1) k}\right\}
$$

5. If $\sum_{k=1}^{N} \Pi_{k}^{\triangleright}=M=n \in \mathbb{N}^{*}$, this parameterization will lead to a projection matrix for any matrix $\Omega: \forall \Omega: \lambda^{N}(\Omega)=1^{n}$.

An implementation of this parameterization appears in the appendix (Section A.3).

### 4.4 Reformulating the constraints on matrix parameters $\left\{V^{k}\right\}_{k=1}^{N-1}$

According to Fickus et al. (2013), the matrix $V^{k}$ can be arbitrarily chosen from the matrices that satisfy the following constraints:

1. $V^{k}$ is of size $(M \times M)$,
2. $V^{k}$ is a block diagonal,
3. The number of blocks of $V^{k}$ is equal to the number of distinct eigenvalues of $\lambda^{k}$,
4. A block's size is equal to the order of multiplicity of the corresponding eigenvalue,
5. Each block is a unitary matrix of some kind.

The parameterization of matrices $\left\{V^{k}\right\}_{k=1}^{N-1}$ occurs through the parameterization of the blocks that constitute them and, therefore, through the parameterization of unitary matrices. We did not identify any easily workable parameterization of this type of matrix in the literature (Dita, 1982, 1994; Jarlskog, 2005; Spengler, Huber, and Hiesmayr, 2010). We therefore propose a simplification. We assume that each element of $\lambda^{k}$ is of multiplicity 1 . As a result, matrix $V^{k}$ is unitary diagonal. Its $j^{\text {th }}$ diagonal term can be written $V_{j j}^{k}=\exp \left(2 i \pi \rho_{j k}\right)$, with $\rho_{j k}$ element of [0,1]. Thus, all matrices $\left\{V^{k}\right\}_{k=1}^{N-1}$ can be obtained from a matrix $\rho$ of size $\left(M \times(N-1)\right.$ ), whose coefficients are independent and have value in $[0,1]$. The $k^{\text {th }}$ column of $\rho$ is used to construct the diagonal of $V^{k}$.

Example 4.3 In example 4.2, the matrix $V^{2}$ is constructed from the spectrum $\lambda^{2}$ of matrix $K_{2}^{4, \Pi^{\triangleright},,^{4}}$, which has two strictly positive eigenvalues: $3 / 4$ and $4 / 5$. Because $M=4, \lambda^{2}$ is supplemented with 0 s so it is size 4 and $\lambda^{2}=(0,0,3 / 4,4 / 5)^{\top} . V^{2}$ has three blocks of respective sizes 2.1 and 1 , corresponding to the multiplicities seen in $\lambda^{2}$. Thus, we have

$$
V^{2}=\left(\begin{array}{llll}
a & b & 0 & 0 \\
c & d & 0 & 0 \\
0 & 0 & e & 0 \\
0 & 0 & 0 & f
\end{array}\right) \xrightarrow[\text { simplification }]{\rightarrow} V^{2}=\left(\begin{array}{cccc}
\exp \left(2 i \pi \rho_{12}\right) & 0 & 0 & 0 \\
0 & \exp \left(2 i \pi \rho_{22}\right) & 0 & 0 \\
0 & 0 & \exp \left(2 i \pi \rho_{32}\right) & 0 \\
0 & 0 & 0 & \exp \left(2 i \pi \rho_{42}\right)
\end{array}\right)
$$

The simplification is inconsequential because the block associated with the eigenvalue 0 is not, in practice, used in the calculations of Fickus et al. (2013). The simplification could be consequential if, for a $\lambda^{k}$, there were an eigenvalue of multiplicity greater than 1 in the interval $] 0,1[$. If $\Omega$ is random, with each coefficient following a uniform distribution on $[0,1]$, an intuition is that this event is of zero measurement. The values of the coefficients $\rho_{32}$ and $\rho_{42}$ that led to $K_{3}^{4, \Pi^{\triangleright}, 1^{4}}$ are not given. To find them, it would be necessary to use the reciprocal of Fickus et al. (2013), the principle of which is presented in the longer version of the article.

Remark 4.3 In the algorithm of Fickus et al. (2013), the nature of $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$, complex or real, depends on the choice of $V^{k}$. If at least one matrix $V^{k}$ is complex, $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$ will be complex. Conversely, $K_{N}^{M, \Pi^{\triangleright}, \lambda^{N}}$ will be real if and only if all matrices $V^{k}$ are real. Based on our parameterization, this case corresponds to a choice of $\rho$ with value in $\left\{0 ; \frac{1}{2} ; 1\right\}^{(M \times(N-1))}$.

Remark 4.4 In the following, the notation $K_{N}^{\Pi^{\triangleright}}(M, \Omega, \rho)$ will refer to a Hermitian matrix with diagonal $\Pi^{\triangleright}$ and constructed according to the method of Fickus et al. (2013) with the parameters $(M, \Omega, \rho)$. The reference to $\lambda^{N}$ is omitted because this quantity is considered a parameter in the same way as the other spectra.

## 5. Matrices $K_{N}^{\Pi^{\triangleright}}(M, \Omega, \rho)$ for specific values of $\Omega$

### 5.1 Where we get back to $\boldsymbol{P}^{\Pi^{\triangleright}}$

The construction of $P^{\Pi}$, in Loonis and Mary (2019), relies the sequential modification of the columns of a matrix of size $(M(=n) \times N)$ using rotation matrices, which are also unitary matrices (algorithm 3.1). This approach is consistent with the spirit of the Fickus et al. method. (2013). However, the formalism of the two methods does not appear to be directly equivalent. In the theorem below, we specify some links between the two approaches.

Theorem 5.1 Let $\Pi^{\triangleright}$ be a vector of size $N$, such that $0<\Pi_{k}^{\triangleright}<1, \sum_{k=1}^{N} \Pi_{k}^{\triangleright}=n \in \mathbb{N}^{*}$ and $\Pi_{1}^{\triangleright} \geq \ldots \Pi_{k}^{\triangleright} \geq \ldots \geq$ $\Pi_{N}^{\triangleright}$. For any value of the parameter $\rho$,

$$
K_{N}^{\Pi^{\triangleright}}\left(M=n, \Omega=0^{(n \times N)}, \rho\right)=H^{N}(\rho) \odot P^{\Pi^{\triangleright}},
$$

where $K_{N}^{\Pi^{\triangleright}}\left(M=n, \Omega=0^{n \times N}, \rho\right)$ is constructed based on the method of Fickus et al. (2013), $P^{\Pi^{\triangleright}}$ is the matrix defined by Loonis and Mary (2019) from the vector $\Pi^{\triangleright}$ and $H^{N}(\rho)$ is an Hermitian matrix of size $(N \times N)$ where all diagonal terms equal 1 and non-diagonal terms have a modulus of 1 .

For point (b) in example 3.1, this theorem indicates that the method of Fickus et al. (2013), with parameters $\left(M=n=4, \Omega=0^{(4 \times 10)}\right)$, directly provides the matrix $P^{\Pi^{\triangleright}}$, up to $H_{N}$ for all $\rho$. According to this theorem, the modulus of the off-diagonal terms of $K^{\Pi^{\triangleright}}\left(M=n, \Omega=0^{n \times N}, \rho\right)$ does not depend on $\rho$ because they equal $P_{k l}^{\Pi^{\triangleright}} H_{k l}^{N}(\rho) \overline{H_{k l}^{N}(\rho)} P_{k l}^{\Pi^{\triangleright}}=\left(P_{k l}^{\Pi^{\triangleright}}\right)^{2}$. In terms of surveys, the variance of $\hat{t}_{y}$ is dependent on $\Delta_{k l}$ and thus on the modulus of the $K_{k l}^{\Pi^{\triangleright}}$ when the design is determinantal, according to (3.7). The implication is that there is nothing to expect from changing the parameter $\rho$ to change the variance obtained from $P^{\Pi^{\triangleright}}$. This is a very special feature among Hermitian matrices, and among matrices of the family $P^{\Pi}$. When $\Omega$ is fixed, different from $0^{(M \times N)}$, it is empirically observed that the variance of the estimators is usually affected by variations of $\rho$.

Remark 5.1 For a matrix $P^{\Pi}$, constructed from any vector $\Pi$, in other words, whose coefficients are not necessarily sorted, there is a matrix $\Omega^{\Pi}$ that leads to $\Sigma^{\Pi^{\top}} P^{\Pi} \Sigma^{\Pi}$ with the method of Fickus et al. (2013). There is no reason, a priori, for $\Omega^{\Pi}$ to be of the form $0^{(n \times N)}$. This is particularly true for the matrix $P^{\Pi}$ of point (a) in example 3.1 and its reordered version given in example 4.1.

### 5.2 Where we discover $\boldsymbol{Q}^{{ }^{\square}}$

The theorem below shows that it is also possible to find an explicit formulation of $K^{\Pi^{\triangleright}}(M=$ $\left.n, \Omega=1^{(n \times N)}, \rho\right)$, except for a few coefficients.

Theorem 5.2 Let $\Pi^{\triangleright}$ be a vector of size $N$, such that $0<\Pi_{k}^{\triangleright}<1, \sum_{k=1}^{N} \Pi_{k}^{\triangleright}=n \in \mathbb{N}^{*}$ and $\Pi_{1}^{\triangleright} \geq \ldots \Pi_{k}^{\triangleright} \geq \ldots \geq$ $\Pi_{N}^{\triangleright}$. Let $\left\{\left\{\lambda_{j}^{k}\right\}_{j=1}^{n}\right\}_{k=n}^{N}$ be a sequence of reals such that $\lambda^{N}=1^{n}$ and

$$
\left\{\begin{array}{lll}
\lambda_{1}^{k}=\lambda_{n}^{k+1}-\Pi_{k+1}^{\triangleright} & k=n, \ldots, N-1,  \tag{5.1}\\
\lambda_{j}^{k}=\lambda_{j-1}^{k+1}, & j=2, \ldots, n, k=n, \ldots, N-1 .
\end{array}\right.
$$

It is assumed that $\Pi^{\triangleright}$ is such that one of the following two conditions is satisfied:

- C1: only 0 and 1 can appear multiple times in the $\lambda^{k}, k=n, \ldots, N$.
- $\quad C 2: \Pi_{k}=n / N, k=1, \ldots, N$.

For any value of the parameter $\rho$, we then have

$$
K^{\Pi^{\triangleright}}\left(M=n, \Omega=1^{n \times N}, \rho\right)=H^{N}(\rho) \odot Q^{\Pi^{\triangleright}}
$$

where $H^{N}(\rho)$ is a Hermitian matrix of size $(N \times N)$, where all diagonal coefficients equal 1 and nondiagonal coefficients have a modulus of 1. The matrix $Q^{\Pi^{\triangleright}}$ is such that

1. for $k \geq n, \lambda^{k}$ yields the $n$ largest eigenvalues of the main submatrix of order $k$ of $Q^{\Pi^{\triangleright}}$, for which no more than $n$ are strictly positive
2. for any $(k, l) \in[n+1, N]^{2}$,
(a) under $C 1: Q_{k l}^{\Pi^{\triangleright}}=\sqrt{\Pi_{k}^{\triangleright} \Pi_{l}^{\triangleright}} 1(l \equiv k \bmod n)$
(b) under $C 2$ : $Q_{k l}^{\Pi}=n / N 1(l \equiv k$ ou $k-N \equiv N-(l+1) \bmod 2 n)$
3. under C2, if $n$ divides $N$, the formulae (5.1) and point 1 are true for $0<k<n$, the point 2 (b) is true for $k \leq n$ or $l \leq n$.

Figure 5.1 shows an example of matrix $Q^{\Pi^{\triangleright}}$ under $C 1$. Table 5.1 shows the organization of the spectra of the main submatrices of $Q^{\Pi^{\triangleright}}$ based on $\Pi^{\triangleright}$. The equivalents for $C 2$, including in the case $n$ divides $N$, are provided in the appendix (Section A.7).

For $C 1$, the constraint on the multiplicities, set out in the theorem, implies that in each column there are different values except, possibly, for the eigenvalues 0 and 1 . If that is not the case, inextricable calculation difficulties arise. When $k<n$, the expression of the eigenvalues is more difficult to find. As a result, we do not arrive at an explicit formula of some coefficients of the matrices $Q^{\Pi^{\triangleright}}$ as a function of $\Pi^{\triangleright}$, except under $C 2$ and $n$ divides $N$.

Figure 5.1 $Q^{\Pi^{\triangleright}}$ under $C 1$ for $n=2, N=8$.


The symbol • indicates that the explicit formula of the coefficient in question is unknown.

Table 5.1
Eigenvalues $\lambda_{j}^{k}$ of the main submatrices of $Q^{\Pi^{\triangleright}}$, for $n=M=5, k=N-11, \ldots, N$.

| $\boldsymbol{k}$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{N - 1 1}$ | $\mathbf{N - 1 0}$ | $\mathbf{N}-\mathbf{9}$ | $\mathbf{N - \mathbf { 8 }}$ | $\mathbf{N} \mathbf{- 7}$ | $\mathbf{N - 6}$ |
| $1-\Pi_{N-10}^{\triangleright}-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ | $1-\Pi_{N-9}^{\triangleright}-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-8}^{\triangleright}-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-7}^{\triangleright}-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-6}^{\triangleright}-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ |
| $1-\Pi_{N-9}^{\triangleright}-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-8}^{\triangleright}-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-7}^{\triangleright}-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-6}^{\triangleright}-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ | $1-\Pi_{N-4}^{\triangleright}$ |
| $1-\Pi_{N-8}^{\triangleright}-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-7}^{\triangleright}-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-6}^{\triangleright}-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ | $1-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-3}^{\triangleright}$ |
| $1-\Pi_{N-7}^{\triangleright}-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-6}^{\triangleright}-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ | $1-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-2}^{\triangleright}$ |
| $1-\Pi_{N-6}^{\triangleright}-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N-5}^{\triangleright}-\Pi_{N}^{\triangleright}$ | $1-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-1}^{\triangleright}$ |
| $\mathbf{N - 5}$ | $\mathbf{N}-\mathbf{4}$ | $\mathbf{N - 3}$ | $\mathbf{N - 2}$ | $\mathbf{N} \mathbf{- 1}$ | $\mathbf{N}$ |
| $1-\Pi_{N-4}^{\triangleright}$ | $1-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N}^{\triangleright}$ | 1 |
| $1-\Pi_{N-3}^{\triangleright}$ | $1-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N}^{\triangleright}$ | 1 | 1 |
| $1-\Pi_{N-2}^{\triangleright}$ | $1-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N}^{\triangleright}$ | 1 | 1 | 1 |
| $1-\Pi_{N-1}^{\triangleright}$ | $1-\Pi_{N}^{\triangleright}$ | 1 | 1 | 1 | 1 |
| $1-\Pi_{N}^{\triangleright}$ | 1 | 1 | 1 | 1 | 1 |

## 6. Applications

### 6.1 The data

The samples from Insee (Institut national de la statistique et des études économiques) household surveys are usually selected based on a two-stage design, when the collection method is face-to-face. In the first stage, primary units (PUs) are selected, from which the households to survey are then selected. The PUs consist of groups of the smallest nearby municipalities, with at least 2,000 principal residences. The sampling of PUs is stratified according to the 22 former metropolitan regions. Here we are interested in the sole first-stage selection of the PUs in the Provence-Alpes-Côte d'Azur region. They total $N=148$ and are
described by auxiliary variables yielding, for each PU, the total population by sex, age, the total amount of certain incomes or the total number of dwellings by category (vacant, secondary residence, etc.). Inclusion probabilities are proportional to the number of principal residences of the PUs.

### 6.2 A lower bound for the variance of the estimators in the determinantal situation?

When a sampling design is of a fixed size, the variance of the estimators, given by (3.6), takes the following particular form:

$$
\begin{equation*}
\operatorname{var}\left(\hat{t}_{y}\right)=-\frac{1}{2} \sum_{k} \sum_{l \neq k}\left(\frac{y_{k}}{\pi_{k}}-\frac{y_{l}}{\pi_{l}}\right)^{2} \Delta_{k l} . \tag{6.1}
\end{equation*}
$$

In the determinantal case, (3.7) becomes

$$
\begin{equation*}
\operatorname{var}\left(\hat{t}_{y}\right)=\frac{1}{2} \sum_{k} \sum_{l \neq k}\left(\frac{y_{k}}{K_{k k}}-\frac{y_{l}}{K_{l l}}\right)^{2}\left|K_{k l}\right|^{2} . \tag{6.2}
\end{equation*}
$$

The determinantal variance will be low if the $K_{k l}$ modulus is low for the values of $y_{k} / K_{k k}$ distant from those of $y_{l} / K_{l l}$. Assuming that the population is sorted according to $y_{k} / \Pi_{k}=y_{k} / K_{k k}$, this situation is observed for the matrix $P^{\Pi}$ because we have, for $|k-l|$ large, $\pi_{k l} \simeq \Pi_{k} \Pi_{l}$ and therefore $\left|K_{k l}\right|^{2}=\Pi_{k} \Pi_{l}-$ $\pi_{k l} \simeq 0$ (see Section 3.4.2). Empirical results from Loonis and Mary (2019) suggest that matrices $P^{\Pi}$, constructed on a population sorted by $y_{k} / \Pi_{k}$, perform well in terms of variance. Theorems 3.1 and 5.1 also show that matrices $P^{\Pi}$ have special properties among all fixed-diagonal projection matrices, even though, in the case of theorem 5.1, the scope is limited to diagonals of the form $\Pi^{\triangleright}$.

In this section, we characterize the performance of the designs associated with $P^{\Pi}$ among all fixed-size determinantal designs and inclusion probabilities $\Pi$. We limit ourselves to the case of real kernels. For this, we conduct the following experiment:

- We create nested subpopulations of size $N=20,40,60,80,100,120$ associated with samples of size $n=3,6,9,12,15,18$.
- We consider three auxiliary variables, representing respectively the total amount of salaries, the number of tenants and the number of homeowners.
- For each variable, population and sample size,
- we minimize (6.2), in $K$, among the real projection matrices with diagonal $\Pi$. We use optimization algorithms on manifolds (Absil et al., 2009; Boumal et al., 2014). We succinctly present the main principles of these procedures in the Appendix (Section A.8).
- we calculate (6.2) for a matrix $K_{x^{9}}=P_{x^{4}}^{\Pi}$, where $P_{x^{q}}^{\Pi}$ was constructed on a population that was previously sorted by the values of $x_{k}^{q} / \Pi_{k}$.

Figure 6.1 shows that the scatterplot obtained by crossing the previous two values, for each variable and sample size, is aligned with the first bisector. From this, we empirically deduce that the variance obtained with $P_{x^{4}}^{\Pi}$ corresponds to a minimum of the variance of $\hat{x}_{x^{9}}$, among determinantal designs. The proof of this result, or its refutation, seems to be out of reach. We therefore propose conjecture 6.1 below.

Conjecture 6.1 Let y be a positive variable and $\Pi$ be a vector of $] 0,1\left[{ }^{N}\right.$ such that $\sum_{k=1}^{N} \Pi_{k}=n \in \mathbb{N}^{*}$. Let $\mathcal{P}$ be a determinantal sampling design with inclusion probabilities $\pi_{k}=\Pi_{k}, k=1, \ldots, N$ and $\hat{t}_{y}$ be the Horvitz-Thompson estimator of the total $t_{y}$ of $y$ on that design. Without losing generality, $\Pi$ and $y$ are such that $y_{1} / \Pi_{1} \leq \ldots \leq y_{k} / \Pi_{k} \leq \ldots \leq y_{N} / \Pi_{N}$, then

$$
\operatorname{var}\left(\hat{t}_{y}\right) \geq y^{\top} D_{\Pi}^{-1}\left[\left(I_{N}-P^{\Pi}\right) \odot P^{\Pi}\right] D_{\Pi}^{-1} y,
$$

where $y^{\top} D_{\Pi}^{-1}\left[\left(I_{N}-P^{\Pi}\right) \odot P^{\Pi}\right] D_{\Pi}^{-1} y$ is the variance of $\hat{t}_{y}$ obtained with design $\operatorname{DSD}\left(P^{\Pi}\right)$, with $P^{\Pi}$ matrix of Loonis and Mary (2019).

Figure 6.1 Comparison, for three auxiliary variables and six population sizes, of the variation coefficients for variances obtained by minimizing (6.2) into $K$ and for a matrix $P_{x^{q}}^{\Pi}$ sorted based on the values of $\boldsymbol{x}_{k}^{q} / \Pi_{k}$.


In the classic configuration of a fixed-size design with constant inclusion probabilities, the previous conjecture makes it possible to find intuitive interpretations, when $n$ divides $N$.

According to Section 3.4.2, if $\Pi_{k}=n / N$ and if $n$ divides $N, P^{\Pi}$ is a block diagonal with $n$ identical blocks. The coefficients of each block of size $(N / n \times N / n)$ all equal $n / N$. The associated design involves selecting 1 individual in each of the $n$ strata of size $N / n$. After sorting based on the values of $y$, the first stratum $U_{1}$ combines the first $N / n$ individuals, the second stratum $U_{2}$ combines the next $N / n$ individuals,
and so on. Applying the variance formulae of the Horvitz-Thompson estimator for a stratified design (Särndal, Swensson, \& Wretman, 2003) leads to the lower bound, in this case, to be equal to

$$
y^{\top} D_{\Pi}^{-1}\left[\left(I_{N}-P^{\Pi}\right) \odot P^{\Pi}\right] D_{\Pi}^{-1} y=\frac{N^{2}}{n^{2}}\left(1-\frac{n}{N}\right) \sum_{h=1}^{n} S_{y U_{h}}^{2},
$$

where $S_{y U_{h}}^{2}=N /(N-n) \sigma_{y U_{h}}^{2}$ and $\sigma_{y U_{h}}^{2}$ is the variance of $y$ in stratum $U_{h}$.
If the total you want to estimate is $N_{c}$, the number of individuals who have a given characteristic $c$, the underlying variable $y$ equals 1 if the individual has the characteristic, otherwise it equals 0 . After sorting this population based on this variable, in $n-1$ strata, the variance of $y$ will be zero because $y$ will still be 0 or 1. In a single stratum, there will be $r_{c} 1$ values, where $r_{c}$ is the remainder of the Euclidean division of $N_{c}$ by $N / n$, and $N / n-r_{c} 0$ values. The lower bound is then $r_{c}\left(N / n-r_{c}\right)$.

## 7. Conclusion

In this article, we propose a workable parameterization of the kernels of determinantal sampling designs. We show that the family $P^{\Pi}$, originally constructed by Loonis and Mary (2019) with the sole objective of having an example of a fixed-sized determinantal design with given first-order inclusion probabilities, turns out to have unexpected statistical properties. For any $k=1, \ldots, N$, it minimizes the variability of the sample size that falls within the fields $D^{k}=\{1, \ldots, k\}$ (theorem 3.1). It is directly associated with a very particular value of the parameter $\Omega$ in the construction of Fickus et al. (2013) (theorem 5.1). Finally, it appears empirically as a solution to a problem of minimizing the variance of the Horvitz-Thompson estimator (Section 6.2). This ubiquity leads to the conjecture that the lower bound of the variance of the HorvitzThompson estimator, among determinantal designs, is expressed as a function of $P^{\Pi}$ (conjecture 6.1).

These results are obtained at the cost of theoretical or computational complexity, which can seem stimulating or bewildering. Bridges have been built between sampling theory and other fields, such as probabilities, algebra or semidefinite optimization. These bridges can be a source of new theoretical studies. However, the concepts and methods used are not part of the ordinary toolbox of survey statisticians of public institutes, even of the author initially. One of the challenges for developing studies around determinantal designs therefore lies in education, which can be facilitated through the provision of $R$ programs that make the various results tangible. Another possibility is, in the future, to focus on more practical applications. As such, indirect sampling, the search for alternatives to the Horvitz-Thompson estimator, or spatial sampling appear to be promising areas.

## Acknowledgements

The author thanks Alina Matei and two anonymous referees for the relevance of their comments, which led to substantial improvements in this version of the article. Any remaining errors, omissions, typos and blunders are the responsibility of the author.

## Appendix

## A. 1 Determinantal designs and population order

Matters of population order often arise in the article. Below, we provide a summary of the various findings, as well as some elements to assist in their interpretation. We first reiterate that exchanging two rows (or two columns) of a square matrix leads to multiplying its determinant by -1 .

1. Determinantal designs are not dependent on population order. We consider a population $U=\{1, \ldots, N\}$ and a determinantal design $\operatorname{DSD}(K)$. Sorting $U$ involves applying a permutation to it, the matrix of which is notated $\Sigma^{\top}$. The design $\operatorname{DSD}(K)$ on $U$ is then equivalent to the design $\operatorname{DSD}\left(\Sigma^{\top} K \Sigma\right)$ on the sorted population. This property results from the fact that an even number of exchanges of rows and columns are applied to the matrix $K$. Therefore, the determinants of the extracted matrices are not changed.
2. An algorithm for constructing particular kernels can depend on the order of the units. This is true for matrices $P^{\Pi}$. Two different population orders will lead to two matrices $P^{\Pi_{1}}$ and $P^{\Pi_{2}}$. In general, there will be no permutation matrix such that $P^{\Pi_{1}}=\Sigma^{\top} P^{\Pi_{2}} \Sigma$. The designs $\operatorname{DSD}\left(P^{\Pi_{1}}\right)$ and $\operatorname{DSD}\left(P^{\Pi_{2}}\right)$ are not equivalent. However, the previous remark applies to each of them.
3. The algorithm by Fickus et al. (2013) is not dependent on the order. Let $K^{\Pi}$ be a contracting Hermitian matrix. For numerical reasons, the algorithm of Fickus et al. (2013) will construct the matrix $K^{\Pi^{\triangleright}}=\Sigma^{\top} K^{\Pi} \Sigma$, where $\Sigma^{\top}$ transforms $\Pi$ into $\Pi^{\triangleright}$. To find the original matrix, simply take $K^{\Pi}=\Sigma K^{\Pi^{\triangleright}} \Sigma^{\top}$.
4. The connection between the algorithm of Fickus et al. (2013) and the family $P^{\Pi}$ arises from the fact that $P^{\Pi^{\triangleright}}=K\left(n, 0^{(n \times N)}, \rho\right)$, up to matrix $H_{N}(\rho)$. Theorem 5.1 does not say that there is $P^{\Pi}=\Sigma K^{\Pi^{\triangleright}}\left(n, 0^{(n \times N)}, \rho\right) \Sigma^{\top}$, for any $\Pi$, even up to matrix $H_{N}$. However, there are indeed many unknown parameters such that $P^{\Pi}=\Sigma K^{\Pi^{\triangleright}}\left(n, \Omega^{\Pi}, \rho^{\Pi}\right) \Sigma^{\top}$.
5. The lower bound of conjecture 6.1 is a matrix $P^{\Pi}$ constructed on a population sorted in such a way that $y_{1} / \Pi_{1} \leq \ldots \leq y_{k} / \Pi_{k} \leq \ldots \leq y_{N} / \Pi_{N}$.

## A. 2 Programs

The applications and simulations in this article were mostly obtained using SAS programs. We are gradually making their retranscription in $R$ available, experimentally, on the Insee Lab at
https://github.com/InseeFrLab/Determinantal-Sampling-Designs.
In particular, these will gradually be made available:

1. The matrix $P^{\Pi}$ and an eigenvector base constructed with algorithm 3.1
2. The matrix $K^{\Pi^{\triangleright}}(M, \Omega, \rho)$ constructed following the method described in Section A. 5 and an eigenvector base as obtained at the very end of algorithm A. 2
3. The selection algorithm of Lavancier et al. (2015) for selecting samples based on the determinantal designs associated with the kernels defined in the previous points.

There will be programs for points not covered here, such as the reciprocal of (Fickus et al., 2013) or the construction of periodic determinantal designs with constant inclusion probabilities, such as described in Loonis and Mary (2019). A longer version of this article, which presents the determinantal designs in more detail and provides the proofs of algorithm 3.1, theorems 3.1, 5.1, 5.2, and proposition A.1, is available at the same address.

## A. 3 Constructing the sequence $\left\{\lambda^{k}\right\}_{k=1}^{k=N-1}$ from $\Omega$

According to Fickus et al. (2013), $\lambda_{j}^{k}$ can be arbitrarily chosen in the interval $\left[A_{j}^{k}, B_{j}^{k}\right]$, with

$$
\begin{gather*}
A_{j}^{k}=\max \left\{\lambda_{j-1}^{k+1}, \sum_{s=1}^{j} \lambda_{s}^{k+1}-\sum_{s=1}^{j-1} \lambda_{s}^{k}-\Pi_{k+1}^{\triangleright}\right\},  \tag{A.1}\\
B_{j}^{k}=\min \left\{\lambda_{j}^{k+1}, \min _{i=j, \ldots, M}\left\{\sum_{s=M-i+1}^{k} \Pi_{s}^{\triangleright}-\sum_{s=j}^{i-1} \lambda_{s}^{k+1}-\sum_{s=1}^{j-1} \lambda_{s}^{k}\right\}\right\}, \tag{A.2}
\end{gather*}
$$

for $k=1, \ldots, N-1, j=1, \ldots, M$ and where $\lambda_{0}^{k}=0$ and, by convention, the sums over empty sets equal 0 .
Here we describe how to construct the sequence $\left\{\lambda^{k}\right\}_{k=1}^{k=N-1}$ from the formulae (A.1) and (A.2) and proposition 4.1. The vector $\Pi^{\triangleright}$, of size $N=10$, is such that $\Pi_{k}^{\triangleright}=\frac{6(11-k)}{10^{\triangleright} 11}, k=1, \ldots, 10$ and $\sum_{k=1}^{N} \Pi_{k}^{\triangleright}=3$. We choose $M=7$ and set the value of $\lambda^{10}$ at $\left(\frac{3}{10}, \frac{3}{10}, \frac{3}{10} \frac{3}{10}, \frac{6}{10} \frac{6}{10} \frac{6}{10}\right)^{\top}$. The matrix $\Omega$ is of size $(7 \times 10)$. We choose to have all its coefficients equal 0.5 for the first nine columns, which are the ones used in the calculations of $\left\{\lambda^{k}\right\}_{k=1}^{9}$. Table A. 1 gives the final result for this sequence of eigenvalues. We then show how to arrive at the particular value $357 / 704$ for $\lambda_{5}^{5}$. According to proposition 4.1, this value results from the calculation $A_{5}^{5}+\Omega_{5,5}\left(B_{5}^{5}-A_{5}^{5}\right)=0,5\left(A_{5}^{5}+B_{5}^{5}\right)$ with $A_{5}^{5}=219 / 400$ and $B_{5}^{5}=909 / 1,760$. Table A. 2 shows how $A_{5}^{5}=219 / 400$ is arrived at by applying formula (A.1). Table A. 3 shows how to arrive at $B_{5}^{5}=909 / 1,760$ by applying formula (A.2). Since the logic is the same, we don't show how to apply the formulae from proposition (A.1), which would make it possible to construct $\lambda^{10}$ from the 10th column of $\Omega$. However, we specify that the values in this latter column do not necessarily equal 0.5 .

Table A. 1
Final values of $\lambda_{j}^{k}$.

|  |  | $\boldsymbol{k}$ |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{j}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\mathbf{8}$ | $\mathbf{9}$ | $\mathbf{1 0}$ |
| $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | $\frac{3}{22}$ | $\frac{12}{55}$ | $\frac{3}{11}$ | $\frac{3}{10}$ |
| 2 | 0 | 0 | 0 | 0 | 0 | $\frac{93}{440}$ | $\frac{111}{440}$ | $\frac{63}{220}$ | $\frac{3}{10}$ | $\frac{3}{10}$ |
| 3 | 0 | 0 | 0 | 0 | $\frac{213}{880}$ | $\frac{3}{11}$ | $\frac{129}{440}$ | $\frac{3}{10}$ | $\frac{3}{10}$ | $\frac{3}{10}$ |
| 4 | 0 | 0 | 0 | $\frac{927}{3,520}$ | $\frac{501}{1,760}$ | $\frac{261}{880}$ | $\frac{3}{10}$ | $\frac{3}{10}$ | $\frac{3}{10}$ | $\frac{3}{10}$ |
| 5 | 0 | 0 | $\frac{5,583}{14,080}$ | $\frac{669}{1,408}$ | $\frac{357}{704}$ | $\frac{909}{1,760}$ | $\frac{21}{40}$ | $\frac{6}{11}$ | $\frac{63}{110}$ | $\frac{3}{5}$ |
| 6 | 0 | $\frac{13,623}{28,160}$ | $\frac{909}{1,760}$ | $\frac{477}{880}$ | $\frac{3,969}{7,040}$ | $\frac{2,001}{3,520}$ | $\frac{63}{110}$ | $\frac{129}{220}$ | $\frac{3}{5}$ | $\frac{3}{5}$ |
| 7 | $\frac{6}{11}$ | $\frac{15,561}{28,160}$ | $\frac{7,881}{14,080}$ | $\frac{4,041}{7,040}$ | $\frac{4,113}{7,040}$ | $\frac{2,073}{3,520}$ | $\frac{261}{440}$ | $\frac{3}{5}$ | $\frac{3}{5}$ | $\frac{3}{5}$ |
| Total | $\frac{6}{11}$ | $\frac{57}{55}$ | $\frac{81}{55}$ | $\frac{102}{55}$ | $\frac{24}{11}$ | $\frac{27}{11}$ | $\frac{147}{55}$ | $\frac{156}{55}$ | $\frac{162}{55}$ | 3 |
| $\Pi_{k}^{\triangleright}$ | $\frac{6}{11}$ | $\frac{27}{55}$ | $\frac{24}{55}$ | $\frac{21}{55}$ | $\frac{18}{55}$ | $\frac{3}{11}$ | $\frac{12}{55}$ | $\frac{9}{55}$ | $\frac{6}{55}$ | $\frac{3}{55}$ |

Table A. 2
How to obtain $\lambda_{5}^{5}=\frac{1}{2}\left(A_{5}^{5}+B_{5}^{5}\right)=\frac{357}{704}$ in Table A. 1 calculating $A_{5}^{5}$ with formula (A.1).

| $\lambda_{4}^{6}$ | $\sum_{s=1}^{5} \lambda_{s}^{6}$ | $\sum_{s=1}^{4} \lambda_{s}^{5}$ | $\Pi_{6}^{\triangleright}$ | b-c-d | $\boldsymbol{A}_{5}^{5}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| (a) | (b) | (c) | (d) | (e) | $\boldsymbol{\operatorname { m a x } ( \boldsymbol { a } , \boldsymbol { e } )}$ |
| $\frac{261}{880}$ | $\frac{93}{440}+\frac{3}{11}+\frac{261}{880}+\frac{909}{1,760}$ | $\frac{213}{880}+\frac{501}{1,760}$ | $\frac{3}{11}$ | $\frac{219}{440}$ | $\frac{219}{440}$ |

Table A. 3
How to obtain $\lambda_{5}^{5}=\frac{1}{2}\left(A_{5}^{5}+B_{5}^{5}\right)=\frac{357}{704}$ in Table A. 1 calculating $B_{5}^{5}$ with formula (A.2).

| $\boldsymbol{i}$ | $\lambda_{5}^{6}$ | $\sum_{s=8-i}^{5} \Pi_{s}^{\triangleright}$ | $\sum_{s=5}^{i-1} \lambda_{s}^{6}$ | $\mathbf{g - h - c}$ |
| :---: | :---: | :---: | :---: | :---: |
| 5 | $\frac{909}{1,760}$ | $\frac{24}{55}+\frac{21}{55}+\frac{18}{55}$ | 0 | $\frac{99}{160}$ |
| 6 | $\frac{909}{1,760}$ | $\frac{27}{55}+\frac{24}{55}+\frac{21}{55}+\frac{18}{55}$ | $\frac{909}{1,760}$ | $\frac{261}{440}$ |
| 7 | $\frac{909}{1,760}$ | $\frac{6}{11}+\frac{27}{55}+\frac{24}{55}+\frac{21}{55}+\frac{18}{55}$ | $\frac{909}{1,760}+\frac{2,001}{3,520}$ | $\frac{2,007}{3,520}$ |
| $B_{5}^{5}=\min \left(\frac{909}{1,760}, \frac{2,007}{3,520}, \frac{261}{440}, \frac{99}{160}\right)=\frac{909}{1,760}, \lambda_{5}^{5}=\frac{1}{2}\left(\frac{219}{440}+\frac{909}{1,760}\right)=\frac{357}{704}$ |  |  |  |  |

## A. 4 Expression of the constraints on $\lambda^{N}$

Here we give an expression of the constraints applying to $\lambda^{N}$ in the same form as those applying to $\left\{\lambda^{k}\right\}_{k=1}^{N-1}$. The difference is that $\lambda^{N}$ is subject only to the Schur-Horn theorem.

Proposition A. 1 Let $\Pi^{\triangleright}$ be a vector of $] 0,1\left[{ }^{N}\right.$ such that $\Pi_{1}^{\triangleright} \geq \ldots \geq \Pi_{k}^{\triangleright} \geq \ldots \geq \Pi_{N}^{\triangleright}$ and $\sum_{k=1}^{N} \Pi_{k}^{\triangleright}=\mu$. Let $M$ be an integer such that $\mu \leq M \leq N$ and $\lambda^{N}$ be a vector of $\left.] 0,1\right]^{M}$. There is a positive semidefinite Hermitian matrix with diagonal $\Pi^{\triangleright}$ and whose strictly positive eigenvalues are given by $\lambda^{N}$ if and only if

$$
\left\{\begin{array}{l}
\lambda_{j}^{N} \in\left[A_{j}^{N}, B_{j}^{N}\right], j=1, \ldots, M  \tag{A.3}\\
A_{j}^{N}=\max \left\{\lambda_{j-1}^{N}, \mu-\sum_{s=1}^{j-1} \lambda_{s}^{N}-(M-j)\right\} \\
B_{j}^{N}=\min _{i=1, \ldots, M-j+1}\left\{\frac{\mu-\sum_{s=1}^{j-1} \lambda_{s}^{N}-\sum_{s=1}^{M-j-i+1} \Pi_{s}^{\triangleright}}{i}\right\},
\end{array}\right.
$$

where $\lambda_{0}^{N}=0$ and, by convention, the sums over empty sets equal 0.

## A. 5 The method by Fickus et al. (2013)

The general method is based on two subalgorithms for constructing the various quantities used in equations (A.7) and (A.8). For a given vector $\Pi^{\triangleright}$, it is assumed that values were set for the parameters $M, \Omega, \rho$, which made it possible to construct the sequences $\left\{\lambda^{k}\right\}_{k=1}^{N}$ and $\left\{V^{k}\right\}_{k=1}^{N-1}$. In its original form, the method of Fickus et al. (2013) also depends on any unitary matrix $U^{1}$ of size $(M \times M)$. This matrix does not directly influence the final matrices. It influences only the choice of one of their eigenvector bases. $U^{1}=I_{M}$ will be taken in practice.

## Algorithm A. 1 (Fickus et al. [2013])

For $k=2$ to $N$,

1. set $E_{1}^{k}=E_{2}^{k-1}=\{1, \ldots, M\}$
2. for $j=1$ to $M$,

- if $\lambda_{j}^{k-1} \in\left\{\lambda_{E_{1}^{k}}^{k}\right\}$, where $\left\{\lambda_{E_{1}^{k}}^{k}\right\}$ is the set of distinct values of the subvector of $\lambda^{k}$ indexed by $E_{1}^{k}$, then
- $E_{2}^{k-1}=E_{2}^{k-1} \backslash\{j\}$ ( $\backslash$ meaning "deprived of" here)
- $E_{1}^{k}=E_{1}^{k} \backslash\left\{j^{\prime}\right\}$, where $j^{\prime}=\min \left\{j^{\prime \prime} \in E_{1}^{k} \mid \lambda_{j^{\prime \prime}}^{k}=\lambda_{j}^{k-1}\right\}$

3. construct $\bar{E}_{2}^{k-1}$ and $\bar{E}_{1}^{k}$, complementary in $\{1, \ldots, M\}$ of $E_{2}^{k-1}$ and $E_{1}^{k}$.

## Algorithm A. 2 (Fickus et al. [2013])

1. Set $\varphi^{1}=\sqrt{\Pi_{1}^{\triangleright}} u^{1}$ where $u^{1}$ is the first column of $U^{1}$.
2. for $k=2$ to $N$,

- construct the sets $E_{1}^{k}, E_{2}^{k-1}, \bar{E}_{1}^{k}, \bar{E}_{2}^{k-1}$, such that $E_{1}^{k} \cup \bar{E}_{1}^{k}=E_{2}^{k-1} \cup \bar{E}_{2}^{k-1}=\{1, \ldots, M\}$ and $E_{1}^{k} \cap \bar{E}_{1}^{k}=E_{2}^{k-1} \cap \bar{E}_{2}^{k-1}=\varnothing$ based on the principles of algorithm A. 1
- let $r_{k}=\operatorname{card}\left(E_{1}^{k}\right)=\operatorname{card}\left(E_{2}^{k-1}\right)$
- let $E_{1}^{\prime k}=(M+1) 1^{r_{k}}-E_{1}^{k}\left(\right.$ resp. $\left.E_{2}^{\prime k-1}\right), \bar{E}_{1}^{\prime k}=(M+1) 1^{M-r_{k}}-\bar{E}_{1}^{k}\left(\right.$ resp. $\left.\bar{E}_{2}^{\prime k-1}\right)$
- let $\sigma_{1}^{k}\left(\right.$ resp. $\left.\sigma_{2}^{k}\right)$ be the unique permutation of $\{1, \ldots, M\}$ increasing in $E_{1}^{\prime k}$ and $\bar{E}_{1}^{* k}$ (resp. $E_{2}^{\prime k-1}$ and $\bar{E}_{2}^{\prime k-1}$ ) and such that $\sigma_{1}^{k}(j) \in\left\{1, \ldots, r_{k}\right\}$ for any $j \in E_{1}^{\prime k}$ (resp. $\sigma_{2}^{k-1}(j) \in\left\{1, \ldots, r_{k}\right\}$ for any $j \in E_{2}^{k-1}$ ). Let $\Sigma_{1}^{k}$ (resp. $\Sigma_{2}^{k-1}$ ) be the associated permutation matrices
- let $R^{k}=\mathcal{J}_{V_{k}}\left(\lambda_{E_{2}^{k-1}}^{k-1}, \lambda_{E_{1}^{k}}^{k}\right)$ be a matrix of size $\left(r_{k} \times 2\right)$, where $\lambda_{E_{2}^{k-1}}^{k-1}\left(\right.$ resp. $\left.\lambda_{E_{1}^{k}}^{k}\right)$, the vector extracted from $\lambda^{k-1}$ (resp. $\lambda^{k}$ ) whose rows are indexed by $E_{2}^{k-1}$ (resp. $\left.E_{1}^{k}\right)$, and $\mathcal{J}_{r_{k}}$ refers to the anti-diagonal matrix of size $r_{k}$ :

$$
\mathcal{J}_{3}=\left(\begin{array}{lll}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{array}\right)
$$

- let $v^{k}, w^{k}$ be two vectors of size $r_{k}$ and $W^{k}$ be a matrix $r_{k} \times r_{k}$ such that

$$
\begin{align*}
& v_{i}^{k}=\sqrt{-\frac{\prod_{i=1}^{v_{k}}\left(R_{i 1}^{k}-R_{i 2}^{k}\right)}{\prod_{\substack{i=1 \\
i=l_{k} \\
i=1}}^{i_{i}}\left(R_{i 1}^{k}-R_{i 1}^{k}\right)}}  \tag{A.4}\\
& w_{i}^{k}=\sqrt{\frac{\prod_{\substack{i=1 \\
v_{k}}}^{\prod_{i-1}^{i=l_{k}}\left(R_{i 2}^{k}-R_{i 1}^{k}\right)}}{\left.i_{i=1}^{k}-R_{i 2}^{k}-R_{i^{\prime} 2}^{k}\right)}} ;  \tag{A.5}\\
& W^{k}=\left(e_{r_{k}} \odot R_{.2}^{k^{\top}}-e_{r_{k}}^{\top} \odot R_{.1}^{k}\right)^{[-1]} \odot\left(v^{k} w^{k^{\top}}\right) ; \tag{A.6}
\end{align*}
$$

where $[-1]$ refers to the matrix inverse in the sense of the Hadamard product

- set

$$
\begin{gather*}
\varphi^{k}=U^{k-1} V^{k-1} \Sigma_{2}^{k-1^{\top}}\left[\begin{array}{c}
v^{k} \\
0_{M-r_{k}}
\end{array}\right],  \tag{A.7}\\
U^{k}=U^{k-1} V^{k-1} \Sigma_{2}^{k-1 \top}\left[\begin{array}{cc}
W^{k} & 0_{\left(r_{k}, M-r_{k}\right)} \\
0_{\left(M-r_{k}, r_{k}\right)} & I_{\left(M-r_{k}, M-r_{k}\right)}
\end{array}\right] \Sigma_{1}^{k}, \tag{A.8}
\end{gather*}
$$

and construct $\Phi^{k}$ the matrix whose columns are the $\left\{\varphi^{s}\right\}_{s=1}^{s=k}$.
The theorem of Fickus et al. (2013) states that, for $k=2, \ldots, N$,

- $U^{k}$ is a basis of orthonormalized eigenvectors of $\Phi^{k} \bar{\Phi}^{k \top}$ whose spectrum is $\lambda^{k}$
- $K_{k}^{\Pi^{\triangleright}}=\bar{\Phi}^{k^{\top}} \Phi^{k}$ is a positive semidefinite Hermitian matrix whose diagonal is $\left(\Pi_{1}^{\triangleright}, \ldots, \Pi_{k}^{\triangleright}\right)^{\top}$ and whose strictly positive eigenvalues are given by $\lambda^{k}$.

In practice, we are interested in the result at the end of the process, in other words, for $k=N$. In this case, the matrix $K_{N}^{\Pi^{\triangleright}}$ has the desired properties. The authors demonstrate that, by varying the different parameters, all the matrices with a fixed diagonal and spectrum $\lambda^{N}$ are obtained. Conversely, any matrix of this type can be constructed based on this procedure.

It is inferred from these results that a basis of orthonormalized eigenvectors of $K_{N}^{\Pi^{\triangleright}}$ is given by $\bar{\Phi}^{N^{T}} U^{N} D_{\lambda^{-N N}}^{-\frac{1}{2}}$. This basis can be used directly as input to the selection algorithm of Lavancier et al. (2015) in the case of projection matrices, and with a slight adaptation for determinantal designs of random size (Loonis and Mary 2019).

## A. 6 Constructing the elements of algorithms A. 1 and A. 2

We show here how to obtain the various quantities of algorithms A. 1 and A. 2 from the example in Table A. 1 for $k=9$ and $M=7$.

$$
\begin{gathered}
E_{2}^{8}=\{1,2,5,6\} \quad E_{1}^{9}=\{1,4,5,7\} \\
E_{2}^{\prime 8}=\{2,3,6,7\} \quad E_{1}^{\prime 9}=\{1,3,4,7\} \\
k=9, r_{9}=4, R^{9}=\left(\begin{array}{llll}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{array}\right)\left(\begin{array}{cc}
\frac{12}{55} & \frac{3}{11} \\
\frac{63}{220} & \frac{3}{10} \\
\frac{6}{11} & \frac{63}{110} \\
\frac{129}{220} & \frac{3}{5}
\end{array}\right)=\left(\begin{array}{cc}
\frac{129}{220} & \frac{3}{5} \\
\frac{6}{11} & \frac{63}{110} \\
\frac{63}{220} & \frac{3}{10} \\
\frac{12}{55} & \frac{3}{11}
\end{array}\right) \\
v_{1}^{9}=\sqrt{\left(\begin{array}{l}
\left(\frac{3}{5}-\frac{129}{220}\right)\left(\frac{63}{110}-\frac{129}{220}\right)\left(\frac{3}{10}-\frac{129}{220}\right)\left(\frac{3}{11}-\frac{129}{220}\right) \\
\bar{E}_{2}^{\prime 8}=\left\{\frac{129}{220}\right)
\end{array}\right.} \\
\left(\begin{array}{l}
\left.\frac{63}{220}-\frac{129}{220}\right)\left(\frac{12}{55}-\frac{129}{220}\right) \\
2 \\
2 \\
3 \\
3 \\
\sigma_{2}^{8}\left(\begin{array}{l}
5 \\
4 \\
5 \\
6 \\
7
\end{array}\right)=\left(\begin{array}{l}
5 \\
1 \\
2 \\
6 \\
7 \\
3 \\
3
\end{array}\right),\binom{E_{1}^{9}}{4}\left(\begin{array}{l}
1 \\
2 \\
3 \\
4 \\
5 \\
6 \\
7
\end{array}\right)=\left(\begin{array}{l}
1 \\
5 \\
2 \\
3 \\
6 \\
7 \\
4
\end{array}\right) .
\end{array}\right.
\end{gathered}
$$

It should be noted that $\sigma_{2}^{8}$ (resp. $\sigma_{1}^{9}$ ) is indeed increasing in $E_{2}^{\prime 8}$ (resp. $E_{1}^{\prime 9}$ ). In addition, for any $j \in E_{2}^{\prime 8}$ (resp. $j \in E_{1}^{\prime 9}$ ), we indeed have $\sigma_{2}^{8}(j) \in\left\{1, \ldots, r_{k}\right\}$ (resp. $\sigma_{1}^{9}(j) \in\left\{1, \ldots, r_{k}\right\}$ ). For a given permutation $\sigma$, the associated matrix is such that

$$
\begin{gathered}
\Sigma_{i j}=\left\{\begin{array}{lllll}
1 & \text { if } i=\sigma(j) \\
0 & \text { otherwise }
\end{array}\right. \\
\Sigma_{2}^{8}=\left(\begin{array}{lllllll}
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0
\end{array}\right), \Sigma_{1}^{9}=\left(\begin{array}{lllllll}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}\right) .
\end{gathered}
$$

## A. 7 Matrices $Q^{\Pi^{\triangleright}}$ for $\Pi^{\triangleright}=n / N$

Table A. 4
Eigenvalues $\lambda_{j}^{k}$ of the main sub-matrices of $Q^{\Pi^{\triangleright}}$, for $\Pi_{k}=n / N$ and $n$ does not divide $N$ : Example of $n=5$ and $N=12$.

|  |  |  | $\boldsymbol{k}$ |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\mathbf{8}$ | $\mathbf{9}$ | $\mathbf{1 0}$ | $\mathbf{1 1}$ | $\mathbf{1 2}$ |
| 0 | 0 | 0 | 0 | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 |
| 0 | 0 | 0 | $\bullet$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 |
| 0 | 0 | $\bullet$ | $\bullet$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 | 1 |
| 0 | $\bullet$ | $\bullet$ | $\bullet$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 | 1 | 1 |
| $\frac{n}{N}$ | $\bullet$ | $\bullet$ | $\bullet$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 | 1 | 1 | 1 |

Table A. 5
Eigenvalues of the main submatrices of $Q^{\Pi^{\triangleright}}$, for $\Pi_{k}=n / N$ and $n$ divides $N$ : Example of $n=4$ and $N=12$.

|  |  |  |  | $\boldsymbol{k}$ | $\mathbf{0}$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\mathbf{8}$ | $\mathbf{9}$ | $\mathbf{1 0}$ | $\mathbf{1 1}$ | $\mathbf{1 2}$ |
| 0 | 0 | 0 | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 |
| 0 | 0 | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 |
| 0 | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 | 1 |
| $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{2 n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | $1-\frac{n}{N}$ | 1 | 1 | 1 | 1 |

Figure A. $1 Q^{\Pi^{\triangleright}}$ under $\boldsymbol{C} 2$.
(a) General case: $n=3, N=10, \Pi_{k}=n / N$

$$
\begin{aligned}
Q^{\Pi^{\triangleright}}= & \left(\begin{array}{cccccccc}
n / N & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & n / N & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & n / N & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & n / N & 0 & 0 & n / N & 0 \\
\bullet & \bullet & \bullet & 0 & n / N & n / N & 0 & 0 \\
\bullet & \bullet & \bullet & 0 & n / N & n / N & 0 & 0 \\
\bullet & \bullet & \bullet & n / N & 0 & 0 & n / N & 0 \\
\bullet & \bullet & \bullet & 0 & 0 & 0 & 0 & n / N
\end{array}\right) \\
(b) & \text { Specific case: } n \text { divides } N: n=2, N=10, \Pi_{k}=n / N
\end{aligned}
$$

$$
Q^{\square^{\triangleright}}=\left(\begin{array}{cccccccc}
n / N & 0 & 0 & n / N & n / N & 0 & 0 & n / N \\
0 & n / N & n / N & 0 & 0 & n / N & n / N & 0 \\
0 & n / N & n / N & 0 & 0 & n / N & n / N & 0 \\
n / N & 0 & 0 & n / N & n / N & 0 & 0 & n / N \\
n / N & 0 & 0 & n / N & n / N & 0 & 0 & n / N \\
0 & n / N & n / N & 0 & 0 & n / N & n / N & 0 \\
0 & n / N & n / N & 0 & 0 & n / N & n / N & 0 \\
n / N & 0 & 0 & n / N & n / N & 0 & 0 & n / N
\end{array}\right) .
$$

The symbol • indicates that the explicit formula of the coefficient in question is unknown.

## A. 8 Optimization among real fixed-diagonal projection matrices

Among real matrices of size $(N \times N)$, we attempt to solve a problem of the following type:

$$
\operatorname{Min}_{K} \sum_{q=1}^{Q} \frac{x^{q^{\top}}\left(I_{N} \odot K\right)^{-1}\left[K \odot\left(I_{N}-K\right)\right]\left(I_{N} \odot K\right)^{-1} x^{q}}{t_{x^{q}}^{2}} \quad \text { s.c. }\left\{\begin{array}{l}
K=K^{\top} \\
\operatorname{diag}(K)=\Pi \\
K^{2}=K .
\end{array}\right.
$$

$K$ is a projection matrix, so there is a basis $V$ of orthonormalized vectors of size ( $N \times n$ ) such that $K=V V^{\top}$ and $V^{\top} V=I_{n}$. It is possible to rewrite the problem in the following penalized form:

$$
\operatorname{Min}_{V} \sum_{V=I_{n}} \sum_{q=1}^{Q} \frac{q^{\top^{\top}}\left(I_{N} \odot V V^{\top}\right)^{-1}\left[V V^{\top} \odot\left(I_{N}-\overline{V V^{\top}}\right)\right]\left(I_{N} \odot V V^{\top}\right)^{-1} x^{q}}{t_{x^{q}}^{2}}+r \operatorname{Trace}\left(\left(V V^{\top}-D_{\Pi}\right) \odot\left(V V^{\top}-D_{\Pi}\right)\right) .
$$

The objective of penalization via the trace function is for the diagonal of the optimal matrix to indeed be equal to $\Pi$. The set of real matrices $V$ such that $V^{\top} V=I_{n}$ is the Grassmannian manifold. The problem becomes a penalized optimization problem on varieties for which there are algorithms and powerful
problem-solving software, as long as the population size remains reasonable, around a few hundred (Absil et al., 2009; Boumal et al., 2014; Townsend et al., 2016). For the penalization parameter $r$, we set $r=10^{i}$, for $i=0, \ldots, 10$ and retained the one that minimizes, to the optimum set at $i$, the function $10^{i} \operatorname{Trace}\left(\left(V V^{\top}-D_{\Pi}\right) \odot\left(V V^{\top}-D_{\Pi}\right)\right)$.

## References

Absil, P.-A., Mahony, R. and Sepulchre, R. (2009). Optimization algorithms on matrix manifolds. Optimization Algorithms on Matrix Manifolds. Princeton University Press.

Boumal, N., Mishra, B., Absil, P.-A. and Sepulchre, R. (2014). Manopt, a matlab toolbox for optimization on manifolds. The Journal of Machine Learning Research, 15(1), 1455-1459.

Brändén, P., and Jonasson, J. (2012). Negative dependence in sampling. Scandinavian Journal of Statistics, 39(4), 830-838.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm for balanced sampling. Computational Statistics, 21(1), 53-62.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. Biometrika, 91(4), 893912.

Dhillon, I.S., Heath Jr, R.W., Sustik, M.A. and Tropp, J.A. (2005). Generalized finite algorithms for constructing hermitian matrices with prescribed diagonal and spectrum. SIAM Journal on Matrix Analysis and Applications, 27(1), 61-71.

Dita, P. (1982). Parametrisation of unitary matrices. Journal of Physics A: Mathematical and General, 15(11), 3465.

Dita, P. (1994). On the parametrisation of unitary matrices by the moduli of their elements. Communications in Mathematical Physics, 159(3), 581-591.

Fickus, M., Mixon, D.G. and Poteet, M.J. (2013). Constructing finite frames with a given spectrum. Finite Frames, Springer, 55-107.

Horn, A. (1954). Doubly stochastic matrices and the diagonal of a rotation matrix. American Journal of Mathematics, 76(3), 620-630.

Horn, R.A., and Johnson, C.R. (1991). Topics in Matrix Analysis. Cambridge Univ. Press.
Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685.

Hough, J.B., Krishnapur, M., Peres, Y. and Virág, B. (2006). Determinantal processes and independence. Probability Surveys, 3, 206-229.

Jarlskog, C. (2005). A recursive parametrization of unitary matrices. Journal of Mathematical Physics, 46(10), 103508.

Kadison, R.V. (2002). The pythagorean theorem: I. the finite case. Proceedings of the National Academy of Sciences, 99(7), 4178-4184.

Lavancier, F., Møller, J. and Rubak, E. (2015). Determinantal point process models and statistical inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(4), 853-877.

Leuenberger, M., Eustache, E., Jauslin, R. and Tillé, Y. (2022). Balancing a sample almost perfectly. Statistics \& Probability Letters, 180, 109229.

Loonis, V., and Mary, X. (2019). Determinantal sampling designs. Journal of Statistical Planning and Inference, 199, 60-88.

Lyons, R. (2003). Determinantal probability measures. Publications Mathématiques de l'Institut des Hautes Études Scientifiques, 98, 167-212.

Macchi, O. (1975). The coincidence approach to stochastic point processes. Advances in Applied Probability, 83-122.

Pemantle, R., and Peres, Y. (2014). Concentration of lipschitz functionals of determinantal and other strong rayleigh measures. Combinatorics, Probability and Computing, 23(01), 140-160.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). Model Assisted Survey Sampling. Springer Science \& Business Media.

Scardicchio, A., Zachary, C.E. and Torquato, S. (2009). Statistical properties of determinantal point processes in high-dimensional euclidean spaces. Physical Review E, 79(4), 041108.

Schur, J. (1911). Bemerkungen zur theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. Journal für die reine und angewandte Mathematik, 140, 1-28.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 5(1194), 127.

Soshnikov, A. (2000). Determinantal random point fields. Russian Mathematical Surveys, 55(5), 923.

Soshnikov, A. (2002). Gaussian limit for determinantal random point fields. Annals of Probability, 171187.

Spengler, C., Huber, M. and Hiesmayr, B.C. (2010). A composite parameterization of unitary groups, density matrices and subspaces. Journal of Physics A: Mathematical and Theoretical, 43(38), 385306.

Tillé, Y. (2001). Théorie des sondages. Dunod, Paris, 13, 23-26.

Townsend, J., Koep, N. and Weichwald, S. (2016). Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. arXiv preprint arXiv:1603.03236.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society: Series B (Methodological), 15(2), 253-261.

Yuan, M., Su, C. and Hu, T. (2003). A central limit theorem for random fields of negatively associated processes. Journal of Theoretical Probability, 16(2), 309-323.

# Design-based conformal prediction 

Jerzy Wieczorek ${ }^{\mathbf{1}}$


#### Abstract

Conformal prediction is an assumption-lean approach to generating distribution-free prediction intervals or sets, for nearly arbitrary predictive models, with guaranteed finite-sample coverage. Conformal methods are an active research topic in statistics and machine learning, but only recently have they been extended to non-exchangeable data. In this paper, we invite survey methodologists to begin using and contributing to conformal methods. We introduce how conformal prediction can be applied to data from several common complex sample survey designs, under a framework of design-based inference for a finite population, and we point out gaps where survey methodologists could fruitfully apply their expertise. Our simulations empirically bear out the theoretical guarantees of finite-sample coverage, and our real-data example demonstrates how conformal prediction can be applied to complex sample survey data in practice.


Key Words: Conformal prediction; Machine learning; Cross validation; Predictive modeling; Complex sample survey designs.

## 1. Introduction

What is conformal prediction? Conformal prediction, also called conformal inference, is a family of general-purpose approaches to constructing prediction intervals or prediction sets, which can be wrapped around almost any predictive modeling algorithm. Specifically, imagine that we have fit a function $\hat{f}_{n}$ to a set of training data $\left(X_{i}, Y_{i}\right)$ for $i=1, \ldots, n$ that allows us to make a point prediction for a new observation's response value $Y_{n+1}$ when $\hat{f}_{n}$ is evaluated at the covariate value $X_{n+1}$, and our goal is to report a level 1- $\alpha$ prediction set for $Y_{n+1}$. Let us require only that all $n+1$ of the points ( $X_{i}, Y_{i}$ ) are an exchangeable sample from some common distribution $P$, and that the algorithm used to fit $\hat{f}_{n}$ treats these points symmetrically (which rules out e.g. algorithms that give more weight to more-recent data over time). (A sequence of random variables is exchangeable when its joint probability distribution would not change if you permuted the random variables. More precisely, a sequence $X_{1}, X_{2}, \ldots$ is exchangeable if for each $n$ and each permutation $\pi$ of $\{1, \ldots, n\}$, the joint distribution of $\left(X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$ is just the same as the distribution of $\left(X_{1}, \ldots, X_{n}\right)$ (Durrett, 2019). Examples of exchangeability include iid sequences as well as simple random sampling (with or without replacement).) If so, we can use standard conformal prediction methods to construct $\hat{C}_{n}$ which are either prediction intervals (for a regression problem) or prediction sets (for a classification problem), such that $\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n}\left(X_{n+1}\right)\right] \geq 1-\alpha$, where the probability is taken over repeated sampling of all $n+1$ points. In Section 2 we describe two standard approaches to conformal prediction and describe the intuition behind how they work.

Although survey data analysis has traditionally focused on estimation and testing, predictive modeling with complex sample survey data is not uncommon. To name just a few examples: Hong and He (2010) use the Second Longitudinal Study on Aging (National Center for Health Statistics, 2016) to fit a predictive
model that can be used to monitor functional mobility status among the elderly. Kshirsagar, Wieczorek, Ramanathan and Wells (2017) use 2015 Living Conditions Monitoring Survey data (Central Statistical Office, Zambia, 2015) to build a model to classify households by poverty level, at a range of poverty thresholds. Krebs, Reeves and Baggett (2019) use Forest Inventory and Analysis data (Bechtold and Patterson, 2005) to build random-forest models that can predict understory vegetation structure, intending for these models to be implemented in the Forest Vegetation Simulator (Crookston and Dixon, 2005). In each case, there is interest in using the model to make unit-level predictions, not just to estimate regression coefficients or population means. Yet, with the exception of Hong and He (2010), many such papers have not been able to provide prediction intervals or sets.

Why is this interesting to survey methodologists? For practitioners, survey data analysts who build predictive models - like those mentioned above - could apply conformal methods to provide design-based prediction intervals or sets (Section 3). Such prediction intervals or sets would have guaranteed coverage, even for novel machine learning algorithms. Conformal methods may not add much value to older methods like linear regression, which already has well-known Gaussian-based prediction intervals and has already been adapted to account for complex survey designs. But for newer methods - such as prediction algorithms that have not yet been adapted to work optimally with survey designs, or non-probabilistic algorithms that do not come with "built-in" prediction intervals such as the simulation models of LeRoy and Schafer (2021) - we can apply conformal prediction methods that guarantee coverage based on the sampling design alone. Those who collect and pre-process the survey data may also find uses for conformal methods, for instance in nonresponse prediction, imputation, or data cleaning (Section 5).

More philosophically, the principles behind conformal methods align with complex survey sampling inference. [a] Conformal methods and many design-based methods provide exact, finite-sample coverage guarantees, not asymptotic or approximate guarantees. [b] Unlike traditional model-based approaches (such as Gaussian-errors prediction intervals for regression), conformal prediction does not require assumptions about the distribution $P$ - only exchangeable sampling. Likewise, design-based methods do not require assumptions about how the data values are distributed in the population - only knowledge of the sampling design. Finally, [c] conformal guarantees hold even if the predictive model is not a good fit to $P$ (though of course the prediction intervals or sets may be larger then), and many model-assisted methods in designbased inference also have guarantees that hold whether or not the model is a good fit to the population.

Hence, we believe this is a good opportunity to cross-pollinate ideas between the survey methodology and machine learning research communities. The growth of conformal prediction indicates that machine learning practitioners are seeking methods that are guaranteed to work based only on the sampling design, not on the distribution of the data values. Survey methodologists, having the right expertise to meet that demand, may enjoy applying their skillset to this new challenge.

Why is this not already in use for survey sampling? Despite the apparent affinities, conformal methods have not previously been studied under a complex sample survey framework. Conformal prediction has
been under development for several decades, initially by Vladimir Vovk and colleagues (Vovk, Gammerman and Shafer, 2022) and more recently by a wide range of statistical and machine learning researchers. Conformal methods are starting to be deployed in real-world settings, such as the Washington Post's 2020 presidential election tracker, which reported prediction intervals of the votes for each party that were updated in real time as voting districts slowly reported their results (Cherian and Bronner, 2020). Yet the requirement of exchangeability has made traditional conformal methods inapplicable for complex sampling designs other than simple random sampling (with or without replacement).

However, recently Tibshirani, Barber, Candes and Ramdas (2019) extended conformal prediction to what they call "weighted exchangeable" sampling, while Dunn, Wasserman and Ramdas (2022) derived several "hierarchical" conformal methods that can apply to cluster sampling. The present work builds on these two papers to begin providing conformal guarantees for complex sampling designs.

Briefly, instead of assuming that all $n+1$ data points are exchangeable, we will assume that the $n$ training points came from a known complex sampling design on a finite population, and the test point was selected uniformly at random from the same finite population. The latter assumption is simply a way of phrasing the guarantee of marginal coverage across the finite population. In Section 3, we summarize the complex survey settings in which such conformal prediction methods are currently known to have coverage guarantees.

### 1.1 Related work

Prediction intervals or tolerance regions have a long history in statistics; for a recent review, see Tian, Nordman and Meeker (2022). Section 13.3.3 of Vovk et al. (2022) connects conformal methods to much of this earlier work. However, conformal prediction is currently a broad and very active research area, and we encourage interested readers to look into the following general resources: Angelopoulos and Bates (2022) provide an introductory review article; Vovk and colleagues maintain a list of their own recent work at http://alrw.net/; Manokhin (2022) maintains a frequently updated list of conformal papers, tutorials, and software; and COPA, the Symposium on Conformal and Probabilistic Prediction with Applications, is an annual conference on conformal prediction and its extensions: https://copa-conference.com/.

For dependent data, our paper builds largely on the work of Dunn et al. (2022) and Tibshirani et al. (2019). Fong and Holmes (2021) give a distinct, Bayesian approach to conformal prediction for hierarchical data. Barber, Candès, Ramdas and Tibshirani (2022) relax the usual conformal-prediction requirement that the fitted prediction model must treat data points symmetrically, and Fannjiang, Bates, Angelopoulos, Listgarten and Jordan (2022) address "feedback covariate shift", where the choice of which test data to sample depends on results from the training data, which could be useful in addressing adaptive sampling for surveys (Thompson, 1997). Conformal methods for nonstationarity over time (Chernozhukov, Wüthrich and Yinchu, 2018; Oliveira, Orenstein, Ramos and Romano, 2022), including distribution drift (Gibbs and

Candes, 2021), would be useful in developing conformal methods for panel surveys. Recently, Lunde (2023) developed conformal methods for network data under non-uniform sampling, with implications for respondent-driven sampling in surveys.

Few papers have applied conformal methods to survey samples. Bersson and Hoff (2022) do use conformal prediction for small area estimation problems which rely on survey data. However, they work in a model-based framework and assume that sampled units are exchangeable within each small area, rather than focusing on design-based inference under a complex sampling design. Romano, Patterson and Candes (2019) and several followup papers (Sesio and Candes, 2020; Sesio and Romano, 2021; Feldman, Bates and Romano, 2021; Bai, Mei, Wang, Zhou and Xiong, 2022) illustrate conformal prediction methods using unitlevel records from a complex survey: the Medical Expenditure Panel Survey (Agency for Healthcare Research and Quality, 2017). However, they appear to ignore the documented sampling design and treat it as exchangeable.

Apart from conformal methods, there has been increasing interest in applying machine learning methods to complex survey data. Dagdoug, Goga and Haziza (2022) summarize the current state of the art for using survey data with many machine learning algorithms: penalized regression (McConville, Breidt, Lee and Moisen, 2017), k nearest neighbors (Yang and Kim, 2019), random forests (Dagdoug, Goga and Haziza, 2021), and others. However, along with Sande and Zhang (2021), they focus on estimation of means and totals, rather than on predicting individual response values. Even when such algorithms have been adapted to make individual predictions that account for the survey design, many still do not offer procedures for creating prediction intervals, which design-based conformal methods could provide.

### 1.2 Our contributions

We introduce survey methodologists to key definitions and intutions behind standard conformal methods under exchangeable sampling as well as their extensions to complex sampling (Section 2). Furthermore, we derive exact, finite-sample, design-based coverage guarantees for applying conformal inference to data from several fundamental sampling designs, building on results from the weighted exchangeable or hierarchical frameworks (Section 3).

Next, we illustrate our design-based guarantees through simulations and a real-data example (Section 4). We show that results can differ on real data depending on whether or not conformal methods account for the survey design, and we show how conformal methods are affected by different sampling designs. We then discuss several practical considerations and future challenges to be addressed in applying conformal methods to survey data (Section 5). We also suggest ways that conformal methods might be useful in conducting surveys. Finally, we invite survey researchers to contribute to the literature on conformal methods (Section 6). We anticipate that advances from the design-based perspective will turn out to be useful to the general community of conformal researchers.

## 2. Introduction to conformal inference

### 2.1 Definitions

Because one of our main contributions is a corollary of results in Tibshirani et al. (2019), we restate without proof some of their notation, definitions, and results here and in Section 3.

Let Quantile $(\beta ; F)$ denote the level $\beta$ quantile of distribution $F$, so that for $Y \sim F$,

$$
\text { Quantile }(\beta ; F)=\inf \{y: \mathbb{P}\{Y \leq y\} \geq \beta\} .
$$

We allow for distributions on the augmented real line, $\mathbb{R} \cup\{\infty\}$. We use $v_{1: n}=\left\{v_{1}, \ldots, v_{n}\right\}$ to denote a multiset, meaning that it is unordered and can allow the same element to appear several times. We use $\delta_{a}$ to denote a point mass at the value $a$. If $v_{1: n}$ is an exchangeable sample, its empirical probability measure is $n^{-1} \sum_{i=1}^{n} \delta_{v_{i}}$, and the level $\beta$ quantile of its empirical distribution is Quantile $\left(\beta ; v_{1: n}\right)$ which is the $\lceil\beta n\rceil$ smallest value in $v_{1: n}$, where $\lceil\cdot\rceil$ is the ceiling function.

Now we can state the general-purpose "quantile lemma" that forms the foundation for conformal inference in the exchangeable setting. For instance, if we apply this lemma directly to data from an exchangeable sample of scalar random variables $V_{1}, \ldots, V_{n}$, we obtain a level $\beta$ one-sided prediction interval for a new observation $V_{n+1}$.

Lemma 1 (Tibshirani et al. (2019)). If $V_{1}, \ldots, V_{n+1}$ are exchangeable random variables, then for any $\beta \in(0,1)$, we have

$$
\mathbb{P}\left\{V_{n+1} \leq \operatorname{Quantile}\left(\beta ; V_{1: n} \cup \infty\right)\right\} \geq \beta .
$$

Furthermore, if ties between $V_{1}, \ldots, V_{n+1}$ occur with probability zero, then the above probability is upper bounded by $\beta+1 /(n+1)$.

In order to use this lemma to do conformal prediction for nontrivial regression or classification problems, we must also choose a score function $\mathcal{S}$, which takes the following arguments: a point ( $x, y$ ), and a multiset $Z$ (meaning that $\mathcal{S}$ must treat the points in $Z$ as unordered). The points in $Z$ will typically be the $n$ observations $Z_{i}=\left(X_{i}, Y_{i}\right)$ for $i=1, \ldots, n$, possibly along with one additional hypothetical observation or test case; $(x, y)$ is typically one of the points in $Z . \mathcal{S}$ should return a real value, such that lower values indicate that $(x, y)$ "conforms" to $Z$ better. Finally, use $\mathcal{S}$ to define nonconformity scores

$$
\begin{equation*}
V_{i}^{(x, y)}=\mathcal{S}\left(Z_{i}, Z_{1: n} \cup\{(x, y)\}\right), i=1, \ldots, n, \text { and } V_{n+1}^{(x, y)}=\mathcal{S}\left((x, y), Z_{1: n} \cup\{(x, y)\}\right) . \tag{2.1}
\end{equation*}
$$

The reason we require $Z$ to be an unordered multiset is to ensure that if the observations $\left(X_{i}, Y_{i}\right)$ are exchangeable, then the scores $V_{i}^{(x, y)}$ will be too.

For instance, in a regression context, we might choose $\mathcal{S}((x, y), Z)=|y-\hat{f}(x)|$ where $\hat{f}$ is some regression function fitted using all of $Z$, including $(x, y)$. If the absolute-residual nonconformity score
$V_{n+1}^{(x, y)}$ is small relative to all of the scores $V_{i}^{(x, y)}$, this suggests that $(x, y)$ conforms well to the overall trend in $Z$.

In a classification setting, we might choose $\mathcal{S}((x, y), Z)=1-\hat{f}(x)_{y}$ where $\hat{f}$ is the probability of class $y$ estimated by some classification function fitted using all of $Z$, again including $(x, y)$. A small value of this nonconformity score $V_{n+1}^{(x, y)}$ suggests that $(x, y)$ conforms well to $Z$, in the following sense: This observation with covariates $x$ has a high estimated probability of being in class $y$ based on trends in $Z$, and it does indeed belong to class $y$.

We will also use following common abbreviations: SRS (simple random sampling), WR (with replacement), WOR (without replacement), PPS (probability proportional to size sampling). $N$ denotes a population size and $n$ typically denotes a sample size, except for the "split conformal" methods (defined in Section 2.4) which split the data into a "proper training set" of size $m$ and a "calibration set" of size $n$.

### 2.2 Intuition under exchangeable sampling

Quantile lemma: The lower bound in Lemma 1 has a simple rationale. First, let $\hat{q}_{n+1}$ be the $\lceil\beta(n+1)\rceil$ smallest value in $V_{1:(n+1)}$, and let $\hat{q}_{\text {conf }}$ be the $\lceil\beta(n+1)\rceil$ smallest value in $V_{1: n} \cup \infty$. By exchangeability, the rank of $V_{n+1}$ among all the $V_{i}$ is uniformly distributed over $\{1, \ldots, n+1\}$, so $V_{n+1}$ is at or below $\hat{q}_{n+1}$ with probability exactly $\beta$. Then since $\hat{q}_{\text {conf }} \geq \hat{q}_{n+1}, V_{n+1}$ is at or below $\hat{q}_{\text {conf }}$ with probability at least $\beta$. Note that these probabilities are marginal (over exchangeable sampling of all the $V_{1}, \ldots, V_{n+1}$ together) - not conditional on the first $n$ observations nor conditional on the test observation.

This is an exact finite-sample result that only relies on exchangeability. By contrast, if we used instead $\hat{q}_{n}$, the $\lceil\beta n\rceil$ smallest value in $V_{1: n}$, then it may be an asymptotically good estimate of the population quantile, but it would require stronger conditions on the data distribution. Even then, the probability that $V_{n+1}$ is at or below $\hat{q}_{n}$ would be only approximately $\beta$.

To see why $\hat{q}_{\text {conf }} \geq \hat{q}_{n+1}$, imagine an empirical cumulative distribution function (eCDF) plot of just the first $n$ datapoints, with step heights of $1 / n$ at each observed value. (See the top left subplot of Figure 2.1.) If we add one more datapoint, the step heights of the eCDF will be $1 /(n+1)$ instead, and the lower $\beta$ quantile will be $\hat{q}_{n+1}$ as above. It may be larger or smaller than the lower $\beta$ quantile of the first $n$ datapoints, depending on how large the last datapoint is. The lower $\beta$ quantile is largest if we choose a $(n+1)^{\text {th }}$ value that is larger than any of the first $n$ observed values, say at $\infty$. That is because if the added datapoint is larger than any of the others, it pushes the rest of the eCDF down, so the horizontal line with $y$-intercept of $\beta$ crosses the eCDF at an $x$ value further to the right. In this most extreme case, the lower $\beta$ quantile is $\hat{q}_{\text {conf }}$ as above, and therefore $\hat{q}_{\text {conf }} \geq \hat{q}_{n+1}$. (See the bottom left subplot of Figure 2.1.) By contrast, if instead of $\infty$ we had chosen to add a value smaller than some of the first $n$ observations, it would have pushed part of the eCDF upward and may have caused the lower $\beta$ quantile to become smaller than $\hat{q}_{n+1}$.

Figure 2.1


Notes: Top left: eCDF for an iid sample with values $\{1,2,3,4\}$. At each value there is a vertical jump of $1 / n=0.25$. The $75^{\text {th }}$ percentile is at 3 . Bottom left: same eCDF but padded with an extra value at $\infty$. Now at each value there is a vertical jump of $1 / n+1=0.20$. The $75^{\text {th }}$ percentile is between 3 and 4 , so we round up to 4 .

Top right: eCDF for a survey sample with values $\{1,2,3,4\}$ and corresponding survey weights $\{4,3,2,1\}$. At each value there is a vertical jump proportional to its survey weight. The $75^{\text {th }}$ percentile is between 2 and 3 , so we round up to 3 .
Bottom right: same eCDF but padded with an extra value at $\infty$, corresponding to a not-yet-observed unit in the population that would have a survey weight of 3 if it were to be sampled. The vertical jumps at each value are still proportional to the survey weights, but now rescaled to make room for the weight of the extra value at $\infty$. The $75^{\text {th }}$ percentile is between 3 and 4 , so we round up to 4 .

Put another way, the probability that \{we take an exchangeable sample of size $n+1$ in which the last observation is greater than the $\lceil(1-\alpha)(n+1)\rceil$ smallest of the first $n$ observations $\}$ is at most $\alpha$. This is a strict inequality, not approximate or asymptotic.

Regression prediction intervals: Next, we can apply this quantile lemma in a regression setting. (In Section 2.5 we describe an approach for classification problems.) Imagine that we are given a fixed regression function $\hat{f}$ for making predictions of some real-valued random variable $Y$ from some covariate vectors $X$ for data from this population. (For the moment, assume $\hat{f}$ is not data-dependent; perhaps it was chosen a priori.) Also imagine that we have some residuals from applying $\hat{f}$ to a "calibration set" of $n$ different data points $\left(X_{1}, Y_{1}\right), \ldots,\left(X_{n}, Y_{n}\right)$. (We recognize that the term "calibration" traditionally has a completely different meaning in survey sampling. In this paper, we will use "calibration" only in this conformal inference sense.) Finally, we have one other covariate vector $X_{n+1}$, drawn exchangeably from the same population, but we do not know its $Y$ value. We want to get a level $(1-\alpha)$ prediction interval for $Y_{n+1}$ at this test-case $X_{n+1}$.

We can simply apply the "quantile lemma" logic to the $n$ absolute residuals from the calibration set. These absolute residuals are our nonconformity scores. If we let $\hat{q}$ be the $\lceil(1-\alpha)(n+1)\rceil$ smallest absolute residual, then $\hat{C}\left(X_{n+1}\right)=\hat{f}\left(X_{n+1}\right) \pm \hat{q}$ is a prediction interval for $Y_{n+1}$ with guaranteed coverage of at least $1-\alpha$.

Marginal vs. conditional coverage: This coverage is marginal across all samples of size $(n+1)$, with a size- $n$ calibration set plus one new test case. It is not conditional on the specific test case $X_{n+1}$ we chose; it assumes that both the calibration set and the test case together are exchangeable. Nonetheless, it is an exact finite-sample result. Barber, Candès, Ramdas and Tibshirani (2021) show that although no conformal method could generally guarantee "exact conditional coverage" (conditioning on the exact value of $X$ ), certain relaxed versions of conditional coverage are achievable. Angelopoulos and Bates (2022) review approaches to assessing and controlling various forms of conditional coverage, noting that marginal coverage alone may be insufficient, e.g. if it happens to be achieved by low coverage in a rare but important subpopulation and high coverage elsewhere. In Appendix A we review links between conditional coverage and the related idea of "adaptive" prediction regions.

### 2.3 Intuition under complex sampling

When the data are not exchangeable, Tibshirani et al. (2019) have extended conformal prediction to a setting called "covariate shift", in which the distribution of $X$ is differerent for the training and/or calibration sets than for the test set, but the conditional distribution of $Y \mid X$ remains the same. They define a condition called "weighted exchangeability" and show how to make conformal inference work for such data. In the present paper, we will show that certain classic finite-population sampling designs can be treated as a special case of covariate shift, and therefore the results of Tibshirani et al. (2019) apply.

Specifically, imagine we sample $n$ cases with replacement from a finite population with known but unequal sampling probabilities; for instance, we may be using PPS sampling. Assume also that we take a SRS of just one test case from the full finite population. Now, we need a "survey weighted quantile lemma" that tells us how to find an adjusted quantile $\hat{q}$ of the complex sample, such that a test case nonconformity score is no larger than $\hat{q}$ with probability at least $1-\alpha$. We will plug the $n$ complex sample observations from our calibration set into a regression function $\hat{f}$ and find $\hat{q}$ for the absolute residuals. Then the probability of our test case $\left(X_{n+1}, Y_{n+1}\right)$ having a larger absolute residual is at most $\alpha$, and so $\hat{f}\left(X_{n+1}\right) \pm \hat{q}$ is a level 1- $\alpha$ prediction interval for $Y_{n+1}$.

In Section 3 we will show that in this unequal-probabilities setting, it is enough to mimic the exchangeable setting, except that instead of using an equal-weights eCDF to get the quantiles, we use a survey-weighted eCDF; see e.g. Section 5.11 of Särndal, Swensson and Wretman (1992). The additional observation at $\infty$ will simply be assigned the known sampling weight that the test case ( $X_{n+1}, Y_{n+1}$ ) would have had, under the complex sampling design used to sample the first $n$ units. See the right half of Figure 2.1 for an illustration.

We acknowledge that there is a long-running debate in the survey sampling literature about whether and how to use sampling weights for model fitting and inference (Fienberg, 2010; Lumley and Scott, 2017). Our goal in the present paper is not to take a stance in this debate, but simply to show how conformal inference can be applied if the data analyst is taking a design-based perspective. In a model-based analysis, if the design features can justifiably be ignored, standard conformal inference methods may be used.

### 2.4 Split vs. full conformal

Above, for simplicity we have assumed that a $\hat{f}$ has been provided for us. More typically, we will need to fit $\hat{f}$ using data from the sample at hand. In the "split conformal" approach (Lei, G'Sell, Rinaldo, Tibshirani and Wasserman, 2018), also called "inductive conformal" (Papadopoulos, Proedrou, Vovk and Gammerman, 2002), we start with an exchangeable dataset of $m+n$ observations, and we split them at random into a "proper training set" of size $m$ to be used for training $\hat{f}$, plus a calibration set of size $n$ as described above. In this situation, coverage is still marginal over the calibration set plus one test case, but now it is conditional on the proper training set.

An optimal sample splitting ratio $m / n$ for split conformal is not known. Larger $m / n$ should lead to better estimates of $\hat{f}$, and therefore shorter conformal prediction interval lengths on average. On the other hand, larger $m / n$ also leads to a smaller calibration set, and therefore more-variable conformal prediction interval lengths.

If we do not wish to lose statistical efficiency by splitting our data, we can use a more computationallyintensive "full conformal" approach (Vovk et al., 2022). In that approach, we no longer need a separate calibration set, and we let $n$ denote the total number of our complete-data training cases. We also have one test case with only the covariates $X_{n+1}$ known. Then we repeat the following process for many $y$ values:

- Choose a hypothetical response value $y \in \mathbb{R}$.
- Fit $\hat{f}_{y}$ to an augmented dataset in which we pretend this $y$-value is correct: $\left(X_{1}, Y_{1}\right), \ldots$, $\left(X_{n}, Y_{n}\right),\left(X_{n+1}, y\right)$. Find the $n+1$ nonconformity scores, e.g. the absolute residuals: $R_{y, i}=$ $\left|Y_{i}-\hat{f}_{y}\left(X_{i}\right)\right|$ for $i=1, \ldots, n$ and $R_{y, n+1}=\left|y-\hat{f}_{y}\left(X_{n+1}\right)\right|$. Find their $1-\alpha$ quantile: $\hat{q}_{y}$ is the $\lceil(1-\alpha)(n+1)\rceil$ smallest value of $R_{y, i}$ for $i=1, \ldots, n+1$.
- If $R_{y, n+1} \leq \hat{q}_{y}$, we say $y$ "conforms" to the rest of the data, and we add $y$ to our prediction interval.

In terms of notation, refer back to (2.1). Full conformal nonconformity scores are calculated using $Z=Z_{1: n} \cup\left\{\left(X_{n+1}, y\right)\right\}$ and refitting a new prediction model $\hat{f}_{y}$ for each new $y$ or $X_{n+1}$. By contrast, split conformal nonconformity scores use a fixed $\hat{f}$ conditional on the proper training set. Each of the first $n$ nonconformity scores is calculated only using that data point $\left(X_{i}, Y_{i}\right)$ and the fixed $\hat{f}$, and there is no need to calculate a $(n+1)^{\text {th }}$ nonconformity score. Under the notation in (2.1), we allow $Z$ to be ignored and $V_{n+1}^{(x, y)}$ to be undefined for split conformal.

Compared to split conformal, the full conformal procedure reduces the variance in the reported prediction interval endpoints. On the other hand, full conformal is far more computationally intensive than split conformal, as $\hat{f}_{y}$ has to be refit for every candidate $y$ at a given test point $X_{n+1}$, and the entire interval needs to be refit for every test point.

In empirical comparisons, Lei et al. (2018) found that split conformal and full conformal often produce very similar prediction intervals. They recommend using split conformal, on the grounds that it is faster to compute with little loss of efficiency, although full conformal avoids randomness in the data split.

### 2.5 Classification problems and prediction sets

For brevity, we focus on regression prediction intervals, but classification prediction sets are another common use case. For instance, consider split conformal inference for a multi-class probabilistic classifier. Examples range from simple logistic regression to deep neural networks. First, fit the classifier to the proper training set to get a function $\hat{f}(x)_{y}$, whose outputs are the estimated probabilitity of class $y$ for an input $x$ . Then, using the calibration set, each calibration "residual" or nonconformity score $V_{i}$ is calculated as $1-\hat{f}\left(X_{i}\right)_{Y_{i}}$ for each calibration observation $i=1, \ldots, n$. Next, find the corrected $1-\alpha$ quantile of these probabilities: let $\hat{q}$ be the $\lceil(1-\alpha)(n+1)\rceil$ smallest value of $V_{i}$ for $i=1, \ldots, n$. Finally, for a new test case $X_{n+1}$, find its estimated probability for each class $y$, and choose all the classes whose nonconformity scores are below the corrected quantile: $\hat{C}\left(X_{n+1}\right)=\left\{y: 1-\hat{f}\left(X_{n+1}\right)_{y} \leq \hat{q}\right\}$. Romano, Sesia and Candes (2020) and Angelopoulos and Bates (2022) discuss refinements to this approach.

## 3. Methods

For the reader's convenience, we restate several of the key results from Tibshirani et al. (2019) below. Next we prove that we can apply these results to unequal-probability sampling with replacement. We follow by discussing methods for sampling without replacement, cluster sampling based on Dunn et al. (2022), stratified sampling, and post-stratification.

Most of the methods below apply both to full and split conformal. However, for split conformal, we recommend using the design-based approach of Wieczorek, Guerin and McMahon (2022) to split survey data into a proper training set and a calibration set. This way, both the proper training set and the calibration set will mimic the original sampling design, which ensures that the methods below are safe to apply to your calibration set. By contrast, a simple random split can cause the calibration set to have different properties than a clustered or stratified sampling design.

### 3.1 Previous results for covariate shift (Tibshirani et al., 2019)

Tibshirani et al. (2019) define the "covariate shift" setting as follows:

$$
\begin{gather*}
\left(X_{i}, Y_{i}\right) \stackrel{\text { iid }}{\sim} P=P_{X} \times P_{Y \mid X}, i=1, \ldots, n, \\
\left(X_{n+1}, Y_{n+1}\right) \sim \tilde{P}=\tilde{P}_{X} \times P_{Y \mid X}, \text { independently. } \tag{3.1}
\end{gather*}
$$

Note that the conditional distribution of $Y \mid X$ remains the same as the marginal distribution of $X$ changes.
Assuming that $P_{X}$ and $\tilde{P}_{X}$ are known, we can define likelihood ratio weight functions $w=\mathrm{d} \tilde{P}_{X} / \mathrm{d} P_{X}$. We use these to define a second set of weights:

$$
\begin{equation*}
p_{i}^{w}(x)=\frac{w\left(X_{i}\right)}{\sum_{j=1}^{n} w\left(X_{j}\right)+w(x)}, \quad i=1, \ldots, n, \text { and } \quad p_{n+1}^{w}(x)=\frac{w(x)}{\sum_{j=1}^{n} w\left(X_{j}\right)+w(x)} . \tag{3.2}
\end{equation*}
$$

In this setting, we can state a weighted, nonexchangeable counterpart to Lemma 1. Although Tibshirani et al. (2019) give and prove a more general version, here we only state a version tailored to the covariate shift setting.

Lemma 2. Assume data from the model (3.1). Assume $\tilde{P}_{X}$ is absolutely continuous with respect to $P_{X}$, and denote $w=\mathrm{d} \tilde{P}_{X} / \mathrm{d} P_{X}$. For any $\beta \in(0,1)$,

$$
\mathbb{P}\left\{V_{n+1} \leq \text { Quantile }\left(\beta ; \sum_{i=1}^{n} p_{i}^{w}(x) \delta_{V_{i}}+p_{n+1}^{w}(x) \delta_{\infty}\right)\right\} \geq \beta,
$$

where $V_{i}^{(x, y)}, i=1, \ldots, n+1$ are as defined in (2.1), and $p_{i}^{w}, i=1, \ldots, n+1$ are as defined in (3.2).
Proof. Apply Lemma 3 of Tibshirani et al. (2019) in the covariate shift setting of (3.1).
This weighted quantile lemma allows conformal inference in the covariate shift setting.
Corollary 1 (Tibshirani et al. (2019)). Assume data from the model (3.1). Assume $\tilde{P}_{X}$ is absolutely continuous with respect to $P_{X}$, and denote $w=\mathrm{d} \tilde{P}_{X} / \mathrm{d} P_{X}$. For any score function $\mathcal{S}$, and any $\alpha \in(0,1)$, define for $x \in \mathbb{R}^{d}$,

$$
\hat{C}_{n}(x)=\left\{y \in \mathbb{R}: V_{n+1}^{(x, y)} \leq \operatorname{Quantile}\left(1-\alpha ; \sum_{i=1}^{n} p_{i}^{w}(x) \delta_{V_{i}^{(x, y)}}+p_{n+1}^{w}(x) \delta_{\infty}\right)\right\},
$$

where $V_{i}^{(x, y)}, i=1, \ldots, n+1$ are as defined in (2.1), and $p_{i}^{w}, i=1, \ldots, n+1$ are as defined in (3.2). Then $\hat{C}_{n}$ satisfies $\mathbb{P}\left\{Y_{n+1} \in \hat{C}_{n}\left(X_{n+1}\right)\right\} \geq 1-\alpha$.

This is the weighted "full conformal" approach. For the "split conformal" approach, we restate part of Section A. 3 from the supplement to Tibshirani et al. (2019). Let $\left(X_{1}^{0}, Y_{1}^{0}\right), \ldots,\left(X_{m}^{0}, Y_{m}^{0}\right)$ be a proper training set of size $m$, used for fitting the regression function $\hat{f}_{0}$. Also let $\left(X_{1}, Y_{1}\right), \ldots,\left(X_{n}, Y_{n}\right)$ be the calibration set of size $n$ and let $\left(X_{n+1}, Y_{n+1}\right)$ be the test case. Then weighted split conformal prediction is a special case of Corollary 1 in which $\hat{f}_{0}$ is treated as fixed, and e.g. if we use the absolute-residual score function, the prediction interval simplifies to

$$
\hat{C}_{n}(x)=\hat{f}_{0}(x) \pm \text { Quantile }\left(1-\alpha ; \sum_{i=1}^{n} p_{i}^{w}(x) \delta_{\left|Y_{i}-\hat{f}_{0}\left(X_{i}\right)\right|}+p_{n+1}^{w}(x) \delta_{\infty}\right),
$$

with weights $p_{i}^{w}$ as defined in (3.2). By Corollary 1, this has coverage at least $1-\alpha$ conditional on the proper training set $\left(X_{1}^{0}, Y_{1}^{0}\right), \ldots,\left(X_{m}^{0}, Y_{m}^{0}\right)$. Similar results apply for other score functions and for classification problems.

Finally, Remark 3 of Tibshirani et al. (2019) notes that the results above still hold if the likelihood ratio weights have an unknown normalization constant, i.e. if $w \propto \mathrm{~d} \tilde{P}_{X} / \mathrm{d} P_{X}$, because this constant cancels out in the final weights in (3.2).

### 3.2 Unequal probability sampling with replacement

Now, consider independently sampling WR from a finite population where each unit has unequal but known sampling probabilities, such as in PPS sampling. Assume no stratification, clustering, or other design constraints. Our goal is to use this survey sampling design to get prediction intervals that cover $Y$ for most of the $N$ units in the finite population.

Below, we show that this is a special case of the covariate shift setting of Tibshirani et al. (2019). Their distribution $P_{X}$ is replaced with the sampling probabilities for each unit in the finite population, and $\tilde{P}_{X}$ is replaced with a uniform distribution over the same population units. The only randomness is in the sampling; all covariates and response variables are fixed in the population, as is usual in design-based inference.

Let $j$ index the universe of population units from 1 to $N$. Once a sample $S$ is selected, it can be written as a length- $n$ vector of sampled unit IDs, so that the elements of $S$ take values in $\{1, \ldots, N\}$. Often, survey researchers find it easier to work with the random vector $S$ by rewriting it as a length- $N$ vector of sampling indicators $Z_{j}=\{1$ if $j \in S$, and 0 otherwise $\}$. However, in the covariate shift setting, it will be easier to work with $S$ directly. While using $j$ to index the population units, we use $i$ for indexing the elements of $S$. For example, $S_{1}=4$ would mean that the first unit we sampled was the fourth member of the population.

Lemma 3. Assume our training sample is a sample WR of size $n$ where each unit is drawn independently from a finite population of size $N$, with possibly-unequal but nonzero and known sampling probabilities $\pi_{j}$ for each population ID $j \in 1, \ldots, N$. Also assume our test case is a single observation sampled uniformly at random from the population. Let $S$ be a vector whose first $n$ elements are the population IDs of our $n$ training cases, and whose last element is the population ID for our test case.

- Let our finite population consist of $N$ units, each with its own fixed data vector $\left(X_{j}, Y_{j}\right)$ and sampling probability $\pi_{j}$ for $j=1, \ldots, N$. The conditional distribution $P_{(X, Y) \mid S}$ is a deterministic lookup table: Once we have sampled a random $S_{i}$ for some $i \in\{1, \ldots, n+1\}$, we observe its corresponding covariate and response values $\left(X_{S_{i}}, Y_{S_{i}}\right)$.
- Let the training distribution $P_{S}$ correspond to our complex survey design, which consists of $n$ iid draws from a categorical distribution whose categories are simply the unit IDs $j \in\{1, \ldots, N\}$, with known probabilities $\pi_{j}$. (The sampling indicator vector $Z$ is a draw from a multinomial
distribution (sampling WR) or a multivariate hypergeometric distribution (sampling WOR), where the $j^{\text {th }}$ element counts how often that unit was selected. Sampling from a categorical distribution is identical, but instead of recording counts, we record each of the selected unit IDs individually (Tu, 2014).)
- Let the "covariate-shifted" test distribution $\tilde{P}_{S}$ consist of one draw from the same set of unit IDs as the training distribution, but with uniform probability: let $\tilde{\pi}_{j}=1 / N$ for all $j=1, \ldots, N$.

Then this is a special case of the covariate shift setting, with $S$ and $(X, Y)$ playing the roles that $X$ and $Y$ respectively played in (3.1):

$$
\begin{align*}
\left(S_{i},\left(X_{S_{i}}, Y_{S_{i}}\right)\right) & \stackrel{\text { iid }}{\sim} P=P_{S} \times P_{(X, Y) \mid S}, i=1, \ldots, n, \\
\left(S_{n+1},\left(X_{S_{n+1}}, Y_{S_{n+1}}\right)\right) & \sim \tilde{P}=\tilde{P}_{S} \times P_{(X, Y) \mid S}, \text { independently }, \tag{3.3}
\end{align*}
$$

and $w\left(S_{i}\right) \equiv 1 / \pi_{S_{i}}$.
Proof. In the training data sampling design, units are sampled with replacement from a fixed population, and thus they are iid. Even though individual population units have different sampling probabilities, each unit in a training sample is drawn from the same categorical distribution. Also, because each $\left(X_{j}, Y_{j}\right)$ is paired with a fixed $j$ for all $j=1, \ldots, N$, the conditional distribution $P_{(X, Y) \mid S}$ is the same for training and test data. Hence, these training and test distributions match the requirements in (3.1).

Furthermore, since $\pi_{j}$ is nonzero for all population units, we have $w=\mathrm{d} \tilde{P}_{S} / \mathrm{d} P_{s}=(1 / N) / \pi_{S_{i}} \propto 1 / \pi_{S_{i}}$. By Remark 3 of Tibshirani et al. (2019), it is safe to ignore the normalization constant and set $w\left(S_{i}\right)=1 / \pi_{S_{i}}$ directly.

Corollary 2. Assume that $n$ training cases and one test case are drawn as described in Lemma 3. Then Lemma 2, the weighted full conformal results from Corollary 1, and the weighted split conformal results from Section A. 3 of Tibshirani et al. (2019) hold, with $p_{i}^{w}$ defined by using the inverse-probability sampling weights $1 / \pi_{s_{i}}$ for the likelihood ratio weight function $w$ in (3.2).

Proof. In this setting, $P_{S}$ and $\tilde{P}_{S}$ are discrete distributions with identical support, so $\tilde{P}_{S}$ is absolutely continuous with respect to $P_{S}$. All other conditions of Corollary 1 and Section A. 3 of Tibshirani et al. (2019) are met by Lemma 3 .

As suggested by the intuition in Section 2.3, we carry out conformal inference by replacing the exchangeable eCDF with a survey-weighted eCDF in which the first $n$ observations (the training sample) have their usual inverse-probability sampling weights. The $(n+1)^{\text {th }}$ sample (the test case) is assigned the known sampling weight $1 / \pi_{S_{n+1}}$ that population unit $S_{n+1}$ would have had under the original sampling design. The assumption of uniform sampling of test cases simply lets us guarantee marginal coverage across the entire finite population. We do need to know what $\pi_{j}$ would have been for every unit in the population,
which may be reasonable for the organization carrying out the survey but not for end users of the data; Section 5.1 discusses ways to address this issue.

In a situation with no covariates, we can replace $X$ with a constant in Lemma 3, and apply Lemma 2 to get a prediction interval for $Y$ that is not conditional on $X$, although it may depend on $\pi_{S_{n+1}}$.

### 3.3 Unequal probability sampling without replacement

SRSWOR is exchangeable, so the usual conformal methods apply directly. But other kinds of sampling WOR are not a special case of the covariate shift setting above, because the training data are no longer independent. While Tibshirani et al. (2019) do provide more general results under a relaxed condition they call "weighted exchangeability", it is not immediately clear that this condition can account for sampling WOR.

However, if $n \ll N$, then statistical properties derived under sampling WR are often fairly good approximations to the actual properties under assuming sampling WOR. Our simulations in Section 4.2, in which we sample WOR, suggest that this is likely to hold true for conformal methods as well.

### 3.4 Cluster sampling

We cannot apply the covariate shift results to cluster sampling, because the data are not independent. Cluster sampling with unequal probabilities will require further research.

However, in the special case where the clusters themselves are sampled by SRS and the ultimate units are sampled by SRS within each cluster, then we can apply the methods of Dunn et al. (2022). Their paper is framed in terms of a more general two-layer hierarchical setting. They do not explicitly consider a finitepopulation setting, but their assumptions do allow for it (except for some restrictions in their CDF pooling method).

In the framework of Dunn et al. (2022), let $P_{1}, \ldots, P_{k} \sim \Pi$ be $k$ random distributions drawn iid from $\Pi$. From each of the sampled distributions $P_{\ell}$ for $\ell=1, \ldots, k$, we draw $n_{\ell}$ iid observations $\left(X_{\ell 1}, Y_{\ell 1}\right), \ldots$, ( $X_{\ell n_{i}}, Y_{\ell n_{i}}$ ).

The corresponding setup in survey sampling would be cluster sampling, where our finite population of size $N$ is partitioned into a fixed number $K$ of clusters or Primary Sampling Units (PSUs), and we take a sample of these clusters. SRSWR from a finite population is a special case of iid sampling. Hence, if we take a SRSWR of $k<K$ clusters $P_{1}, \ldots, P_{k}$ from the finite population, and then take a SRSWR of $n_{\ell}$ ultimate units from cluster $\ell$ for each $\ell=1, \ldots, k$, then this is a special case of the setup above, and we can apply most of the results in Dunn et al. (2022). We outline one of their approaches briefly here, and the others in Appendix B, but refer readers to their full paper for details.

Although Dunn et al. (2022) state their results in terms of iid sampling, it seems likely that they could be relaxed to exchangeable sampling of the distributions $P_{\ell}$ as well as exchangeable sampling within each
$P_{\ell}$. If so, these results would also apply to SRSWOR, not just SRSWR. Similarly, we conjecture that some of their results could be extended to the "weighted exchangeable" setting of Tibshirani et al. (2019).

Subsampling: By subsampling, Dunn et al. (2022) change the sampling design to become exchangeable. Start with the design above, but then subsample our dataset by choosing one unit at random from each cluster. Then any test case from any new cluster is exchangeable with our subsample. We can treat the $k$ subsampled training cases and the one test case as being generated exchangeably by the process: "Take a cluster at random, then take one observation at random from that cluster", and it is valid to use standard conformal methods.

However, although this process guarantees exact marginal coverage $1-\alpha$ across training sets, it ignores most of the data and leads to wider variability in achieved coverage from training set to training set. An alternative is to carry out repeated subsampling $B$ times and combine the results appropriately across subsamples. Dunn et al. (2022) show how to do this in a way that is guaranteed to have coverage of $1-2 \alpha$, but in practice tends to achieve coverage close to $1-\alpha$.

### 3.5 Stratified sampling

Again, we cannot apply the covariate shift results to stratified sampling, because even though strata are independent of each other, the $n$ samples are not independent.

However, we can safely apply the conformal methods from previous subsections within each stratum separately, if within each stratum independently we have used one of the sampling methods with conformal guarantees. In other words, if the full population is partitioned into $H$ strata, we can treat each stratum $h=1, \ldots, H$ as its own population. To form a prediction interval for a test case from stratum $h$, we apply conformal methods to only the $n_{h}$ training cases from that stratum. Clearly this will guarantee conditional coverage by stratum. If the same coverage level is used simultaneously across all strata, it will also guarantee marginal coverage. This is an example of "group-balanced conformal prediction" or "object-conditional Mondrian conformal prediction" (Vovk, 2013; Vovk et al., 2022; Angelopoulos and Bates, 2022).

This stratum-by-stratum approach will lead to a loss of statistical efficiency, since each stratum's conformal quantiles will be estimated using a sample size $n_{h}<n$. On the other hand, in some situations, prediction intervals may be more useful if we allow their sizes to vary by stratum than if their size has to be constant across strata. Further, guaranteeing coverage conditional on stratum may be more desirable than only guaranteeing marginal coverage, which could be achieved by overcoverage in some strata at the expense of undercoverage in others.

### 3.6 Post-stratification

Imagine our first $n$ observations were drawn SRSWR, but we wish to post-stratify after data collection, and the population size $N_{h}$ of each post-stratum is known. We could reweight each sampled observation
by the relative stratum sizes: unit-level post-stratification weights are proportional to $N_{h} / n_{h}$, where $n_{h}$ is random rather than fixed in advance (Lohr, 2021).

We cannot use such weights for conformal inference and retain our exact finite-sample guarantees under the justifications in the present paper, because it would induce dependence between the training set and test case. However, Fannjiang et al. (2022) extend conformal methods to allow for "feedback covariate shift", where the test distribution is allowed to depend on the observed training data, and this may be a promising direction for future work on post-stratified conformal prediction.

In the meantime, we can treat post-stratification as an approximation to estimating the covariate-shift likelihood ratio weights. Although we lose the exact conformal guarantees, using such an approximation would be just as reasonable as the estimation of covariate-shift weights in general. Specifically, imagine we are sampling SRSWR from two different finite populations: a training population of size $M$, and a test population of size $N$. Both populations have the same $P_{(X, Y) \mid S}$ and the same set of post-strata $1, \ldots, H$, but different (and known) post-stratum sizes $M_{h}, N_{h}$ for $h=1, \ldots, H$. In both cases, our sampling design is equivalent to first choosing a post-stratum at random with probability proportional to post-stratum size, then a unit from within that post-stratum at random.

Now, we apply Lemma 3 - except that we let $P_{S}$ and $\tilde{P}_{S}$ depend on the post-stratum ID $h \in 1, \ldots, H$, not the population unit ID. Then $d P_{S} / d \tilde{P}_{S}=\left(M_{h} / M\right) /\left(N_{h} / N\right)$, so $w\left(S_{i}\right) \propto N_{h_{i}} / M_{h_{i}}$. So far, we have exact guarantees. If we now assume that we do not actually know the true post-stratum sizes for the training population, we can replace $M_{h} / M$ with training-sample estimates $n_{h} / n$ and get post-stratification weights $w\left(S_{i}\right) \propto N_{h_{i}} / n_{h_{i}}$. If we further assume that both populations are actually the same, we are now justified in using conformal methods with the standard post-stratification weights. Our guarantees are approximate only because we estimated the "training" post-stratum sizes.

Similar types of weighting could also be developed in order to apply conformal inference when we do not we wish to assume that test cases will be sampled uniformly but with some other sampling design. For instance, instead of guaranteeing prediction interval coverage across people, perhaps we want to guarantee coverage across visits to the doctor, and we use a sampling distribution to encode our knowledge of different people's propensities to visit the doctor.

## 4. Examples

Our R code and knitted RMarkdown output are available at https://github.com/ColbyStatSvyRsch/ surveyConformal-paper-code.

### 4.1 Real data

We have claimed that conformal methods may work better when they account for the sampling design of the data. As a simple demonstration, we turn to an extract of the Medical Expenditure Panel Survey or

MEPS (Agency for Healthcare Research and Quality, 2017), which is a nationally representative survey about the cost and use of health care among the U.S. civilian noninstitutionalized population.

We chose the MEPS because it has already been used as a benchmark dataset in several conformal inference papers, starting with Romano et al. (2019) and followed by others (Sesio and Candes, 2020; Sesio and Romano, 2021; Feldman et al., 2021; Bai et al., 2022). In each of these papers, the authors randomly partition MEPS data into proper training, calibration, and test sets, then report the coverage and length of conformal prediction intervals (PIs) for various models across many such random partitions. However, none of these papers report accounting for the complex sampling design of MEPS, which includes stratification, clustering, and oversampling of selected subgroups.

At present, we do not attempt a full correction of these earlier analyses of MEPS. We only wish to illustrate that there can be noticeable differences in the conformal PIs depending on whether or not we account for the sampling design, even in a very simple analysis. We use a portion of the public-use dataset for calendar year 2015. We subset to only those respondents who filled out the self-administered questionnaire (SAQ) portion of the survey, and we use the person-level weight variable designed to be used with the SAQ for persons age 18 and older during the interview.

In the poster associated with Romano et al. (2019), available at https://github.com/yromano/cqr/blob/ master/poster/CQR_Poster.pdf, the authors explain that they are predicting "health care utilization, reflecting \# visits to doctor's office/hospital". Following their GitHub code, we define a "utilization" response variable as the sum of five counts for 2015: total number of office-based visits reported; total number of reported visits to hospital outpatient departments; count of all emergency room visits reported; total number of nights associated with hospital discharges; and total number of days where home health care was received from any type of paid or unpaid caregiver.

Unlike the earlier conformal analyses of MEPS, we do take into account the public-use variables for strata, PSUs, and person-level weights. In the 2015 SAQ-eligible subset that we work with, there are 165 strata, and most have 2 or 3 PSUs. First we drop the 4 strata which had no observations in either PSU 1 or PSU 2. Next, for simplicity, we set aside every observation whose PSU is labeled 3 (regardless of stratum) and treat them as our test set. We treat the rest (PSUs 1 and 2) as our overall training set. We split this training data into proper-training and calibration sets under two different approaches. The first approach is a 50/50 SRS split that ignores the survey design. The second approach is to form a random split by PSU within each stratum, so that in each stratum independently, we randomly assign either PSU 1 to propertraining and PSU 2 to calibration or vice versa. These approaches result in proper-training and calibration sets with around 10,000 people each and a test set with 1,659 people.

For each split, we fit a linear regression model to the proper-training set to predict utilization from a subset of the covariates used by Romano et al. (2019): age; sex; indicators for diabetes diagnosis, private insurance coverage, and public insurance coverage; and quantitative summaries of answers to the Physical Component Summary (PCS), the Mental Component Summary (MCS), and the Kessler Index (K6) of
non-specific psychological distress. Higher PCS and MCS scores represent better health, while lower K6 scores represent less distress. We calculate PIs for each test case by combining it with the calibration set and finding conformal quantiles. Developing conformal methods for designs with only one or two PSUs per stratum is still an open problem, not yet addressed by the methods of Sections 3.4 and 3.5. Collapsing strata into pseudo-strata could be a reasonable solution if we had subject matter knowledge of the strata, but the public-use MEPS data uses anonymized stratum IDs. For this reason, our quantiles do not use these methods to handle the clustering and stratification, but they do apply the survey weights as in Section 3.2.

Because we ended up with such large proper-training and calibration sets, relative to the smaller test set, we only saw small differences between the exchangeable and design-based conformal approaches at moderate PI levels. On the other hand, we have enough proper-training and calibration data to estimate $99 \%$ PI levels too, and there we do see substantial differences in the average PI length. Our main takeaways, based on Tables 4.1 and 4.2:

1. When we used a conformal pipeline that assumed exchangeability (data splits at random; no weights in model-fitting; no weights in the conformal quantiles on the calibration set), we tended to over-cover. If we also ignored the weights when using the test set to estimate coverage, these calculations underestimated just how much over-coverage there was. By taking survey-weighted means on the test set, we found slightly higher estimates of coverage. We believe these higher estimates are more appropriate, since the survey-weighted means ought to generalize to the rest of the population better than unweighted means do.
2. When we did use design-based methods (design-based splits; design-based and survey-weighted model fits; and survey-weighted conformal quantiles on the calibration set), this reduced our over-coverage a little, according to the survey-weighted means of coverage on the test set. Similarly, it also made our PIs a little narrower (around one to three fewer utilizations/year) for moderate PI levels, and substantially narrower (around forty fewer utilizations/year) for $99 \%$ PIs.

This brief MEPS example demonstrates that the estimated PI coverages and lengths can differ when conformal methods account for the survey design. In the next subsection, we study these effects in more detail, by repeatedly sampling under known sampling designs from a complete finite population. We also use smaller sample sizes, to see more pronounced differences between using vs. ignoring the survey design.

Table 4.1
Linear regression models' PI coverage, estimated on the MEPS dataset.

| Split/fit/conformal | Test set | $\mathbf{8 0 \%}$ PI covg | $\mathbf{9 0 \%}$ PI covg | $\mathbf{9 5 \%}$ PI covg | 99\% PI covg |
| :--- | :--- | :--- | :--- | :--- | :--- |
| SRS | SRS | $(0.824,0.826)$ | $(0.910,0.911)$ | $(0.951,0.951)$ | $(0.993,0.994)$ |
| SRS | Svy-wtd | $(0.836,0.838)$ | $(0.914,0.915)$ | $(0.950,0.951)$ | $(0.995,0.995)$ |
| Svy-wtd | Svy-wtd | $(0.829,0.831)$ | $(0.911,0.913)$ | $(0.949,0.950)$ | $(0.992,0.993)$ |

Notes: SRS = simple random sampling; PI = prediction interval; MEPS = Medical Expenditure Panel Survey.
Coverages reported as approximate $95 \%$ CIs for the average, based on 100 random proper-training/calibration splits at each setting, using the same test set each time. When data splits, proper-training-set model fits, and calibration-set conformal quantiles ignored the survey design, we over-covered (especially for lower PI levels); but when test-set estimates of coverage also ignored the sampling design, they underestimated the amount of overcoverage, compared to test-set estimates that did account for the sampling design. However, when splits, fits, and conformal quantiles accounted for the survey design, there was slightly less over-coverage.

Table 4.2
Linear regression models' PI lengths, estimated on the MEPS dataset.

| Split/fit/conformal | Test set | $\mathbf{8 0 \%}$ PI length | $\mathbf{9 0 \%}$ PI length | $\mathbf{9 5 \%}$ PI length | 99\% PI length |
| :--- | :--- | :---: | :---: | :---: | :---: |
| SRS | Either | $(28.9,29.2)$ | $(43.9,44.3)$ | $(60.2,60.7)$ | $(250.5,256.6)$ |
| Svy-wtd | Svy-wtd | $(27.3,27.7)$ | $(40.9,41.4)$ | $(57.4,58.0)$ | $(202.0,211.1)$ |

Notes: $\quad$ SRS = simple random sampling; PI = prediction interval; MEPS = Medical Expenditure Panel Survey.
PI lengths reported as approximate $95 \%$ CIs for the average, based on 100 random proper-training/calibration splits at each setting, using the same test set each time. When data splits, proper-training-set model fits, and calibration-set conformal quantiles ignored the survey design, our PI lengths tended to be slightly larger than when splits, fits, and conformal quantiles did account for the survey design - or much larger when the PI level is very high. In the Table's first row, it does not matter whether or not test-set estimates were surveyweighted, because these PI lengths are constant across test-set cases for a given data split and PI level.

### 4.2 Simulations

For our design-based simulations, we used the Academic Performance Index (API) data (California Department of Education, 2018) from R's survey package (Lumley, 2021). The apipop dataset contains information on 37 variables for all 6194 California schools (elementary, middle, or high school) with at least 100 students. The dataset vintage is not documented, but appears to be the 1999-2000 academic year, since the data includes API scores for each school for 1999 and 2000.

We used the apipop dataset as the finite population, and repeatedly took samples (with around $n=200$ ultimate sampling units) using various designs. When we evaluated our results on test sets, we used the entire finite population - including those cases that had already been used to fit models or find conformal quantiles - because this corresponds to the guarantees that our paper makes in Section 3. Simulation details:

- In all simulations, sampling was done without replacement. Although our results in Section 3 assume sampling with replacement, we conjectured that sampling without replacement would still lead to conformal coverage close to nominal, and we wanted to check this empirically.
- All simulations were run 1,000 times. All $95 \%$ confidence intervals are calculated as the estimate $\pm 2$ times the SD over $\sqrt{1,000}$.
- After dropping the rows with missing values for enroll and mobility, the full "finite population" consisted of the 6,153 schools without missing values for any variables used in the simulations.
- Most simulations used a sample size of $n=200$ schools. However, the cluster samples had $n \approx 200$ schools on average but varied across samples. The regression model simulations took PPS samples of size $m+n=400$, then split the samples at random into a proper training set of $m=200$ and a calibration set of $n=200$.
- The response variable was usually api00, the school's API in 2000. The exception is Table 4.5 , where the response variable was enroll, the same variable used to create the PPS weights. For the nonregression simulations, we simply sought "unsupervised" prediction intervals for the marginal distribution of the response variable (with no covariates). For the regression simulations, we found quantiles of $|y-\hat{f}(x)|$ on the calibration set and sought prediction intervals for the response variable at the covariate values for each unit in the population.
- PPS sampling probabilities (if used) were usually proportional to enroll, the number of students enrolled at the school. The exception is parts of Tables 4.8 and 4.9 , where PPS probabilities were proportional to 1 plus the square root of the residuals from the full-population linear regression model, in order to see the effects of over-sampling cases that are hard to fit well. Conformal quantiles for the PPS simulations were calculated as in Section 3.2.
- Clusters (if used) were based on dnum, the school district number. Cluster samples always took a SRS of 24 school districts. 24 was chosen because it led to an average of $n=198$ schools (close to the $n=200$ used in other sampling designs). For the "survey-design-aware" results in Table 4.6, we calculated quantiles using the "subsampling once" method as in Section 3.4, while the design-unaware results ignored clustering and calculated quantiles on the whole dataset.
- Strata (if used) were based on stype, the school type. Stratified samples always took 100 elementary, 50 middle, and 50 high schools, with an SRS within each school type. For the "survey-design-aware" results in Table 4.7, we calculated quantiles separately by stratum as in Section 3.5, while the designunaware results ignored strata and calculated quantiles on the whole dataset.
- Linear regression models always predicted apio0 using a linear combination of ell (the percentage of English language learners), meals (the percentage of students eligible for subsidized meals), and mobility (the percentage of students for whom this is the first year at the school).

Our main takeaways:

1. Across many settings, using the naive quantile (the $\lceil n \alpha\rceil$ order statistic) instead of the conformal quantile (the $\lceil(n+1) \alpha\rceil$ order statistic) tended to give slight undercoverage. The conformal quantile helped partly to fix this; but in non-SRS settings it was not enough of a fix on its own.
2. For SRS designs, the conformal quantile lemma worked as advertised. See Table 4.3.
3. For PPS designs, ignoring the weights gave slight undercoverage when weights were not highly informative about the response variable. On the other hand, ignoring the weights led to extreme overcoverage when weights were highly informative. In both cases, weighted conformal quantiles fixed the problem. See Tables 4.4 and 4.5.
4. For clustered designs, as well as for stratified designs, ignoring the design undercovered but accounting for the design (including conformal-quantile padding) did fix it. The one exception was for one of the cluster-design simulations, where the design-based conformal PIs did not reach the target coverage. This might have been due to the small number of clusters, large variation in cluster sizes, and our choice of "subsampling once" as the conformal method. See Tables 4.6 and 4.7.
5. For simple models and split-conformal inference, if the weights were not highly informative about model variables or the fit of the model, then it did not make much difference whether or not the weights were used for quantiles. But when the weights were informative, we saw that unweighted conformal quantiles over-covered (and PIs were too wide). Using survey-weights in model-fitting was not enough to fix it, but weighting the quantiles was. See Tables 4.8 and 4.9.

Table 4.3
SRS. Average PI coverage of api00, at two different PI levels, under 1,000 SRS samples of $\boldsymbol{n}=200$ each from API dataset.

| Conformal? | $\mathbf{8 0 \%}$ PI coverage | 90\% PI coverage |
| :--- | :---: | :---: |
| no | $(0.794,0.801)$ | $(0.945,0.949)$ |
| yes | $(0.800,0.807)$ | $(0.949,0.953)$ |

Notes: Coverages reported as $95 \%$ confidence intervals. Naive quantiles undercover, but conformal quantiles achieve target coverage.

Table 4.4
Uninformative PPS. Average PI coverage of api00, at two different PI levels, under 1,000 PPS samples of $n=200$ each from API dataset where probability $\propto$ enroll

| Survey-weighted? | Conformal? | $\mathbf{8 0 \%}$ PI coverage | 90\% PI coverage |
| :--- | :--- | :---: | :---: |
| no | no | $(0.752,0.760)$ | $(0.927,0.932)$ |
| no | yes | $(0.759,0.767)$ | $(0.935,0.940)$ |
| yes | no | $(0.795,0.803)$ | $(0.944,0.949)$ |
| yes | yes | $(0.803,0.811)$ | $(0.953,0.958)$ |

Notes: Coverages reported as $95 \%$ confidence intervals. Naive quantiles under-cover; conformal quantiles alone or survey-weighting alone do not fix it; but survey-weighted conformal quantiles achieve target coverage.

Table 4.5
Informative PPS. Average PI coverage of enroll, at two different PI levels, under $\mathbf{1 , 0 0 0}$ PPS samples of $n=200$ each from API dataset where probability $\propto$ enroll.

| Survey-weighted? | Conformal? | $\mathbf{8 0 \%}$ PI coverage | 90\% PI coverage |
| :--- | :--- | :---: | :---: |
| no | no | $(0.933,0.935)$ | $(0.986,0.987)$ |
| no | yes | $(0.934,0.937)$ | $(0.988,0.989)$ |
| yes | no | $(0.792,0.798)$ | $(0.946,0.949)$ |
| yes | yes | $(0.796,0.802)$ | $(0.948,0.951)$ |
| Notes: | Coverages reported as $95 \%$ confidence intervals. Naive quantiles | over-cover; conformal quantiles alone do not fix it, while survey- |  |

Notes: Coverages reported as $95 \%$ confidence intervals. Naive quantiles over-cover; conformal quantiles alone do not fix it, while surveyweighting alone under-covers; but survey-weighted conformal quantiles achieve target coverage.

Table 4.6
Clustering. Average PI coverage of api00, at two different PI levels, under $\mathbf{1 , 0 0 0}$ clustered samples of 24 clusters ( $n \approx 200$ ) each from API dataset.

| Survey-design? | Conformal? | $\mathbf{8 0 \%}$ PI coverage | 90\% PI coverage |
| :--- | :--- | :---: | :---: |
| no | no | $(0.785,0.805)$ | $(0.934,0.943)$ |
| no | yes | $(0.791,0.811)$ | $(0.940,0.949)$ |
| yes | no | $(0.799,0.817)$ | $(0.916,0.930)$ |
| yes | yes | $(0.799,0.817)$ | $(0.959,0.968)$ |

Notes: Coverages reported as $95 \%$ confidence intervals. Due to high variability in cluster sizes, these $95 \%$ CIs are wider than in previous tables, but overall trend is generally similar to other tables: Naive quantiles appear likely to under-cover; conformal quantiles alone or survey-design-aware analyses alone do not necessarily fix it; but survey-design-aware conformal quantiles achieve target coverage.

Table 4.7
Stratification. Average PI coverage of api00, at two different PI levels, under $\mathbf{1 , 0 0 0}$ stratified samples of $n=200$ each from API dataset.

| Survey-design? | Conformal? | $\mathbf{8 0 \%}$ PI coverage | 90\% PI coverage |
| :--- | :--- | :---: | :---: |
| no | no | $(0.769,0.777)$ | $(0.933,0.938)$ |
| no | yes | $(0.775,0.783)$ | $(0.940,0.944)$ |
| yes | no | $(0.787,0.795)$ | $(0.939,0.944)$ |
| yes | yes | $(0.800,0.808)$ | $(0.953,0.956)$ |
| Notes: Coverages reported as $95 \%$ confidence intervals. Naive quantiles under -cover; conformal quantiles alone or survey-design-aware analyses |  |  |  | alone do not fix it; but survey-design-aware conformal quantiles achieve target coverage.

Table 4.8
Linear regression models' PI coverage. Average PI coverage of api00, at two different PI levels, under $\mathbf{1 , 0 0 0}$ PPS samples of $\boldsymbol{n}=\mathbf{2 0 0}$ each from API dataset.

| PPS probs | Svy-wtd conformal? | Svy-wtd regression? | 80\% PI covg | 90\% PI covg |
| :---: | :---: | :---: | :---: | :---: |
| enroll | no | no | (0.807, 0.810) | (0.955, 0.957) |
| enroll | yes | no | (0.803, 0.807) | (0.954, 0.956) |
| residuals | no | no | (0.874, 0.876) | (0.973, 0.974) |
| residuals | no | yes | (0.875, 0.878) | (0.973, 0.974) |
| residuals | yes | no | (0.798, 0.802) | (0.950, 0.951) |
| residuals | yes | yes | (0.798, 0.801) | (0.950, 0.951) |
| Notes:Cove  <br>  diffe <br>  conf <br>  fit su | as $95 \%$ confidence inter or not we weight the co over-cover, and this is fi regression models. | For weights proportiona mal quantiles. For inform by survey-weighted confo | 11 (uninformativ ghts proportional tiles; but it makes | gression), it mak residuals, un-w rence whether or |

Table 4.9
Linear regression models' PI lengths. Average length of PIs for api00, at two different PI levels, under 1,000 PPS samples of $\boldsymbol{n}=\mathbf{2 0 0}$ each from API dataset.

| PPS probs | Svy-wtd conformal? | Svy-wtd regression? | $\mathbf{8 0 \%}$ PI length | 90\% PI length |
| :--- | :--- | :--- | :---: | :---: |
| enroll | no | no | $(192.6,194.1)$ | $(294.5,297.3)$ |
| enroll | yes | no | $(190.8,192.7)$ | $(293.7,297.2)$ |
| residuals | no | no | $(214.1,215.6)$ | $(328.2,331.5)$ |
| residuals | no | yes | $(211.8,213.3)$ | $(331.4,334.8)$ |
| residuals | yes | no | $(179.4,180.8)$ | $(285.4,287.7)$ |
| residuals | yes | yes | $(175.1,176.3)$ | $(286.1,288.3)$ |

Notes: Lengths reported as $95 \%$ confidence intervals. For weights proportional to enroll (uninformative for the regression), it makes little difference whether or not we weight the conformal quantiles. For informative weights proportional to full-pop residuals, PIs from unweighted conformal quantiles are much wider than PIs from survey-weighted conformal quantiles; but it makes little difference whether or not we fit survey-weighted regression models.

Overall, the survey-conformal quantiles we proposed mathematically in Section 3 also appear to work empirically. Data analysts will likely get coverage closer to nominal when they account for the weights or other survey design features.

## 5. Extensions

We discuss several practical considerations: What if the sampling design does not quite match the situations above? What if the sampling probabilities are not all known? We also suggest other possible use cases for conformal inference in survey methodology.

### 5.1 Practical considerations

Weighting adjustments: Most surveys are not released with inverse-probability sampling weights alone. The final survey weights have been adjusted for nonresponse, post-stratification, and other considerations. Tibshirani et al. (2019) found that their conformal methods still maintained coverage close to nominal even when they only estimated the likelihood ratio weights, instead of using true likelihood ratios. We anticipate
similar outcomes for conformal methods that use adjusted sampling weights instead of the true inverseprobability sampling weights.

Distribution shift: Surveys are typically assumed to be sampled from a well-defined finite population, in a specific time and place. We may not be guaranteed coverage if we make predictions for units at future times or from distinct populations. If model-based (rather than design-based) inference makes more sense for the application at hand, we may still be able to estimate the covariate shift likelihood ratio and apply the conformal techniques of Tibshirani et al. (2019); it will simply no longer be strictly design-based inference, though a joint superpopulation/design-based framework may be a fruitful topic for future work (Isaki and Fuller, 1982; Rubin-Bleuer and Şchiopu-Kratina, 2005; Han and Wellner, 2021).

Unknown sampling probabilities for test cases: Our conformal methods of Section 3.2 require the sampling probabilities for each test case. However, in practice the sampling probabilities are not typically known for every population unit. Even if they are known internally within the survey organization, public release of the entire population's sampling probabilities can increase the risk of unit re-identification and privacy breaches.

Instead, the survey organization could report a set of population categories for which the sampling probabilities are approximately equal within each category, along with their approximate probabilities. There may be categories that are broad enough to minimize privacy risks, but fine enough to approximate the real sampling probabilities well. Or instead of discrete categories, the survey organization could report a kind of generalized variance function or GVF (Wolter, 2007), but used to estimate each population unit's sampling probabilities rather than variances, based on covariates available for each unit. This may be especially reasonable for PPS.

Alternatively, the survey organization could release only the value of the single largest inverseprobability sampling weight for the whole population. Then, using that one weight for every test case during conformal prediction would be strictly conservative. Or as an approximation, users could use the largest sampling weight in the public-use dataset. Users could also try a range of plausible weights for a desired test case, based on the weights of similar in-sample units, and report a sensitivity analysis.

Finally, if a user is interested in predictions for a population unit with the same vector of covariates $X$ as one of the sampled cases in the dataset, they could simply use that sampled case's survey weight.

### 5.2 Other uses for conformal methods in survey methodology

National statistical offices, polling agencies, and any others who collect and pre-process survey data may find their own use cases for conformal methods:

- Build a response propensity model based on internal metadata from past surveys. For the next survey, build conformal prediction sets for each sampled unit's most likely mode of response (or ultimate status of nonresponse). Use these sets to help choose the mode of initial contact for each sampled unit.
- As part of automated quality control checks, use conformal prediction intervals or sets for each variable to flag potential outliers for followup (Bates, Candès, Lei, Romano and Sesia, 2023).
- To impute for item nonresponse, or to generate synthetic microdata, draw at random from a conformal prediction interval or set for that variable.


## 6. Conclusion

There is growing interest in extending machine learning (ML) methods to complex survey data, as well as a distinct body of work around developing prediction intervals for new predictive methods such as those arising from the ML community. While this ongoing work is valuable, design-based conformal inference provides an alternative approach for both needs: We can apply a novel predictive algorithm to complex survey data (even if that algorithm has not been specifically adapted to account for the survey design yet), and automatically get conformal intervals compatible with the fitted prediction function (even if native prediction-interval methods have not been developed for that algorithm yet), and still manage to provide exact, finite-sample, design-based coverage guarantees.

Of course, design-based conformal methods are not a panacea. First, for methods that do have wellunderstood design-based adaptations, the design-based version fitted to complex-survey-design training data is likely to be a better predictive model for the conditional mean and consequently to have narrower conformal prediction intervals than when that method ignores the sampling design. As a simple example, in certain cases a survey-weighted linear model may generalize better than an unweighted linear model fit to the same data, so the survey-weighted model's smaller residuals will lead to narrower conformal prediction intervals than for the unweighted model. In this sense, it is still important to develop survey-weighted equivalents of novel ML algorithms.

Second, for models that do have native prediction intervals, if we can justifiably trust that the underlying model assumptions are met, then we may be able to get narrower or less-variable prediction intervals natively than from conformal inference. And in some cases, native methods can give us conditional prediction intervals rather than marginal ones. In this sense, it is still important to develop native prediction intervals for specific models.

But when either of these conditions is not yet met, conformal inference is a practical and assumptionlean way to fill in the gap. We also note that Lei et al. (2018) found in simulations that conventional methods are so noisy for high-dimensional regression that conformal prediction intervals can actually be narrower than native ones.

We have also discussed several pragmatic limitations to conformal inference with survey data: Sampling weights are not truly known in advance, nor can they be reported for every population unit. Our suggested solutions are only a starting point.

Still, we hope that this paper spurs interest in conformal inference among survey statisticians. Survey data analysts might find these methods directly applicable. Furthermore, research into conformal methods
from the design-based perspective could bring new insights back to the wider statistics/ML community of conformal inference researchers.

As one example, Vovk (2013) shows that the distribution of achieved coverage levels has a Beta distribution across calibration sets when data are exchangeable. In order to gauge how much variability to expect from this Beta distribution, Angelopoulos and Bates (2022) suggest one estimate of the effective sample size $n_{\text {eff }}$ for weighted conformal methods, but their estimate is not appropriate for all sampling designs nor for all estimators. Survey statisticians may be able to suggest better estimates of $n_{\text {eff }}$ or recommend other ways to study and control the variability in achieved coverage for non-exchangeable data.

Other open problems include conformal methods for unequal-probability sampling WOR; unequalprobability cluster sampling; strata with few PSUs; combining strata rather than analyzing each stratum separately; exact coverage guarantees for post-stratification; panel survey designs; or joint superpopu-lation/design-based frameworks. We encourage survey researchers to contribute to conformal methodology and find new areas of application for these methods.

## Acknowledgements

The author thanks the Editor, the Associate Editor, and the anonymous referees for their helpful suggestions which improved this manuscript. The author is also grateful to Ryan Tibshirani, Robin Dunn, Benjamin LeRoy, and Evan Randles for their early feedback on several of the ideas presented here.

## Appendix

## A. Adaptive prediction regions

In the simple approach to conformal prediction for regression described in our paper, we have used a constant-width prediction band everywhere, which might be unrealistic for many scenarios. If the true conditional distribution of $Y \mid X$ is heteroscedastic for instance, marginal coverage will still be correct, but it will be achieved by overcovering at some regions of $X$ and undercovering at others. For "adaptive" alternatives, where the PI is wider at regions of $X$ with more variability in $Y$, see recent work on locallyweighted conformal inference (Lei et al., 2018) and conformalized quantile regression (Romano et al., 2019).

For classification problems, even the simple approach of Section 2.5 produces adaptive prediction sets: for hard test cases where $\hat{f}$ is uncertain about the right class, prediction sets will be larger than for easy test cases where $\hat{f}$ confidently assigns most of the probability to one class. See also Romano et al. (2020) and Angelopoulos and Bates (2022).

Varying PI widths due to survey weighting: Note that in the survey-weighted setting, we will automatically get slightly different prediction interval widths at different test cases, because the survey weights can differ for each unit being predicted. However, the effect of survey weighting is to widen slightly the
prediction intervals for regions of $X$ with a smaller effective sample size. This is distinct from adaptivity to regions with more variability in $Y$.

Loosely, a sampled unit with a larger sampling weight represents more population units, so we are more uncertain about units "like" this one. When we apply Lemma 3 to a test case with a larger weight, its surveyweighted eCDF is pushed down farther than if it had a small weight; so the nonconformity score quantile is estimated farther to the right; so the conformal prediction interval is wider for units with larger weights, all else being equal.

On the other hand, consider optimal allocation designs, in which strata with higher variance of $Y$ are oversampled. When we construct conformal intervals separately within each stratum, the high-variance strata will have their eCDFs padded by a smaller $1 / n_{h}$ than low-variance strata do, so the quantile adjustment is less conservative. But this should be more than offset by the fact that high-variance strata also have wider spread in $Y$, so that ultimately their prediction intervals will end up wider than those of lowvariance strata.

## B. Other conformal methods for cluster samples

Beyond Section 3.4, we briefly note how Dunn et al. (2022)'s other methods relate to cluster designs in survey sampling.

Prediction for an observed cluster: If we only need a prediction interval for new observations from one of the clusters we already sampled, we can simply apply standard conformal methods by only using that cluster's data, which will be exchangeable.

Double conformal: For unsupervised prediction (where we want a prediction interval for $Y$ without conditioning on covariates $X$ ) for a new cluster, one could create conformal prediction intervals separately within each cluster, then combine their endpoints appropriately into a single interval. This lines up with the design-based spirit: construct valid estimates within each cluster, then combine them sensibly across clusters. Dunn et al. (2022) derive such a method that is guaranteed to have coverage at least $1-\alpha$. However, in simulations it overcovers, with coverage of nearly 1 and wider intervals than other approaches.

Pooling CDFs: We could first construct eCDFs within each cluster and average them together into one pooled eCDF. Then we could apply standard conformal methods using this pooled eCDF. Dunn et al. (2022) only prove that this is asymptotically valid, requiring a continuous distribution for $Y$ as well as a growing number of sampled clusters $k \rightarrow \infty$. This is not possible in the traditional design-based setting of a fixed finite population, although we could construct a superpopulation model and a sequence of growing finite populations that satisfies their requirements.

In simulations, Dunn et al. (2022) find that CDF pooling tends to have coverage closest to nominal as well as shortest length of prediction intervals across most settings. But if we do not wish to rely on
asymptotic arguments and continuous $Y$ data, repeated subsampling appears to work better than single subsampling or the double conformal method.

## References

Agency for Healthcare Research and Quality (2017). Medical Expenditure Panel Survey 2015 full year consolidated data file. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp? cboPufNumber=HC-181.

Angelopoulos, A.N., and Bates, S. (2022). A gentle introduction to conformal prediction and distributionfree uncertainty quantification. arXiv preprint arXiv:2107.07511.

Bai, Y., Mei, S., Wang, H., Zhou, Y. and Xiong, C. (2022). Effcient and differentiable conformal prediction with general function classes. arXiv preprint arXiv:2202.11091.

Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2021). The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA, 10(2), 455-482.

Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2022). Conformal prediction beyond exchangeability. arXiv preprint arXiv:2202.13415.

Bates, S., Candès, E., Lei, L., Romano, Y. and Sesia, M. (2023). Testing for outliers with conformal pvalues. The Annals of Statistics, 51(1), 149-178.

Bechtold, W.A., and Patterson, P.L. (2005). The Enhanced Forest Inventory and Analysis Program National Sampling Design and Estimation Procedures.

Bersson, E., and Hoff, P.D. (2022). Optimal conformal prediction for small areas. arXiv preprint arXiv:2204.08122.

California Department of Education (2018). Academic Performance Index. https://www.cde.ca.gov/re/pr/ api.asp.

Central Statistical Office, Zambia (2015). Living Conditions Monitoring Survey (LCMS). https://www.zamstats.gov.zm/portfolio/living-conditions-monitoring-survey-lcms/.

Cherian, J., and Bronner, L. (2020). How the Washington Post estimates outstanding votes for the 2020 presidential election. https://s3.us-east-1.amazonaws.com/elex-models-prod/2020-general/write-up/ election_model_writeup.pdf.

Chernozhukov, V., Wüthrich, K. and Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. Conference On Learning Theory, 732-749. PMLR.

Crookston, N.L., and Dixon, G.E. (2005). The forest vegetation simulator: A review of its structure, content, and applications. Computers and Electronics in Agriculture, 49(1), 60-80.

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

Dagdoug, M., Goga, C. and Haziza, D. (2022). Model-assisted estimation in high-dimensional settings for survey data. Journal of Applied Statistics.

Dunn, R., Wasserman, L. and Ramdas, A. (2022). Distribution-free prediction sets for two-layer hierarchical models. Journal of the American Statistical Association.

Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.

Fannjiang, C., Bates, S., Angelopoulos, A.N., Listgarten, J. and Jordan, M.I. (2022). Conformal prediction under feedback covariate shift for biomolecular design. PNAS, 119(43), e2204569119.

Feldman, S., Bates, S. and Romano, Y. (2021). Improving conditional coverage via orthogonal quantile regression. arXiv preprint arXiv:2106.00394.

Fienberg, S.E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. Journal of Privacy and Confidentiality, 1(2).

Fong, E., and Holmes, C.C. (2021). Conformal Bayesian computation. Advances in Neural Information Processing Systems, 34, 18268-18279.

Gibbs, I., and Candes, E. (2021). Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34, 1660-1672.

Han, Q., and Wellner, J.A. (2021). Complex sampling designs: Uniform limit theorems and applications. The Annals of Statistics, 49(1), 459-485.

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

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77(377), 89-96.

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

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

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094-1111.

LeRoy, B., and Schafer, C. (2021). Conformal prediction for simulation models. ICML 2021 Workshop on Distribution-Free Uncertainty Quantification. https://benjaminleroy.github.io/documents/icml2021/ conformal_prediction_for_simulation_models.pdf.

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

Lumley, T. (2021). survey: Analysis of complex survey samples. R package version 4.1-1.

Lumley, T., and Scott, A. (2017). Fitting regression models to survey data. Statistical Science, 265-278.

Lunde, R. (2023). On the validity of conformal prediction for network data under non-uniform sampling. arXiv preprint arXiv:2306.07252.

Manokhin, V. (2022). Awesome conformal prediction. https://github.com/valeman/awesome-conformalprediction.

McConville, K.S., Breidt, F.J., Lee, T. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. Journal of Survey Statistics and Methodology, 5(2), 131-158.

National Center for Health Statistics (2016). The Second Longitudinal Study of Aging (LSOA II). https://www.cdc.gov/nchs/lsoa/lsoa2.htm.

Oliveira, R.I., Orenstein, P., Ramos, T. and Romano, J.V. (2022). Split conformal prediction for dependent data. arXiv preprint arXiv:2203.15885.

Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002). Inductive confidence machines for regression. European Conference on Machine Learning, 345-356. Springer.

Romano, Y., Patterson, E. and Candes, E. (2019). Conformalized quantile regression. Advances in Neural Information Processing Systems, 32.

Romano, Y., Sesia, M. and Candes, E.J. (2020). Classification with valid and adaptive coverage. arXiv preprint arXiv:2006.02544.

Rubin-Bleuer, S., and Şchiopu-Kratina, I. (2005). On the two-phase framework for joint model and designbased inference. The Annals of Statistics, 33(6), 2789-2810.

Sande, L.S., and Zhang, L.-C. (2021). Design-unbiased statistical learning in survey sampling. Sankhyā $A$, 83(2), 714-744.

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

Sesio, M., and Candes, E.J. (2020). A comparison of some conformal quantile regression methods. Stat, 9(1), e261.

Sesio, M., and Romano, Y. (2021). Conformal prediction using conditional histograms. Advances in Neural Information Processing Systems, 34.

Thompson, S.K. (1997). Adaptive sampling in behavioral surveys. In The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates, volume 167 of NIDA Research Monograph, (Eds., L. Harrison and A. Hughes), 296-319. National Institutes of Health.

Tian, Q., Nordman, D.J. and Meeker, W.Q. (2022). Methods to compute prediction intervals: A review and new results. Statistical Science, 37(4), 580-597.

Tibshirani, R.J., Barber, R.F., Candes, E. and Ramdas, A. (2019). Conformal prediction under covariate shift. Advances in Neural Information Processing Systems, 32.

Tu, S. (2014). The Dirichlet-Multinomial and Dirichlet-Categorical models for Bayesian inference. https://stephentu.github.io/writeups/dirichlet-conjugate-prior.pdf. Unpublished manuscript.

Vovk, V. (2013). Conditional validity of inductive conformal predictors. Machine Learning, 92, 349-376.

Vovk, V., Gammerman, A. and Shafer, G. (2022). Algorithmic Learning in a Random World. Springer, $2^{\text {nd }}$ edition.

Wieczorek, J., Guerin, C. and McMahon, T. (2022). K-fold cross-validation for complex sample surveys. Stat, 11(1), e454.

Wolter, K. (2007). Generalized variance functions. Introduction to Variance Estimation. Springer, $2^{\text {nd }}$ edition.

Yang, S., and Kim, J.K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. The Econometrics of Complex Survey Data, volume 39 of Advances in Econometrics, 209234. Emerald Publishing Limited.

# Sample designs and estimators for multimode surveys with face-to-face data collection 

J. Michael Brick and Jill M. DeMatteis ${ }^{1}$


#### Abstract

Survey researchers are increasingly turning to multimode data collection to deal with declines in survey response rates and increasing costs. An efficient approach offers the less costly modes (e.g., web) followed with a more expensive mode for a subsample of the units (e.g., households) within each primary sampling unit (PSU). We present two alternatives to this traditional design. One alternative subsamples PSUs rather than units to constrain costs. The second is a hybrid design that includes a clustered (two-stage) sample and an independent, unclustered sample. Using a simulation, we demonstrate the hybrid design has considerable advantages.


Key Words: Bias; Multi-stage; Subsampling; Two-phase; Web-push.

## 1. Introduction

Many surveys use web and mail data collection modes due to their relatively low costs, but the overall response rates may be lower than desired, and the estimates may be subject to considerable nonresponse bias (Dillman, 2017; Brick, Kennedy, Flores-Cervantes and Mercer, 2021). Introducing face-to-face (ftf) interviewing as a follow-up mode can substantially increase response rates, although at a higher cost. The increased response rates achieved through ftf follow-up typically also result in reduced nonresponse bias, due to the introduction of the ftf mode resulting in increased response propensities for subgroups that are less likely to respond to the web or mail modes. A sampling approach to make multimode data collection with ftf interviewing more cost-efficient is two-phase sampling. Hansen and Hurwitz (1946) introduced two-phase sampling as a method of dealing with nonresponse based on the two-phase sampling theory of Neyman (1938). Hansen and Hurwitz applied their method with a first phase sample of retail establishments that were sent a mail questionnaire, with non-respondents subsampled for ftf. Since their ftf efforts resulted in virtually $100 \%$ response after weighting for subsampling, the estimates from the survey were unbiased.

A well-known application of this sampling strategy is the American Community Survey (ACS) where a sample of households is selected within tabulation areas and the households are requested to respond to a web or mail questionnaire; a subsample of nonrespondents within the area are followed up fff (U.S. Census Bureau, 2014; U.S. Census Bureau, 2019). Like other surveys today, the ACS does not achieve full response so the survey weights must be further adjusted for nonresponse and the estimates are still subject to potential nonresponse bias.

This article examines two new sample designs for multimode household surveys that uses ftf interviewing to increase response rates. Surveys with this type of multimode design may be more common in the U.S. and Canada. The standard two-phase sampling approach selects a sample of primary sampling units (PSUs) and households within the PSU for the low-cost mode(s) and then subsamples within each

PSU for ftf follow-up. The two new approaches alter the designs for the low cost (web/mail) to improve the efficiency of the estimates while retaining the basic ftf design. Designs that use more expensive modes before the lower cost modes are not considered (e.g., Bayart and Bonnel, 2015). The first new sample design we explore is a variant where a large sample of PSUs is sampled and households in each PSU are recruited by low-cost modes in the first phase, but the fff follow-up is only done in a subsample of PSUs. This approach reduces the clustering effect by spreading the respondents from the first phase sample over a larger number of PSUs. We also consider a second approach that selects two independent samples - one an unclustered sample of households that is recruited by low-cost modes and a second clustered sample where all sampled households are recruited sequentially using all modes. The estimates are created by compositing the data from the two samples. We do not discuss nonsampling errors that may arise when more than one mode data collection is used (e.g., Goodman, Brown, Silverwood, Sakshaug, Calderwood, Williams and Ploubidis, 2022).

In Section 2, we describe the two new sample designs in more detail. We present estimators and describe their properties under different nonresponse models in Section 3. Section 4 describes and gives the results of a simulation study we conducted. We conclude in Section 5 with some discussion of the implications of the design options and estimators, recommendations, and areas for future research.

## 2. Sampling designs

All the applications of two-phase household sampling for nonresponse where ftf interviewing is used that we have identified begin with a first phase sample of PSUs (clustered geographically) and then subsample nonrespondents for the second phase from each of the first phase PSUs. In many cases, the subsampling for nonresponse is an adaptive or responsive design feature rather than an initial design approach (Groves and Heeringa, 2006; Heeringa, Wagner, Torres, Duan, Adams and Berglund, 2004; Wagner, Arrieta, Guyer and Ofstedal, 2014). The first phase is a sample of PSUs with a sample of households from an address-based sampling (ABS) frame. All sampled households are subject to the initial data collection protocol that may involve web and/or mail as a low-cost mode.

Two-phase unit subsampling. The nonrespondents to the first-phase sample are subsampled within each PSU and ftf interviewing is the mode for the second phase. The ACS uses this design. We refer to this standard approach as two-phase unit subsampling. Särndal and Swensson (1987) extended the theory of Hansen and Hurwitz to designs where the first phase sample was not a simple or stratified random sample.

Two statistical issues arise in two-phase unit subsampling design. One concern is the increase in variance of estimates due to subsampling (Kish, 1992) since the weights account for the subsampling. The second issue is the increase in variance due to clustering, assuming a positive intraclass correlation for the characteristics of interest, because all responses are clustered within the sampled PSUs. In addition, a related cost and operational issue is limiting the number of sampled PSUs to make the ftf data collection within the PSUs effective.

Two-phase PSU subsampling. An alternative two-phase design is to select a large number of PSUs and households in the first phase and then select a subsample of the PSUs and all of the households within those PSUs for the second phase. All households sampled in the first phase are recruited by web, but only the nonrespondents in the subsample of PSUs are followed up in the second phase by ftf. We refer to this approach as two-phase PSU subsampling. We have not identified any surveys using this approach. This approach has benefits and concerns similar to those with two-phase unit subsampling, but has the advantage of reducing the clustering effect by spreading the respondents from the first phase sample over a larger number of PSUs.

Hybrid sampling. A different design with fewer restrictions than two-phase PSU subsampling selects two independent samples and then composites the estimates from the two samples to produce final estimates. The first sample is an unclustered sample of households from the ABS frame and those households are recruited only by low-cost modes. The second sample uses a two-stage design to reduce data collection costs, with a sample of PSUs and households within the PSUs. The data are collected from households in this second sample by sequentially using web and then ftf modes. This design builds on both dual frame (Lohr, 2011) and two-phase methods. Like dual frame methods it selects two samples, although from the same frame in this case, and then combines the data from the two samples. Like two-phase sampling, only a subsample of the full sample of households is subject to the full data collection protocol. We refer to this as hybrid two-phase sampling or more concisely as hybrid sampling. When discussing the two samples, we refer to them as the unclustered and clustered samples, respectively, for ease of discussion.

Example. Here, we present an illustration of each design and consider the effects on precision; later, for each of these design alternatives, we will examine the bias of various estimators. As a simple example of the three approaches, suppose the goal is to complete 10,000 household interviews with 70 percent done by web and 30 percent by ftf. For illustration, we ignore nonresponse weighting adjustments and details about costs that are discussed later.

The three approaches for this simple example are illustrated in Figure 2.1. The two-phase unit subsampling approach selects a sample of 200 PSUs with probability proportionate to the number of households, and then an equal number of households is selected in each PSU. All sampled households are pushed to web and a subsample of web nonrespondents in each PSUs are followed by ftf. Assuming a $25 \%$ web response rate and $50 \% \mathrm{ftf}$ response rate, then sampling 140 households per PSU yields 35 web completes. Subsampling roughly 30 of the approximately 105 nonrespondents per PSU results in 15 ftf completes for a total of 50 responses ( 35 by web and 15 by ftf ). The design effect due to subsampling (differential weighting) is approximately 1.44 (Kish, 1992) and reaches this maximum value when the outcome and the weights are uncorrelated. The design effect due to clustering is approximately $1+\delta(\bar{m}-1)$, where $\bar{m}$ is the average number of completed households per PSU and $\delta$ is the intraclass correlation. With 50 completes per PSU and $\delta=0.02$, the clustering design effect is about 2.0. The overall design effect is the product of the weighting and clustering effects, or 2.9 , and results in an effective sample size of about $3,500(10,000 / 2.9)$.

Figure 2.1 Illustration of three follow-up approaches.

## Population

$\square$ PSUs • Households (hhs)


Nonrespondents to Web


Subsample for FtF


Notes: PSUs = primary sampling units; NRs = nonrespondents.

In the two-phase PSU subsampling approach, a sample of 700 PSUs is selected with probability proportional to the number of households, and 200 PSUs are subsampled with equal probability. The number of PSUs for the subsample was chosen to equal the number in the unit subsampling design, and 700 total PSUs allows taking all nonrespondents within the 200 subsampled PSUs without further subsampling. In each PSU an equal probability sample of 40 households is selected. All the sampled households are sent to web, resulting in approximately 7,000 web completes $(700 * 40 * 0.25)$. All of the web nonrespondents in the 200 PSU subsample are followed up ftf, yielding $3,000 \mathrm{ftf}$ completes. The design effect due to weighting is still 1.44. Due to the unequal number of completes per cluster, we use the approximate design effect suggested by Holt (1980), with $m^{\prime}=\sum m_{i}^{2} / \sum m_{i}$ instead of $\bar{m}$ where $m_{i}$ is the number of completes in PSU $i$. In this design $m_{i}=10$ in 500 PSUs and is $m_{i}=25$ in 200 PSUs, so $m^{\prime}=17.5$ and the design effect due to clustering is $1+\delta\left(m^{\prime}-1\right)=1.33$ when $\delta=0.02$. The overall design effect is 1.9 and the effective sample size is just over 5,200, a substantial increase in precision over the traditional two-phase approach.

With the hybrid approach a sample of 200 PSUs (equal to the number of PSUs where ftf is done in the other designs) is selected and 40 households per PSU in the clustered sample, yielding $2,000 \mathrm{web}$ completes and $3,000 \mathrm{ftf}$ completes. The unclustered sample is 20,000 to yield $5,000 \mathrm{web}$ completes. The design effect due to clustering in the clustered sample is 1.48 when $\delta=0.02$ and there is no differential weighting effect (assuming a uniform nonresponse adjustment is applied to all respondents). In all, there are 7,000 web and $3,000 \mathrm{ftf}$ completes. The two samples are composited using, for example, $\lambda=0.7$ for the unclustered sample (since it $70 \%$ of the total) and $1-\lambda=0.3$ for the clustered sample, so the overall design effect is 1.14 . (An optimal compositing factor, which also takes into account the clustering effect on precision for the clustered sample, could be determined and applied; however, for simplicity in this illustration, we chose compositing factors proportional to the number of completes from the particular mode. We examine the effects of alternative compositing factors later in the manuscript.) The effective sample size is 8,770 , a substantial increase in precision over both two-phase approaches.

## 3. Estimation and nonresponse models

## Nonresponse models

In the literature on models of survey nonresponse, two frameworks have emerged: a deterministic framework that partitions the population into two mutually exclusive, exhaustive groups, respondent and nonrespondents (as described on pages 359-363 of Cochran, 1977); and a stochastic framework in which each member of the population has a probability of responding to a particular survey (see, for example, Brick and Montaquila, 2009).

Using the deterministic view of nonresponse under a given data collection protocol, the population contains a set of households who will respond to the web request, a set of households that will respond to the ftf request after not responding on the web, and a set of households that will not respond at all. With this model, the population total is

$$
\begin{equation*}
Y=N\left[\gamma_{W} \bar{Y}_{W}+\gamma_{F} \bar{Y}_{F}+\left(1-\gamma_{W}-\gamma_{F}\right) \bar{Y}_{N}\right], \tag{3.1}
\end{equation*}
$$

where $\gamma_{W}$ is proportion responding by web, $\gamma_{F}$ is proportion responding by ftf but not by web, and $\bar{Y}_{W}$, $\bar{Y}_{F}$, and $\bar{Y}_{N}$ are the corresponding population means of the characteristic of interest for the web, ftf , and nonresponding sets. The deterministic view with the fixed partitioning of the population by response mode under the data collection protocol is conceptualized from a post-data collection perspective, where repeated implementations of the same protocol could be used to determine these constants.

A stochastic model that aligns with this deterministic model assumes each unit in the population has a response propensity vector with the first element the probability of responding by web and the second element the probability of responding by ftf and not by web, say $\phi_{k}=\left(\phi_{k, W}, \phi_{k, F}\right)^{\prime}$, where $0 \leq \phi_{k, W} \leq 1$; $0 \leq \phi_{k, F} \leq 1 ; 0<\phi_{k, W}+\phi_{k, F} \leq 1$. Note that if we define $\phi_{k, F \mid W^{c}}$ to be the conditional probability unit $k$ responds by ftf given that they did not respond by web, then $\phi_{k, F}=\left(1-\phi_{k, W}\right) \phi_{k, F \mid W^{c}}$. In addition, taking expectations over this response distribution gives $E_{R} \sum_{k \in U} \phi_{k, W} \equiv \gamma_{W} N ; E_{R} \sum_{k \in U} \phi_{k, W} y_{k} \equiv Y_{W}$ where $U$ denotes the population, $Y_{W}=\gamma_{W} N \bar{Y}_{W}$, and $E_{R} \sum_{k \in U} \phi_{k, F} \equiv \gamma_{F} N ; E_{R} \sum_{K \in U} \phi_{k, F} y_{k} \equiv Y_{F}$ where $Y_{F}=\gamma_{F} N \bar{Y}_{F}$. With full response (i.e., $\phi_{k, F \mid W^{c}}=1$, so that $\phi_{k, W}+\phi_{k, F}=1 \forall k$ ), we have $Y=N\left[\gamma_{W} \bar{Y}_{W}+\gamma_{F} \bar{Y}_{F}\right]$.

## Estimators

First, suppose we follow all web nonrespondents by ftf, i.e., a two-stage sample with no subsampling. The typical approach to estimation is to use the Horvitz-Thompson (HT) estimator with an adjustment for nonresponse. (Here, we consider a single overall adjustment, ignoring more complex nonresponse and calibration adjustments for simplicity.) This estimator of the total is

$$
\begin{align*}
\hat{t}_{1} & =\sum_{k \in S} d_{k} \delta_{k}(W) \hat{R}^{-1} y_{k}+\sum_{k \in S} d_{k} \delta_{k}(F) \hat{R}^{-1} y_{k} \\
& =\hat{N}\left[\frac{\hat{\gamma}_{W}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{W}+\frac{\hat{\gamma}_{F}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{F}\right], \tag{3.2}
\end{align*}
$$

where $S$ denotes the sample, $d_{k}$ is the reciprocal of the probability of selection of household $k$ (accounting for both stages of selection), $\delta_{k}(W)=1$ if household $k$ responds to the web survey and is 0 otherwise, $\delta_{k}(F)=1$ if household $k$ responds to the ftf survey and is 0 otherwise, $\hat{N}=\sum_{k \in S} d_{k}$, and $\hat{R}=\hat{R}_{W}+$ $\left(1-\hat{R}_{W}\right) \hat{R}_{F}$ where $\hat{R}_{W}=\sum_{k \in S} d_{k} \delta_{k}(W) / \sum_{k \in S} d_{k}$ and $\hat{R}_{F}=\sum_{k \in S} d_{k} \delta_{k}(F) / \sum_{k \in S} d_{k}\left(1-\delta_{k}(W)\right)$ are the observed web and ftf response rates. Note that $\hat{R}_{F}$ is the conditional response rate given no response to web. Throughout, we estimate $\hat{\gamma}_{W}=\hat{R}_{W}$ and $\hat{\gamma}_{F}=\left(1-\hat{R}_{W}\right) \hat{R}_{F}$ by the observed response rates where those depend upon the specific design.

The estimator $\hat{t}_{1}$ is unbiased if $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ or if the ftf conditional response rate is $100 \%$. This property is easily shown by taking expectations with respect to both the sampling and response distributions (Särndal and Swensson, 1987). We assume throughout that standard conditions for the appropriate full response estimator to be unbiased also hold.

In contrast to $\hat{t}_{1}$, which applies a constant adjustment $\hat{R}^{-1}$ to all respondents, we also consider an estimator commonly used in the two-phase sampling context, which adjusts only the ftf respondents, as follows:

$$
\begin{align*}
\hat{t}_{2} & =\sum_{k \in S} d_{k} \delta_{k}(W) y_{k}+\sum_{k \in S} d_{k} \delta_{k}(F) \hat{R}_{F}^{-1} y_{k}  \tag{3.3}\\
& =\hat{N}\left[\hat{\gamma}_{W} \bar{y}_{W}+\left(1-\hat{\gamma}_{W}\right) \bar{y}_{F}\right] .
\end{align*}
$$

The estimator $\hat{t}_{2}$ is unbiased if $\bar{Y}_{N}=\bar{Y}_{F}$ or if the ftf conditional response rate is $100 \%$.
Next, we consider the subsampling designs and extend these estimators to incorporate the subsampling. Let $\omega$ denote the conditional probability of selection into the subsample for ftf interviewing. With twophase unit subsampling with the second phase subsampling rate of $\omega(0<\omega \leq 1)$, the extensions of these estimators are

$$
\begin{align*}
\hat{t}_{1} & =\sum_{k \in S} d_{k} \delta_{k}(W) \hat{R}^{-1} y_{k}+\sum_{k \in S} d_{k} \omega^{-1} \delta_{k}(F) \hat{R}^{-1} y_{k} \\
& =\hat{N}\left[\frac{\hat{\gamma}_{W}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{W}+\frac{\hat{\gamma}_{F}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{F}\right] \tag{3.4}
\end{align*}
$$

and

$$
\begin{align*}
\hat{t}_{2} & =\sum_{k \in S} d_{k} \delta_{k}(W) y_{k}+\sum_{k \in S} d_{k} \omega^{-1} \delta_{k}(F) \hat{R}_{F}^{-1} y_{k}  \tag{3.5}\\
& =\hat{N}\left[\hat{\gamma}_{W} \bar{y}_{W}+\left(1-\hat{\gamma}_{W}\right) \bar{y}_{F}\right] .
\end{align*}
$$

In this expression the subsampling rate is accounted for in estimating $\gamma_{W}$ (Table 3.1 shows how this and other estimators are written in terms of weights). Note that the estimators given in equations (3.2) and (3.3) are a special case of those in equations (3.4) and (3.5), respectively, where $\omega=1$.

Table 3.1
Estimators, weights and nonresponse models for two-phase and hybrid sampling.

| Estimator | Design | Respondent weight |  | Nonresponse model |
| :--- | :--- | :--- | :--- | :--- |
| $\hat{t}_{1}$ | Two stage | $d_{k} \hat{R}^{-1}$ | Web respondents in $S_{B}$ | $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ |
|  |  | $d_{k} \omega^{-1} \hat{R}^{-1}$ | Ftf respondents in $S_{B}$ |  |
| $\hat{t}_{2}$ | Two phase | $d_{k}$ | Ftf respondents in $S_{B}$ | $\bar{Y}_{N}=\bar{Y}_{F}$ |
|  |  | $d_{k} \omega^{-1} \hat{R}_{F}^{-1}$ | Ftf respondents in $S_{B}$ |  |
| $\hat{t}_{A}$ | One stage, unclustered | $d_{k} \hat{R}_{W}^{-1}$ | All (web) respondents in $S_{A}$ | $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ |
| $\hat{t}_{d f, 1}$ | Hybrid | $\lambda d_{k} \hat{R}_{W}^{-1}$ | All (web) respondents in $S_{A}$ | $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ |
|  |  | $(1-\lambda) d_{k} \hat{R}^{-1}$ | All respondents in $S_{B}$ |  |
| $\hat{t}_{d f, 2}$ | Hybrid | $\kappa d_{k}$ | All (web) respondents in $S_{A}$ | $\bar{Y}_{N}=\bar{Y}_{F}$ |
|  |  | $(1-\kappa) d_{k}$ | Web respondents in $S_{B}$ |  |
|  |  | $d_{k} \hat{R}_{F}^{-1}$ | Ftf respondents in $S_{B}$ |  |

Note: Ftf = face-to-face.

For the subsampling designs, both estimators are unbiased if the ftf conditional response rate is $100 \%$; $\hat{t}_{2}$ is unbiased if $\bar{Y}_{N}=\bar{Y}_{F}$, while $\hat{t}_{1}$ requires the more stringent condition that $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$. This result applies to both two-phase unit and PSU subsampling. While Särndal and Swensson (1987) did not consider two-phase PSU subsampling, their proof of the unbiasedness applies since the subsampling they consider is not specific to a phase. Thus, the two-phase extension of the estimator $\hat{t}_{2}$ shown in equation (3.5) can be used with either two-phase unit subsampling or two-phase PSU subsampling.

In the PSU subsampling design, the weighting adjustment in $\hat{t}_{2}, \omega^{-1}$, is the ratio of the number of first phase PSUs to the number of PSUs in the second phase. An alternative subsampling adjustment is

$$
\begin{equation*}
\omega_{s}^{-1}=\frac{\sum_{k \in S} d_{k}\left(1-\delta_{k}(W)\right)}{\sum_{k \in \mathrm{sub}-\mathrm{PSU}} d_{k}\left(1-\delta_{k}(W)\right)}, \tag{3.6}
\end{equation*}
$$

where the numerator is the sum of the weights of all sampled cases that did not respond by web and the denominator is the sum of the weights of the web nonrespondents in the subsampled PSUs. This adjustment incorporates the size of the subsampled PSUs, where size is the number of subsampled households.

For hybrid sampling, we use dual frame notation to simplify the presentation. Let $S_{A}$ be an unclustered sample of households, and $S_{B}$ be a two-stage sample with households sampled within PSUs. The protocol for $S_{A}$ just uses web, while for $S_{B}$ web is followed by ftf for all web nonrespondents.

Since $S_{A}$ is a single stage web data collection, the Horvitz-Thompson estimator or (3.2) with only web respondents is

$$
\begin{equation*}
\hat{t}_{A}=\frac{\sum_{k \in S_{A}} d_{k}}{\sum_{k \in S_{A}} d_{k} \delta_{k}(W)} \sum_{k \in S_{A}} d_{k} \delta_{k}(W) y_{k}=\hat{N}_{S_{A}} \bar{y}_{W A}, \tag{3.7}
\end{equation*}
$$

where $\hat{N}_{S_{A}}=\sum_{k \in S_{A}} d_{k}$ and $\bar{y}_{W A}$ is the estimated mean for web respondents based on the unclustered sample $S_{A}$. This unclustered estimator is unbiased if $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ as discussed for $\hat{t}_{1}$.

Rewriting $\hat{t}_{1}$ for a clustered sample $S_{B}$ gives

$$
\begin{equation*}
\hat{t}_{B, 1}=\hat{N}_{S_{B}}\left[\frac{\hat{\gamma}_{W}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{W B}+\frac{\hat{\gamma}_{F}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{F B}\right], \tag{3.8}
\end{equation*}
$$

where $\hat{N}_{S_{B}}=\sum_{k \in S_{B}} d_{k}$, and $\bar{y}_{W B}$ and $\bar{y}_{F B}$ are the estimated means for web respondents and ftf respondents, respectively, based on the clustered sample $S_{B}$.

Compositing the two estimators for the hybrid samples gives (assuming $\hat{N}=\hat{N}_{S_{A}}=\hat{N}_{S_{B}}$ )

$$
\begin{align*}
\hat{t}_{d f, 1} & =\lambda \hat{t}_{A}+(1-\lambda) \hat{t}_{B, 1} \\
& =\hat{N}\left[\lambda \bar{y}_{W A}+(1-\lambda) \frac{\hat{\gamma}_{W}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{W B}+(1-\lambda) \frac{\hat{\gamma}_{F}}{\hat{\gamma}_{W}+\hat{\gamma}_{F}} \bar{y}_{F B}\right] . \tag{3.9}
\end{align*}
$$

This estimator, like (3.2), does not place more weight on the ftf respondents and is unbiased when $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$ as discussed previously. The more stringent nonresponse model is required for unbiasedness because while $\hat{t}_{B, 1}$ estimates the total for the population consisting of both the web and ftf domains, $\hat{t}_{A}$ estimates the total for only the web domain.

Another option is to composite two estimators but to use the two-phase version of (3.5) for the clustered sample. First, we composite the web respondents from $S_{A}$ and $S_{B}$ and then adjust the weight of ftf respondents to account for the remaining nonresponse. This composite estimator (where only the web samples are composited) is

$$
\begin{equation*}
\hat{t}_{d f, 2}=\hat{N} \hat{\gamma}_{W}\left[\kappa \bar{y}_{W A}+(1-\kappa) \bar{y}_{W B}\right]+\hat{N}\left(1-\hat{\gamma}_{W}\right) \bar{y}_{F B} . \tag{3.10}
\end{equation*}
$$

This estimator is unbiased when $\bar{Y}_{N}=\bar{Y}_{F}$ or when the ftf conditional response rate is $100 \%$ using the same arguments as for (3.5). This estimator is efficient since it includes both the clustered and unclustered sample observations. The estimate of $\hat{\gamma}_{W}$ is the observed proportion of web respondents based on both samples.

If the assumed nonresponse model is $\bar{Y}_{N}=\bar{Y}_{F}$ or with full ftf response, then $\hat{t}_{2}$ is an unbiased estimator when the design is either two-phase unit subsampling or two-phase PSU subsampling and $\hat{t}_{d f, 2}$ is an unbiased estimator when the design is hybrid sampling. More efficient estimators ( $\hat{t}_{1}$ and $\hat{t}_{d f, 1}$ ) are available but require the more stringent assumption $\bar{Y}_{N}=\bar{Y}_{F}=\bar{Y}_{W}$.

The only remaining parameters that need to be specified for the hybrid estimators are the compositing factors. The usual approach is a composite factor $\lambda(0 \leq \lambda \leq 1)$ equal to the effective relative sample size and this is often a reasonable approximation. For $\hat{t}_{d f, 1}, \lambda$ might be set to be the ratio of the effective number of respondents in $S_{A}$ divided by the sum of that and the effective number of respondents in $S_{B}$, where the design effect for $S_{B}$ is estimated as described in the earlier example. A similar approach could be used for $\kappa$, but since only the web respondents are being composited, the effective sample sizes are only those for the web respondents. We explore different compositing factors in the simulation and find that reasonable choices have little effect on the variances of the estimates.

## Variance estimation

Variance estimation methods for most of the estimators are well-known or require only minor adjustments to handle the approaches proposed here for nonresponse follow-up (i.e., two-stage, unclustered estimators and hybrid sampling). For example, hybrid sampling is covered by dual-frame estimation theory (Lohr, 2011). Variance estimation for two-phase sampling with complex sampling schemes has been the subject of several recent theoretical developments including Hidiroglou (2001), Hidiroglou, Rao and Haziza (2009) and Beaumont, Béliveau and Haziza (2015).

Variance estimation for two-phase PSU subsampling requires some extensions of Beaumont, Béliveau and Haziza (2015) to ensure the variances are appropriately estimated. The only design we considered here
involves sampling a large number of PSUs in the first phase and then taking all the nonrespondents in the second phase in a subsample of the PSUs. The subsampling of PSUs for the second phase does not depend on the outcomes of the first phase sample (we also assume that every sampled first phase PSU has some nonrespondents). Since this design selects all the nonrespondents in the subsampled PSU it satisfies the invariance and independence conditions (see Särndal, Swensson and Wretman, 1992, pages 134-135) for two-stage sampling. Provided the first stage sampling fraction is negligible, the standard with replacement variance estimation can be applied.

We used Taylor series linearization for our simulation but Beaumont, Béliveau and Haziza (2015) describe how replication methods apply equally well in this situation. The sample design for two-phase PSU subsampling typically selects a stratified sample of PSUs. (Our simulation used an unstratified sample, but the generalization to a stratified sample is straightforward.) In each stratum, the number of PSUs selected is determined so that the units treated as variance strata (or clusters) are balanced with respect to the subsampling for follow-up. This combining of PSUs gives an unbiased estimate of the variance (Lu, Brick and Sitter, 2006). For example, with a $50 \%$ subsample of PSUs for the second phase, one approach is to sample 4 PSUs in each stratum and pair these to form 2 variance units so that each variance unit has one PSU subsampled for follow-up and one with no follow-up. With a $33 \%$ subsample of PSUs, 6 PSUs are sampled per stratum in the first phase and 2 variance units are formed and each contains 2 PSUs not subsampled for follow-up and 1 subsampled for follow-up.

## 4. Simulation study

To examine the three sample designs and estimators, we conducted a simulation study. In this section, we begin by describing the approach we used to generate the populations for this simulation. Next, we lay out our simulation design, including a description of the measures we used to evaluate and compare the designs and estimators. Finally, we present the simulation results, beginning with those for the hybrid design scenarios, followed by a comparison of the subsampling designs to the hybrid design.

## Generating the population

For this simulation, we used the 2015-2019 ACS Public Use Microdata Sample (PUMS) data for the 50 States and D.C., treating all ACS respondents residing in households as our population. ACS data collection begins with a mailed invitation to complete the survey on the web. Nonrespondents are followed up first by mailing a paper questionnaire (mail), then by computer-assisted telephone interviewing (CATI) or computer-assisted personal interviewing (CAPI) for a subsample of nonrespondents ( $\mathrm{ft} f$ ). (Although CATI was dropped as a follow-up mode for ACS after September 2017, that is not relevant to our simulation, as all CATI and CAPI respondents are treated as ftf respondents for our simulation purposes.) On the ACS PUMS files, each household is associated with a geographic cluster called a Public Use Microdata Area (PUMA). We used PUMAs as primary sampling units (PSUs) for the clustered samples. In selecting
clustered samples, we sampled PSUs with probabilities proportional to the number of households in the PSU, and sampled households with equal probabilities within PSUs, where the conditional probabilities of selection were proportional to the reciprocal of the PSU selection probabilities. For each unclustered sample, we sampled households using a simple random sample without replacement.

In the previous section, we discussed the nonresponse model under which the estimator is unbiased for each estimator. To examine each estimator's performance under various response models, we created four different pseudopopulations ( $\mathrm{A}, \mathrm{B}, \mathrm{C}$, and D ) by varying the approach used to identify respondents and nonrespondents. For pseudopopulations A, B, and C, we considered all ACS web respondents to be our web-push respondents. For pseudopopulation A, we designated all ACS mail respondents to be our ftf respondents, and all ACS ftf respondents to be our nonrespondents. For pseudopopulation B, among the ACS mail and ftf respondents, we randomly identified half to be our ftf respondents and the other half to be our nonrespondents. Pseudopopulation C has a $100 \%$ response with the ftf follow-up, where all ACS mail and ftf respondents were our ftf respondents. Our motivation for constructing pseudopopulation D was to create a population with a lower response rate to the initial mode of contact; thus, we defined all ACS mail respondents as our web respondents, and as we did in creating pseudopopulation B, we randomly identified half of the remainder to be our ftf respondents and the other half to be our nonrespondents.

We included several demographic and socioeconomic variables at the household level in this simulation. Figure 4.1 gives brief descriptions of each of the variables and displays the population means of each of the variables by ACS mode of completion, where the full ACS PUMS dataset is the population. For pseudopopulation A, we expect all estimators to be biased to some extent since the nonrespondents differ from both groups of respondents. The ftf follow-up, when properly accounted for in the estimator, would be expected to reduce bias in estimates for variables v1, v3, v8, and v11 based on Figure 4.1. For pseudopopulations B and D, the response model $\bar{Y}_{N}=\bar{Y}_{F}$ holds by design, so estimators based on that response model should be unbiased. For pseudopopulation C, there is no nonresponse after the ftf followup so estimators that appropriately weight the follow-up should be unbiased.

## Simulation design

The simulations involved selecting samples from the various pseudopopulations (depicted in Table 4.1) and, for each variable, computing estimated totals using each of the estimators presented in Section 3. The simulation scenarios are listed in Table 4.2. The scenarios are labeled according to the combination of parameters; e.g., scenario B2U is the scenario that uses pseudopopulation B, selects a clustered sample only, and uses two-phase unit subsampling for ftf nonresponse follow-up.

For scenarios A1A, B1A, C1A, and D1A, each of which examines the hybrid design with different pseudopopulations, we selected clustered and unclustered samples of size 2,500 each, with 50 PSUs for the clustered sample. To allow for a direct comparison of the hybrid design approach (specifically, scenario B 1 A ) to the unit subsampling and PSU subsampling approaches, we included scenarios B2P and B2U.

Figure 4.1 Population means for variable from American Community Survey 2015-2019 PUMS, by mode. Of all responses, web is $\mathbf{4 9 . 3} \%$, mail is $\mathbf{2 9 . 1 \%}$, and ftf is $\mathbf{2 1 . 7 \%}$.


Notes: $\mathrm{ACS}=$ American Community Survey; CAPI $=$ Computer-assisted personal interviewing; CATI $=$ Computer-assisted telephones interviewing.
Variable definitions: v1-high school graduate; v2-some college; v3-Bachelor's or higher; v4-1-person household; v5-2-person household; v6-3 or more person household; v7-Hispanic reference person; v8-Black, nonHispanic reference person; v9-renter; v10-person 60 years or older; v11*-household income/110,000; v12- percent urban.

Table 4.1
Pseudopopulation definitions.

| Pseudo- <br> population | ACS Web Respondents | ACS Mail Respondents | ACS CATI/CAPI Respondents |
| :---: | :---: | :---: | :---: |
|  | Web | Ftf | NR |
| B | Web | $50 \% \mathrm{Ftf}$, | $50 \% \mathrm{Ftf}$, |
|  |  | $50 \% \mathrm{NR}$ | $50 \% \mathrm{NR}$ |
| C | Web | Ftf | Ftf |
| D | $50 \% \mathrm{Ftf}$, | $50 \% \mathrm{Ftf}$, |  |
|  | $50 \% \mathrm{NR}$ | $50 \% \mathrm{NR}$ |  |
| Notes: | ACS $=$ American Community Survey; CAPI = Computer-assisted personal interviewing; CATI $=$ Computer-assisted telephones |  |  | interviewing; $\mathrm{Ftf}=$ face-to-face; $\mathrm{NR}=$ nonrespondent.

More details on the sample sizes and subsampling fractions used in each of these designs are given in the Appendix (Table A.1). These designs were constructed to yield the same expected total number of completes $(3,127.7)$ and the same number by mode.

We ran 5,000 independent iterations of each scenario. For the hybrid design scenarios, we computed each of the estimators shown in Table $3.1\left(\hat{t}_{1}, \hat{t}_{2}, \hat{t}_{A}, \hat{t}_{d f, 1}\right.$, and $\left.\hat{t}_{d f, 2}\right)$. For the two subsampling scenarios,
we computed $\hat{t}_{1}$ and $\hat{t}_{2}$. For the two-phase PSU subsampling scenario (B2P), we also computed the variant on the estimator $\hat{t}_{2}$ that uses the alternative subsampling adjustment $\omega_{s}^{-1}$ given in (3.6).

Table 4.2
Simulation scenarios.

| Scenario | Pseudo-population | Design | Nonresponse follow-up |
| :--- | :---: | :--- | :--- |
| A1A | A | 1 (independent clustered and unclustered samples) | A (all) |
| B1A | B | 1 (independent clustered and unclustered samples) | A (all) |
| C1A | C | 1 (independent clustered and unclustered samples) | A (all) |
| D1A | D | 1 (independent clustered and unclustered samples) | A (all) |
| B2P | B | 2 (clustered sample only) | P (two-phase primary sampling <br> unit (PSU) subsampling) |
| B2U | B | 2 (clustered sample only) | U (two-phase unit subsampling) |

For each iteration, we computed the relative bias (RB), coefficient of variation (CV), relative root mean squared error (RRMSE), and an indicator of whether the confidence interval (CI) covered the population total (for a normal 95 percent CI) for each estimator, and averaged each of those measures across iterations. The relative bias (RB) and RRMSE were computed relative to the population parameter (i.e., by dividing the bias and RMSE, respectively, by the population total). We used Taylor series linearization to compute the variance estimates as discussed previously. The RB, CV, RRMSE, and CI coverage are the measures we used to evaluate the estimators.

## Results: Hybrid design scenarios

Figures 4.2 and 4.3 present the RB and RRMSE results, and the CV and CI results are shown in arXiv:2303.13303v1 [stat.ME], for each of the hybrid design simulation scenarios. For scenario A1A, the scenario with the most inherent bias due to the way the pseudopopulation is defined, the simulation results demonstrate that all of the estimators are biased for at least some characteristics. For characteristics such as $\mathrm{v} 1, \mathrm{v} 3, \mathrm{v} 8$, and v 11 , where bringing in the ftf follow-up is expected to reduce bias, we see that the estimators that appropriately incorporate the ftf follow-up ( $\hat{t}_{2}$ and $\hat{t}_{d f, 2}$ ) are generally less biased; for variables such as v5, v7, and v9, where we would expect the ftf follow-up to increase bias (see Figure 4.1), no increase in bias is apparent for the estimators that incorporate ftf follow-up. In this scenario, CI coverage is generally very poor (well below the nominal 95 percent level) due to the bias in the estimates. The hybrid composite estimator $\hat{t}_{d f, 2}$ is generally comparable to the estimator $\hat{t}_{2}$ in terms of RRMSE.

For pseudopopulations B and D, the web respondents differ from the ftf respondents and the nonrespondents, but the assumption $\bar{Y}_{N}=\bar{Y}_{F}$ holds in expectation. A key difference between these two pseudopopulations is that the expected web response rate is considerably higher in pseudopopulation B than in pseudopopulation D (49 percent vs. 29 percent). As expected with this nonresponse model, the estimators $\hat{t}_{2}$ and $\hat{t}_{d f, 2}$ have the smallest relative bias and the best confidence interval coverage rates. For most
characteristics examined, estimator $\hat{t}_{d f, 2}$ has the smallest relative RRMSE. The estimators that adjust the weights of web respondents for nonresponse, $\hat{t}_{4}, \hat{t}_{d f, 1}$, and $\hat{t}_{1}$, exhibit more bias and poor confidence interval coverage.

Figure 4.2 Relative biases of each estimator for each of the hybrid design scenarios.


With pseudopopulation C, complete response is attained after the ftf follow-up. However, the characteristics of the web respondents generally differ from the characteristics of the ftf respondents. In scenario C 1 A , the relative bias is negligible for $\hat{t}_{2}$ and $\hat{t}_{d f, 2}$, the estimators that assume $\bar{Y}_{N}=\bar{Y}_{F}$. In this scenario, because $R=1$, the estimator $\hat{t}_{1}$ reduces to the Horvitz-Thompson (base-weighted) estimator and exhibits properties similar to $\hat{t}_{2}$ and $\hat{t}_{d f, 2}$. Additionally, the confidence interval coverage is closest to nominal levels for these three estimators, whereas the confidence interval coverage of the other estimators is generally very poor.

As noted in Section 1, a further consideration is the choice of compositing factors for the hybrid estimators. To examine this, we computed $\hat{t}_{d f, 2}$, for each iteration of the hybrid design scenarios two waysonce using the near optimal compositing factor described in Section 2 and once using the compositing factor fixed at $\kappa=0.2$, which is far from the optimum. The results, presented in arXiv:2303.13303v1 [stat.ME], demonstrate that the choice of compositing factor has virtually no effect on the bias; for all scenarios except

A1A, the same is true for the effect of the compositing factor on confidence interval coverage. The choice of compositing factor does have a small effect on CV and RRMSE, with the optimal factor resulting in less variable estimates.

Figure 4.3 Relative root mean squared error of each estimator for each of the hybrid design scenarios.


## Results: Comparison of subsampling designs to hybrid design scenario

Next, we consider the scenarios that involve subsampling for nonresponse follow-up, scenario B2U (subsampling units) and scenario B2P (subsampling PSUs). Because these scenarios involve clustered sample only, the only estimators that are relevant to these scenarios are $\hat{t}_{1}$ and $\hat{t}_{2}$. The hybrid design is an alternative to these subsampling designs, and in the previous section we demonstrated that for the hybrid design, as expected based on dual-frame estimation theory, the composite estimator $\hat{t}_{d f, 2}$ performed best. For all three scenarios, the estimator $\hat{t}_{2}$ has negligible bias while the bias in $\hat{t}_{1}$ is evident. (See Table A. 2 in the Appendix.) Thus, our evaluation involves a comparison of the results involving $\hat{t}_{2}$ for scenarios B2U (subsampling units) and B2P (subsampling PSUs) to the results for $\hat{t}_{d f, 2}$ in scenario B1A (hybrid design, no subsampling). For these comparisons, Figures 4.4 and 4.5 present the RB and RRMSE. The CV and CI coverage results are given in arXiv:2303.13303v1 [stat.ME].

Figure 4.4 Relative biases of each estimator for the hybrid design scenario vs. the subsampling scenarios.


Figure 4.5 Relative root mean squared error of each for the hybrid design scenario vs. the subsampling scenarios.


There are no appreciable differences in bias for $\hat{d}_{d f, 2}$ and $\hat{t}_{2}$ across the designs. The estimator $\hat{t}_{2}$ with the subsampling designs is generally more variable than the composite estimator $\hat{t}_{d f, 2}$ with the hybrid design, based on the CV and RRMSE. Confidence interval coverages are comparable for all three scenarios with their associated estimators.

For the PSU subsampling scenario, we evaluated the variation of the estimator $\hat{t}_{2}$ that uses the alternative subsampling adjustment $\omega_{s}^{-1}$ given in equation (3.4) by comparing it to the estimator that uses the reciprocal of the PSU subsampling rate as the adjustment $\omega^{-1}$. The results of this comparison, presented in arXiv:2303.13303v1 [stat.ME], are that the estimators are comparable with respect to bias, but the alternative estimator performs marginally better with respect to CV, RRMSE, and CI coverage.

Table A. 2 in the Appendix presents summaries, averaged across variables, for all of the estimators considered under each scenario.

## 5. Discussion

As response rates have declined and survey costs have increased, survey researchers have sought approaches to combat these trends. In this paper, we have presented design and estimation approaches for multimode data collection and have used a simulation study to examine their effectiveness. While our focus is on web for the first phase of data collection, our findings and recommendations also pertain to designs in which mail (used alone or in combination with web, as in a web-push approach) is used in the first phase.

Data collection protocols that incorporate ftf interviewing have long been held as a gold standard, typically achieving higher response rates than other modes and reducing bias by eliciting response from subgroups that are generally missed by other modes (e.g., non-telephone households missed by telephone surveys or households without internet access missed by web surveys). However, ftf interviewing is expensive relative to other modes, and as a result has been cost-prohibitive for many studies that have turned, instead, to web, phone, or paper survey administrations, or combinations of these lower-cost modes.

We presented two sample designs and associated estimators that are alternatives to two-phase unit subsampling and may open the door to ftf data collection for some studies; for other studies, the approaches we described may facilitate more discriminating use of ftf interviewing as a way of constraining costs while maintaining high standards for quality. The PSU subsampling approach uses only a clustered sample, selecting initially a larger number of PSUs, but only uses ftf follow-up in a subsample of PSUs. The estimator $\hat{t}_{2}$ performed well for both subsampling designs, whereas $\hat{t}_{1}$ exhibited substantial bias and poor CI coverage. Clearly, trying to reduce the variance by using $\hat{t}_{1}$ is likely to result in biases in many situations.

The hybrid design approach blends two independent samples, an unclustered sample and a clustered sample, with the less costly initial mode(s) of data collection applied to both samples but the more costly ftf nonresponse follow-up used only in the clustered sample. As a result, the hybrid design approach offers the advantage of the same number of completes at a lower cost, while achieving the same overall weighted response rate and bias reduction as a design with complete nonresponse follow-up. With the hybrid design, the hybrid estimator $\hat{t}_{d f, 2}$ is unbiased under the nonresponse model $\bar{Y}_{N}=\bar{Y}_{F}$ or if the ftf achieves $100 \%$ response.

With the combination of parameters used in the simulation study, the composite estimator $\hat{t}_{d f, 2}$ in the hybrid design scenario had a RRMSE that was, on average, 8 percent lower than $\hat{t}_{2}$ in the PSU subsampling design scenario and 14 percent lower than $\hat{t}_{2}$ in the unit subsampling design scenario. All three scenarios
were designed to have about the same cost by having the same expected number of completes by mode and with ftf follow-up administered in the same number of PSUs.

The simulation findings suggest that the hybrid design and $\hat{t}_{d f, 2}$ is superior to either of the two-phase designs in the conditions we examined, and the two-phase PSU subsampling is superior to the two-phase unit subsampling. In addition, the hybrid design has practical benefits, especially when there is uncertainty regarding response rates by mode. With the hybrid design, increasing the sample size in the unclustered sample to deal with a lower-than-expected web response rate is relatively simple. With the two-phase designs, there are more complications such as the cost and schedule implications of the ftf follow-up due to increasing the first-phase sample.

The simulation consistently demonstrated the well-known shortcomings of the estimator $\hat{t}_{1}$ that is commonly used in practice, where there is no distinction in respondents by mode during the computation of weighting adjustments. Implicit in this estimator is the nonresponse model that assumes equality of means among the web respondents, the ftf respondents, and the nonrespondents. This is more restrictive than the model of equality of means between the ftf respondents and the nonrespondents that underlies the estimator $\hat{t}_{2}$ and the hybrid estimator $\hat{t}_{d f, 2}$. Brick et al. (2021) discuss this issue and partially account for imbalances in the respondent composition by using an adaptive mode adjustment. In a study involving web with ftf follow-up, the Brick et al. (2021) approach may be implemented by adjusting both the web and ftf respondents' weights to retain some of the bias reduction qualities of the adjustment of the ftf respondents while reducing variances by adjusting the weights of the web respondents as well. More research is needed to explore this approach with ftf surveys.

We have included discussion of theoretical properties of the estimators in the context of the sample designs presented here. However, as with any simulation, we have not examined every possible scenario. In designing our simulation, we focused on aspects we believed to be most likely to affect the relative performance of the estimators.

The estimators we presented incorporate a single adjustment, effectively treating the sample as a single weighting class. In practice, we would not expect the assumption $\bar{Y}_{N}=\bar{Y}_{F}$ to hold in general, but this assumption might be better approximated within classes, i.e., $\bar{Y}_{N_{C}}=\bar{Y}_{F_{C}}$, where the subscript $C$ denotes the class. If auxiliary variables can be identified and are available for both respondents and nonrespondents such that this nonresponse model holds (at least approximately), then these classes would typically be used in computing the weighting adjustments in order to reduce bias. Further research is needed to extend the estimators we presented to this situation, using weighting class-specific adjustments for nonresponse, for both the hybrid design and the PSU subsampling design. Similarly, calibrated estimators need to be evaluated.

We conclude with a few thoughts about another design we believe warrants further examination - an unclustered design in combination with a sample of existing PSUs - where the design that begins with web and uses ftf follow-up of the web nonrespondents in the sampled PSUs. One possible application of this design is for sample replenishment in a longitudinal study in which the original sample was a clustered sample of PSUs. Another application is where trained staff are available only in an existing sample of PSUs. In such contexts, one could consider selecting an unclustered sample, attempting the survey by web first,
and using ftf follow-up for the web nonrespondents in the unclustered sample that are located within the particular PSUs. We believe the extension of the estimator $\hat{t}_{2}$ shown in (3.5), with a modification to the second term so that the $\omega^{-1}$ is replaced by the reciprocal of the PSU probability of selection, is suitable in this context. Further work is needed to more fully examine this design and the properties of the estimator in this context.

## Acknowledgements

We are grateful to Sharon Lohr for her very helpful thoughts and comments on a draft of this manuscript. We also greatly appreciate the comments and suggestions offered by the Editor, Associate Editor, and three reviewers.

## Supplementary material

In the supplementary material (available at https://doi.org/10.48550/arXiv.2303.13303), we include additional results from the simulation. Specifically, this material presents CVs and CI coverage results for all comparisons, shows the results of the examination of alternative compositing factors, and gives results for the alternative subsampling adjustment in equation (3.4).

## Appendix

Table A. 1
Parameters for nonresponse follow-up designs in simulation.

|  | Total for unclustered sample | Total for clustered sample |  |
| :---: | :---: | :---: | :---: |
|  |  | Per PSU | Overall |
| No subsampling for NRFU (Scenario B1A) |  |  |  |
| Unclustered sample size | 2,500 |  |  |
| \# PSUs |  |  | 50 |
| \# units sampled per PSU |  |  | 50 |
| Expected \# completes |  |  |  |
| Web |  | 24.0 | 1,201.0 |
| Ftf (in all 50 PSUs) |  | 14.5 | 725.7 |
| Total | 1,201.0 | 38.5 | 1,926.7 |
| Two-phase unit subsampling (Scenario B2U) |  |  |  |
| Unclustered sample size | 0 |  |  |
| \# PSUs |  |  | 50 |
| \# units sampled per PSU |  |  | 100 |
| Nonresponse follow-up subsampling fraction |  |  | 0.5 |
| Expected \# completes |  |  |  |
| Web |  | 48.0 | 2,402.1 |
| Ftf (in all 50 PSUs) |  | 14.5 | 725.7 |
| Total | 0 | 62.6 | 3,127.7 |

[^6]Table A.1(continued)
Parameters for nonresponse follow-up designs in simulation.

|  | Total for unclustered sample | Total for clustered sample |  |
| :--- | ---: | ---: | ---: |
| Two-phase PSU subsampling (Scenario B2P) | Per PSU | Overall |  |
| Unclustered sample size | 0 |  |  |
| \# PSUs |  |  |  |
| total PSUs in sample |  |  |  |
| PSUs subsampled for NRFU |  |  |  |
| \# units sampled per PSU |  |  |  |
| Expected \# completes per PSU | 24.0 | 50 |  |
| Web |  | 14.5 | $2,402.1$ |
| Ftf (in only the 50 PSUs subsampled for NRFU) | 0 | 38.5 | 725.7 |
| Total in subsampled PSUs |  | $3,127.7$ |  |
| Expected total \# completes |  |  |  |

Notes: PSU = primary sampling unit; NRFU = nonresponse followup; Ftf = face-to-face.

Table A. 2
Mean of summary measures of estimators by scenario.

| Statistic | Scenario | 2-phase | Unclustered | 2-stage | df-1 | df-2 (opt) | df-2 (not opt) | 2-phase (PS) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| RB | A1A | -3.76\% | -5.36\% | -4.20\% | -4.66\% | -3.73\% | -3.76\% | -0.02\% |
|  | B1A | -0.02\% | -5.34\% | -1.78\% | -3.19\% | 0.01\% | -0.01\% |  |
|  | C1A | -0.03\% | -5.35\% | -0.03\% | -1.79\% | -0.01\% | -0.03\% |  |
|  | D1A | 0.00\% | -2.19\% | -0.49\% | -1.02\% | 0.01\% | 0.00\% |  |
|  | B2P | -0.01\% |  | -1.78\% |  |  |  |  |
|  | B2U | -0.07\% |  | -1.82\% |  |  |  |  |
| CV | A1A | 7.31\% | 4.28\% | 7.19\% | 4.74\% | 5.95\% | 6.69\% | 7.18\% |
|  | B1A | 8.16\% | 4.27\% | 7.70\% | 4.96\% | 6.82\% | 7.55\% |  |
|  | C1A | 7.79\% | 4.27\% | 7.79\% | 5.44\% | 6.39\% | 7.15\% |  |
|  | D1A | 8.51\% | 5.61\% | 8.23\% | 5.98\% | 7.76\% | 8.17\% |  |
|  | B2P | 7.37\% |  | 6.39\% |  |  |  |  |
|  | B2U | 8.15\% |  | 7.54\% |  |  |  |  |
| RRMSE | A1A | 12.78\% | 20.81\% | 13.25\% | 14.40\% | 11.69\% | 12.27\% | 7.97\% |
|  | B1A | 9.12\% | 20.77\% | 11.43\% | 13.58\% | 7.66\% | 8.45\% |  |
|  | C1A | 8.71\% | 20.75\% | 8.71\% | 9.53\% | 7.16\% | 8.00\% |  |
|  | D1A | 9.48\% | 22.47\% | 11.03\% | 12.80\% | 8.66\% | 9.11\% |  |
|  | B2P | 8.18\% |  | 10.27\% |  |  |  |  |
|  | B2U | 9.08\% |  | 11.33\% |  |  |  |  |
| CI | A1A | 72.58\% | 12.86\% | 75.06\% | 39.85\% | 65.24\% | 68.93\% | 95.31\% |
|  | B1A | 94.96\% | 12.89\% | 81.40\% | 39.46\% | 94.66\% | 94.60\% |  |
|  | C1A | 94.05\% | 12.95\% | 94.05\% | 63.57\% | 93.42\% | 93.43\% |  |
|  | D1A | 94.70\% | 18.48\% | 83.41\% | 49.66\% | 94.36\% | 94.30\% |  |
|  | B2P | 94.72\% |  | 75.19\% |  |  |  |  |
|  | B2U | 94.91\% |  | 79.97\% |  |  |  |  |
| ABS(RB) | A1A | 8.85\% | 20.12\% | 9.99\% | 13.16\% | 8.86\% | 8.85\% | 0.07\% |
|  | B1A | 0.09\% | 20.08\% | 6.57\% | 11.95\% | 0.07\% | 0.08\% |  |
|  | C1A | 0.05\% | 20.06\% | 0.05\% | 6.63\% | 0.05\% | 0.05\% |  |
|  | D1A | 0.06\% | 21.03\% | 4.75\% | 9.81\% | 0.08\% | 0.06\% |  |
|  | B2P | 0.06\% |  | 6.64\% |  |  |  |  |
|  | B2U | 0.09\% |  | 6.64\% |  |  |  |  |
| NormCIL | A1A | 1.16 | 0.71 | 1.16 | 0.76 | 0.93 | 1.04 | 1.07 |
|  | B1A | 1.20 | 0.71 | 1.18 | 0.76 | 0.99 | 1.08 |  |
|  | C1A | 1.08 | 0.71 | 1.08 | 0.75 | 0.85 | 0.96 |  |
|  | D1A | 1.23 | 0.93 | 1.18 | 0.85 | 1.12 | 1.16 |  |
|  | B2P | 1.07 |  | 0.95 |  |  |  |  |
|  | B2U | 1.19 |  | 1.13 |  |  |  |  |

[^7]
## References

Bayart, C., and Bonnel, P. (2015). How to combine survey media (web, telephone, face-to-face): Lyon and Rhône-Alps case study. Transportation Research Procedia, 11, 118-135.

Beaumont, J.-F., Béliveau, A. and Haziza, D. (2015). Clarifying some aspects of variance estimation in twophase sampling. Journal of Survey Statistics and Methodology, 3(4), 524-542.

Brick, J.M., Kennedy, C., Flores-Cervantes, I. and Mercer, A.W. (2021). An adaptive mode adjustment for multimode household surveys. Journal of Survey Statistics and Methodology, 10(4), 1024-1047.

Brick, J.M., and Montaquila, J.M. (2009). Nonresponse and weighting, Elsevier. Handbook of Statistics, 29, 163-185.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley \& Sons, Inc.

Dillman, D.A. (2017). The promise and challenge of pushing respondents to the Web in mixed-mode surveys. Survey Methodology, 43, 1, 3-30. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017001/article/14836-eng.pdf.

Goodman, A., Brown, M., Silverwood, R.J., Sakshaug, J.W., Calderwood, L., Williams, J. and Ploubidis, G.B. (2022). The impact of using the Web in a mixed-mode follow-up of a longitudinal birth cohort study: Evidence from the National Child Development Study. Journal of the Royal Statistical Society, Series A, 185(3), 822-850.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society, Series A, 169(3), 439-457.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. Journal of the American Statistical Association, 41(236), 517-529.

Heeringa, S.G., Wagner, J., Torres, M., Duan, N., Adams, T. and Berglund, P. (2004). Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). International Journal of Methods in Psychiatric Research, 13(4), 221-240.

Hidiroglou, M.A. (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.

Hidiroglou, M.A., Rao, J.N.K. and Haziza, D. (2009). Variance estimation in two-phase sampling. Australian \& New Zealand Journal of Statistics, 51(2), 127-141.

Holt, D.H. (1980). Discussion of paper by V. Verma, C. Scott and C. O'Muircheartaigh, Sample designs and sampling errors for the World Fertility Survey. Journal of the Royal Statistical Society, Series $A$, 143(4), 468-469.

Kish, L. (1992). Weighting for unequal Pi. Journal of Official Statistics, 8(2), 183.

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

Lu, W.W., Brick, J.M. and Sitter, R.R. (2006). Algorithms for constructing combined strata variance estimators. Journal of the American Statistical Association, 101(476), 1680-1692.

Neyman, J. (1938). Contribution to the theory of sampling human populations. Journal of the American Statistical Association, 33(201), 101-116.

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. International Statistical Review, 55, 3, 279-294.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer Series in Statistics.
U.S. Census Bureau (2014). American Community Survey design and methodology (January 2014). U.S. Census Bureau. (Available from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html).
U.S. Census Bureau (2019). American Community Survey: Accuracy of the data. U.S. Census Bureau. (Available from https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/ACS_Accuracy_ of_Data_2019.pdf).

Wagner, J., Arrieta, J., Guyer, H. and Ofstedal, M.B. (2014). Does sequence matter in multimode surveys: Results from an experiment. Field Methods, 26(2), 141-155.

# Dealing with undercoverage for non-probability survey samples 

Yilin Chen, Pengfei Li and Changbao Wu ${ }^{1}$


#### Abstract

Population undercoverage is one of the main hurdles faced by statistical analysis with non-probability survey samples. We discuss two typical scenarios of undercoverage, namely, stochastic undercoverage and deterministic undercoverage. We argue that existing estimation methods under the positivity assumption on the propensity scores (i.e., the participation probabilities) can be directly applied to handle the scenario of stochastic undercoverage. We explore strategies for mitigating biases in estimating the mean of the target population under deterministic undercoverage. In particular, we examine a split population approach based on a convex hull formulation, and construct estimators with reduced biases. A doubly robust estimator can be constructed if a followup subsample of the reference probability survey with measurements on the study variable becomes feasible. Performances of six competing estimators are investigated through a simulation study and issues which require further investigation are briefly discussed.


Key Words: Auxiliary information; Calibration method; Convex hull; Doubly robust estimator; Inverse probability weighting; Model-based prediction; Outcome regression; Propensity score; Split population.

## 1. Introduction

Probability survey samples and design-based inference have been widely used in official statistics and many other scientific fields as a standard tool for data collection and analysis. In recent years, however, "there has been a wind of change and other data sources are being increasingly explored" (Beaumont, 2020). One of the major reasons for looking at other data sources is the decreasing response rates for probability survey samples, and the seriousness of the undercoverage problem due to nonresponse as well as challenges in dealing with it for valid statistical inference.

Non-probability survey samples are one of the emerging data sources. Their ascent in popularity started with surveys based on web panels but the more broad definition extends to any volunteer-based and/or convenient samples or even administrative records. Statistical analysis of non-probability survey samples faces many hurdles, with the unknown sample selection and participation mechanism and the unknown coverage of the target population as the most pressing ones. Non-probability samples are biased and do not represent the target population in any tractable way, unlike probability survey samples where the survey design information is available. Valid statistical inferences with non-probability samples require additional auxiliary information at the population level and suitable inferential frameworks. A popular framework is to assume that the required population auxiliary information is available in an existing probability survey sample from the same target population. This two-sample framework has been used in several methodological developments, including the sample matching method (Rivers, 2007) and mass imputation (Kim, Park, Chen and $\mathrm{Wu}, 2021$ ), the weighted logistic regression for propensity score estimation using the

1. Yilin Chen, Pengfei Li and Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. Email: cbwu@uwaterloo.ca.
pooled sample (Valliant and Dever, 2011), the pseudo maximum likelihood method (Chen, Li and Wu , 2020), the pseudo empirical likelihood approach (Chen, Li, Rao and Wu, 2022), and the Bayesian approach (Wisniowski, Sakshaug, Ruiz and Blom, 2020), among others.

Statistical inferences with non-probability samples under the two-sample framework also require another critical assumption: the propensity score or the participation probability is positive for all the units in the target population. This is the so-called positivity assumption and is the foundation for the validity of several estimation methods proposed in the literature; see Section 2 for further discussion. With probability survey samples, this is equivalent to having a complete sampling frame without nonresponses. The positivity assumption is typically violated in practice for non-probability samples, due to limited geographic coverage of the population of interest and/or failing to reach subgroups of the population that are not as easily observable through convenient sampling methods. Violations of the positivity assumption lead to undercoverage problems and invalid results based on existing estimation methods. Undercoverage is a notoriously challenge problem in finite population sampling, and there is an added layer of complications with non-probability survey samples; see Elliott and Valliant (2017) for some extended discussion on the topic.

This paper discusses two typical scenarios of undercoverage in practice for non-probability survey samples: stochastic undercoverage and deterministic undercoverage. We argue in Section 3 that methods developed under the positivity assumption can be applied directly to handle stochastic undercoverage for valid inferences. Deterministic undercoverage involves a subpopulation for which certain crucial information is missing and no rigorous and valid estimation procedures can be developed under the existing twosample framework. In Section 4, we first discuss strategies for mitigating biases due to deterministic undercoverage using existing methods, and identify conditions under which existing methods lead to valid estimation results. We then explore estimation procedures under the split population through a convex hull formulation. We show that the correct specification of the outcome regression model is essential to several estimation procedures and a doubly robust estimator can be constructed if a followup subsample of the reference probability sample with measurements on the study variable can be obtained. Performances of six competing estimators of the finite population mean are evaluated through a simulation study and the results are reported in Section 5. Brief discussions on issues which require further investigation and some concluding remarks are given in Section 6.

## 2. Assumptions and existing approaches

There have been exciting methodological developments in recent years on valid statistical inference with non-probability survey samples. One of the key assumptions used by several authors is the non-zero probability of participation in the non-probability survey of all units in the target population. Let $\mathcal{U}=$ $\{1,2, \ldots, N\}$ be the set of $N$ labelled units for the target population. Let $y_{i}$ and $\mathbf{x}_{i}$ be the values of the study variable $y$ and the vector of auxiliary variables $\mathbf{x}$ for the $i^{\text {th }}$ unit in the population. Estimation
procedures are developed for a univariate $y$ with the focus on the population mean $\mu_{y}=N^{-1} \sum_{i=1}^{N} y_{i}$ but extensions can be made to other inferential problems similar to the theory of the Horvitz-Thompson estimator for design-based inference with probability survey samples.

Let $\mathcal{S}_{A}$ be the set of $n_{A}$ participating units in the non-probability survey sample and $\left\{\left(y_{i}, \mathbf{x}_{i}\right), i \in \mathcal{S}_{A}\right\}$ be the sample dataset. The most crucial feature of non-probability survey samples is the unknown sample inclusion or participation mechanism. The recent literature on the topic assumes that the mechanism is guided by an underlying stochastic process. Let $R_{i}=I\left(i \in \mathcal{S}_{A}\right)$ be the indicator variable for unit $i$ being included in the non-probability sample $\mathcal{S}_{A}$. Let

$$
\pi_{i}^{A}=P\left(i \in \mathcal{S}_{A} \mid \mathbf{x}_{i}, y_{i}\right)=P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}\right), \quad i=1,2, \ldots, N .
$$

The term "propensity scores" from the missing data literature (Rosenbaum and Rubin, 1983) was used for $\pi_{i}^{A}$ by Chen et al. (2020), among several other authors. Some authors preferred to use the term "participation probabilities" for $\pi_{i}^{A}$; see, for instance, Beaumont (2020) and Rao (2021), among others.

### 2.1 Assumptions

The following assumptions have been used in the recent literature on statistical inference with nonprobability survey samples; see, for instance, Wu (2022) and several key references therein.

A1 The sample inclusion and participation indicator $R_{i}$ and the study variable $y_{i}$ are independent given the set of auxiliary variables $\mathbf{x}_{i}$, i.e., $\left(R_{i} \perp y_{i}\right) \mid \mathbf{x}_{i}$.
A2 All the units in the target population have non-zero propensity scores, i.e., $\pi_{i}^{A}>0, i=1$, $2, \ldots, N$.

A3 The indicator variables $R_{1}, R_{2}, \ldots, R_{N}$ are independent given the set of auxiliary variables $\left(\mathbf{x}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{N}\right)$.

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

Assumption A1 is similar to the concept of missing-at-random (MAR) widely used for missing data analysis. Assumption A3 is more of a convenient tool for likelihood-based estimation of propensity scores, and it is not crucial to the validity of several existing estimating procedures ( $\mathrm{Wu}, 2022$ ). Assumption $\mathbf{A 4}$ is on the two-sample framework where auxiliary information on the target population is available from an existing probability survey sample. It is the basic setting for most estimation procedures proposed in the literature on non-probability survey samples.

Assumption $\mathbf{A 2}$ refers to the so-called positivity condition. For probability surveys, this is equivalent to conditions that the sampling frames are complete and there are no hardcore nonrespondents. In other words, the sampled population is identical to the target population, and statistical inferences based on the survey
sample are valid for the target population. In practice, assumption $\mathbf{A 2}$ is often violated for non-probability survey samples due to the voluntary and convenience nature of survey participation, resulting in undercoverage problems and invalid statistical statements on the target population.

### 2.2 Approaches to inference

There are three main approaches to inference using non-probability survey samples under assumptions A1-A4: (i) inverse probability weighting (IPW) based on an assumed model $q$ for the propensity scores; (ii) model-based prediction based on an assumed outcome regression model $\xi$; and (iii) doubly robust (DR) procedures using both the estimated propensity scores and the outcome regression model.

Under assumption A1, the propensity scores $\pi_{i}^{A}=P\left(R_{i}=1 \mid \mathbf{x}_{i}\right)=\pi\left(\mathbf{x}_{i}\right)$ are a function of the auxiliary variables $\mathbf{x}_{i}$ with an unknown form $\pi(\cdot)$. Let $\pi_{i}^{A}=\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)$ be a specified parametric form with unknown model parameters $\boldsymbol{\alpha}$. Under the two-sample setting where the population auxiliary information is supplied by the reference probability sample $\mathcal{S}_{B}$, the pseudo log-likelihood function for $\boldsymbol{\alpha}$ proposed by Chen et al. (2020) is given by

$$
\begin{equation*}
\ell^{*}(\boldsymbol{\alpha})=\sum_{i \in \mathcal{S}_{A}} \log \left(\frac{\pi_{i}^{A}}{1-\pi_{i}^{A}}\right)+\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \log \left(1-\pi_{i}^{A}\right) . \tag{2.1}
\end{equation*}
$$

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\alpha}}$ is the maximizer of $\ell^{*}(\boldsymbol{\alpha})$ and can be obtained as the solution to the pseudo score equations given by $\mathbf{U}(\boldsymbol{\alpha})=\partial \ell^{*}(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}=\mathbf{0}$. If the logistic regression model is assumed for the propensity scores where $\pi_{i}^{A}=1-\left\{1+\exp \left(\mathbf{x}_{i}^{\prime} \boldsymbol{\alpha}\right)\right\}^{-1}$, the pseudo score functions are given by

$$
\begin{equation*}
\mathbf{U}(\boldsymbol{\alpha})=\sum_{i \in \mathcal{S}_{A}} \mathbf{x}_{i}-\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right) \mathbf{x}_{i} . \tag{2.2}
\end{equation*}
$$

The estimated propensity scores are obtained as $\hat{\pi}_{i}^{A}=\pi\left(\mathbf{x}_{i}, \hat{\boldsymbol{\alpha}}\right), i \in \mathcal{S}_{A}$. The inverse probability weighted (IPW) estimator of $\mu_{y}$ is computed as

$$
\begin{equation*}
\hat{\mu}_{\mathrm{yPW}}=\frac{1}{\hat{N}^{A}} \sum_{i \in \mathcal{S}_{A}} \frac{y_{i}}{\hat{\pi}_{i}^{A}}, \tag{2.3}
\end{equation*}
$$

where $\hat{N}^{A}=\sum_{i \in \mathcal{S}_{A}}\left(\hat{\pi}_{i}^{A}\right)^{-1}$ is the estimated population size. The estimator $\hat{\mu}_{\text {yIPW }}$ is consistent under the joint randomization of the propensity score model $q$ and the probability sampling design $p$ for the reference probability sample $\mathcal{S}_{B}$.

The model-based prediction approach to inference also relies heavily on the first assumption. Under assumption A1, the conditional distribution of $y$ given $\mathbf{x}$ for units in the non-probability sample $\mathcal{S}_{A}$ (i.e., $R=1$ ) is the same as the conditional distribution of $y$ given $\mathbf{x}$ for units in the target population. It allows a valid model on $y$ given $\mathbf{x}$ to be built using the non-probability sample dataset $\left\{\left(y_{i}, \mathbf{x}_{i}\right), i \in \mathcal{S}_{A}\right\}$. Under the semiparametric outcome regression model $\xi$ as described in Wu (2022) with the first conditional moment specified as $E_{\xi}\left(y_{i} \mid \mathbf{x}_{i}\right)=m\left(\mathbf{x}_{i}, \boldsymbol{\beta}\right)$, the model parameters $\boldsymbol{\beta}$ can be consistently estimated by $\hat{\boldsymbol{\beta}}$
using the non-probability sample. Let $y_{i}^{*}=m\left(\mathbf{x}_{i}, \hat{\boldsymbol{\beta}}\right)$ be the fitted value of $y$ for unit $i \in \mathcal{S}_{A}$ or predicted value of $y$ for unit $i \notin \mathcal{S}_{A}$. A general form of the model-based prediction estimator of $\mu_{y}$ is computed as

$$
\begin{equation*}
\hat{\mu}_{y \mathrm{MI}}=\frac{1}{\hat{N}^{B}} \sum_{i \in \mathcal{S}_{B}} d_{i}^{B} y_{i}^{*}, \tag{2.4}
\end{equation*}
$$

where $\hat{N}^{B}=\sum_{i \in \mathcal{S}_{B}} d_{i}^{B}$. The subscript "MI" refers to "Mass Imputation", since the estimator is constructed based on the reference probability sample $\mathcal{S}_{B}$ with the unobserved $y$ treated as $100 \%$ missing for the sample and imputed for all the units in the sample. The estimator $\hat{\mu}_{y \mathrm{MI}}$ is consistent under the joint randomization of the outcome regression model $\xi$ and the probability sampling design $p$ for $\mathcal{S}_{B}$.

The doubly robust estimator of $\mu_{y}$ is computed by using the estimated propensity scores $\hat{\pi}_{i}^{A}=\pi\left(\mathbf{x}_{i}, \hat{\boldsymbol{\alpha}}\right)$ and fitted or predicted values $y_{i}^{*}=m\left(\mathbf{x}_{i}, \hat{\boldsymbol{\beta}}\right)$ and is given by (Chen et al., 2020)

$$
\begin{equation*}
\hat{\mu}_{y \mathrm{DR}}=\frac{1}{\hat{N}^{A}} \sum_{i \in \mathcal{S}_{A}} \frac{y_{i}-y_{i}^{*}}{\hat{\pi}_{i}^{A}}+\frac{1}{\hat{N}^{B}} \sum_{i \in \mathcal{S}_{B}} d_{i}^{B} y_{i}^{*} . \tag{2.5}
\end{equation*}
$$

The estimator $\hat{\mu}_{y \mathrm{DR}}$ is consistent under the probability sampling design $p$ for $\mathcal{S}_{B}$ and one of the correctly specified models, $q$ or $\xi$.

## 3. Two practical scenarios with undercoverage

It is known in design-based inference that there exists an unbiased estimator of $\mu_{y}$ in a subclass of the so-called Godambe class of linear estimators if and only if the first order inclusion probabilities are nonzero for all the units in the finite population (Wu and Thompson, 2020). For probability survey samples, zero-inclusion probabilities are the consequences of incomplete sampling frames and nonrespondents, leading to undercoverage problems for the target population.

The positivity assumption $\mathbf{A 2}$ which states that $\pi_{i}^{A}=P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}\right)>0$ for all $i$ in the target population is indeed the same condition for the validity of the IPW estimator, which is adapted from the HorvitzThompson estimator for probability survey samples, under the propensity score model $q$. The positivity assumption used in missing data analysis and causal inference is not an issue since the propensity scores are only defined for units in the sample. For non-probability survey samples, assumption $\mathbf{A 2}$ is often violated in practice for two major reasons: incomplete sampling frame(s) and voluntary participation. The sampling frame(s) used for selecting a non-probability survey sample is typically a convenient list such as a web panel, and it is almost surely incomplete for the target population. Participation in a non-probability survey sample is voluntary and nonresponse and refusals are an inherent part of the recruiting process.

Let $\mathcal{U}$ be the set of $N$ units for the target population. Let $\mathcal{U}_{0}=\left\{i \mid i \in \mathcal{U}\right.$ and $\left.\pi_{i}^{A}>0\right\}$. It is apparent that $\mathcal{U}_{0} \subset \mathcal{U}$ and $\mathcal{U}_{0} \neq \mathcal{U}$ when assumption $\mathbf{A 2}$ is violated. Let $\mathcal{U}_{1}=\left\{i \mid i \in \mathcal{U}\right.$ and $\left.\pi_{i}^{A}=0\right\}$. It follows that $\mathcal{U}=\mathcal{U}_{0} \cup \mathcal{U}_{1}$. Let $N=N_{0}+N_{1}$ where $N_{0}$ and $N_{1}$ are the sizes of the two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$. Let $\mu_{y 0}=N_{0}^{-1} \sum_{i \in \mathcal{U}_{0}} y_{i}$ and $\mu_{y 1}=N_{1}^{-1} \sum_{i \in U_{1}} y_{i}$. We have $\mu_{y}=W_{0} \mu_{y 0}+W_{1} \mu_{y 1}$, where $W_{k}=N_{k} / N$ for $k=0$, 1 . If
$\mathcal{S}_{A}$ is a sample from $\mathcal{U}_{0}$, and $\hat{\mu}_{y A}$ is an "unbiased estimator" based on $\mathcal{S}_{A}$, we usually have $E\left(\hat{\mu}_{y A}\right)=\mu_{y 0}$, and the bias of using $\hat{\mu}_{y A}$ to estimate $\mu_{y}$ is given by

$$
E\left(\hat{\mu}_{y A}\right)-\mu_{y}=W_{1}\left(\mu_{y 0}-\mu_{y 1}\right) .
$$

The two major factors for the amount of bias are (i) the size of the subpopulation (i.e., $N_{1}$ ) not represented by the sample $\mathcal{S}_{A}$; and (ii) the difference (i.e., $\mu_{y 0}-\mu_{y 1}$ ) between all potential participants of the sample and those who have no chances to be included in the sample. We discuss two practical scenarios for undercoverage problems and their implications on inference.

### 3.1 Stochastic undercoverage

The first scenario is termed as stochastic undercoverage, where the non-probability sample $\mathcal{S}_{A}$ is selected from a subpopulation $\mathcal{U}_{0}$ and the $\mathcal{U}_{0}$ itself can be viewed as a random sample from $\mathcal{U}$ (Chen, 2020; Wu, 2022). A typical example for this scenario is when the contact list of an existing large probability survey sample is used to recruit participants for the non-probability survey sample. Another example is when the participants for the non-probability sample are selected from a very large commercial panel, and the composition of the panel mimics the distributions of the target population in terms of key demographical variables. One can argue that it falls into the scenario of stochastic undercoverage. A less obvious example is a convenient sample of respondents recruited from shoppers at local shopping centers over certain period of time. If the target population consists of certain types of consumers in the region and there is a belief that such consumers have a non-trivial chance to visit one of the shopping centers during the time period, then it is also a case of stochastic undercoverage.

Let $D_{i}=1$ if $i \in \mathcal{U}_{0}$ and $D_{i}=0$ otherwise, $i=1,2, \ldots, N$. Noting that $R_{i}=I\left(i \in \mathcal{S}_{A}\right)$, we have

$$
P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}, D_{i}=1\right)>0 \text { and } P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}, D_{i}=0\right)=0
$$

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

$$
\pi_{i}^{A}=P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}\right)=P\left(R_{i}=1 \mid \mathbf{x}_{i}, y_{i}, D_{i}=1\right) P\left(D_{i}=1 \mid \mathbf{x}_{i}, y_{i}\right)>0
$$

for $i=1,2, \ldots, N$. The positivity assumption $\mathbf{A 2}$ is indeed satisfied under scenarios of stochastic undercoverage, and estimation procedures developed under the assumption can be used directly to provide valid inferences. A practical issue with stochastic undercoverage is how to specify a model for propensity scores due to the two-phase arguments involving $R_{i}$ and $D_{i}$. See Section 5 for further discussion.

### 3.2 Deterministic undercoverage

Many non-probability samples are volunteer-based convenience samples, and the potential participants often possess certain characteristics which are unique to the group. For instance, if participation in a survey
requires the use of a computer and access to the Internet, then those who do not have Internet access or never used a computer will have no chance to be included. The severity of undercoverage in this case depends largely on the proportion of the population being excluded.

The subpopulation $\mathcal{U}_{1}=\left\{i \mid i \in \mathcal{U}\right.$ and $\left.\pi_{i}^{A}=0\right\}$ may be conceptually defined through an accessibility function. Let $\Phi\left(\mathbf{x}_{i}\right)$ be a function of $\mathbf{x}_{i}$ that measures the accessibility of unit $i$ to the survey. An individual with a small value of $\Phi\left(\mathbf{x}_{i}\right)$ will have (practically) no access to the survey. More formally, we have $\pi_{i}^{A}=P\left(i \in \mathcal{S}_{A} \mid \mathbf{x}_{i}, y_{i}\right)=0$ if $\Phi\left(\mathbf{x}_{i}\right) \leq c$ for an unknown cut-off value $c$ on accessibility. The two subpopulations can alternatively be defined as

$$
\begin{equation*}
\mathcal{U}_{0}=\left\{i \mid i \in \mathcal{U} \text { and } \Phi\left(\mathbf{x}_{i}\right)>c\right\} \text { and } \mathcal{U}_{1}=\left\{i \mid i \in \mathcal{U} \text { and } \Phi\left(\mathbf{x}_{i}\right) \leq c\right\} . \tag{3.1}
\end{equation*}
$$

The truncation on $\Phi\left(\mathbf{x}_{i}\right)$ to exclude certain units from the non-probability survey can be viewed as a deterministic process, which motivates the use of the term "deterministic undercoverage". An overly simplified example is when $x_{i}$ represents the "age" of unit $i$ and all young individuals (i.e., $x_{i} \leq c$ for a chosen $c$ ) are excluded from the survey.

If the two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ can be clearly identified, valid statistical inferences can be claimed for the subpopulation $\mathcal{U}_{0}$. Extending the results to the target population $\mathcal{U}$ may be possible for certain scenarios but has the risk of overstretching with unrealistic assumptions.

## 4. Strategies for dealing with deterministic undercoverage

Deterministic undercoverage has similarities to frame and nonresponse undercoverage for probability survey samples. There are two major difficulties with inferences on the target population: the identification of the two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ and the lack of information on $\mathcal{U}_{1}$. In this section, we discuss approaches to mitigating biases of estimation due to the undercoverage and potential issues with these methods.

### 4.1 Calibrated IPW approach

Under the positivity assumption $\mathbf{A 2}$ and the specified parametric form $\pi_{i}^{A}=\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)$, the pseudo score functions given by $\mathbf{U}(\boldsymbol{\alpha})=\partial \ell^{*}(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$ from (2.1) can be replaced by a set of unbiased estimating equations (Chen et al., 2020; Wu, 2022)

$$
\begin{equation*}
\mathbf{G}(\boldsymbol{\alpha})=\sum_{i \in \mathcal{S}_{A}} \mathbf{h}\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)-\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right) \mathbf{h}\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right), \tag{4.1}
\end{equation*}
$$

where $\mathbf{h}\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)$ is a user-specified vector of functions with the same dimension of $\boldsymbol{\alpha}$. If we let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})=\mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha})$, the estimating functions given in (4.1) reduce to

$$
\begin{equation*}
\mathbf{G}(\boldsymbol{\alpha})=\sum_{i \in \mathcal{S}_{A}} \frac{\mathbf{x}_{i}}{\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)}-\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \mathbf{x}_{i} . \tag{4.2}
\end{equation*}
$$

Note that $\mathbf{G}(\boldsymbol{\alpha})=\mathbf{0}$ becomes the calibration equations $\sum_{i \in \mathcal{S}_{A}} \mathbf{x}_{i} / \pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)=\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \mathbf{x}_{i}$ with the population totals $\sum_{i=1}^{N} \mathbf{x}_{i}$ estimated by the probability sample $\mathcal{S}_{B}$. Let $\hat{\boldsymbol{\alpha}}_{C}$ be the solution to $\mathbf{G}(\boldsymbol{\alpha})=\mathbf{0}$, where the subscript " " indicates "Calibration". Let $\hat{\pi}_{i}^{C}=\pi\left(\mathbf{x}_{i}, \hat{\boldsymbol{\alpha}}_{C}\right)$. It is assumed that the first component of $\mathbf{x}$ is 1 so that

$$
\hat{N}_{C}^{A}=\sum_{i \in \mathcal{S}_{A}}\left(\hat{\pi}_{i}^{C}\right)^{-1}=\sum_{i \in \mathcal{S}_{B}} d_{i}^{B}=\hat{N}^{B} .
$$

The calibrated IPW estimator of $\mu_{y}$ is computed as $\hat{\mu}_{y \mathrm{IPW}}^{C}=\left(\hat{N}_{C}^{A}\right)^{-1} \sum_{i \in \mathcal{S}_{A}} y_{i} / \hat{\pi}_{i}^{C}$. The term "Calibrated IPW" was first used by Chen (2020). The idea was discussed by several other authors, including Rao (2021).

Under deterministic undercoverage, the parametric form with the restriction $\pi_{i}^{A}=\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)>0$ for all $i$ is clearly misspecified. As a consequence, the conventional IPW estimator $\hat{\mu}_{y \mathrm{yPW}}$ given by (2.3) is no longer consistent. The calibrated IPW estimator $\hat{\mu}_{y \mathrm{yPW}}^{C}$ can reduce the bias if the outcome regression model is linear, i.e., $E_{\xi}\left(y_{i} \mid \mathbf{x}_{i}\right)=\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}$. The calibrated IPW estimator under this scenario is an approximately model-unbiased prediction estimator (with the estimated population totals from the probability sample $\mathcal{S}_{B}$ ) since

$$
E_{p} E_{\xi}\left\{\frac{1}{\hat{N}_{C}^{A}} \sum_{i \in \mathcal{S}_{A}} \frac{y_{i}}{\hat{\pi}_{i}^{C}}\right\}=E_{p}\left\{\frac{1}{\hat{N}_{C}^{A}} \sum_{i \in \mathcal{S}_{A}} \frac{\mathbf{x}_{i}}{\hat{\pi}_{i}^{C}}\right\}^{\prime} \boldsymbol{\beta}=E_{p}\left\{\frac{1}{\hat{N}^{B}} \sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \mathbf{x}_{i}\right\}^{\prime} \boldsymbol{\beta} \doteq E_{\xi}\left(\mu_{y}\right) .
$$

The approximate equal sign in the last step amounts to estimating $N$ by $\hat{N}^{B}$.
A question of both practical and theoretical interest is whether the solution to $\mathbf{G}(\boldsymbol{\alpha})=\mathbf{0}$ exists, where $\mathbf{G}(\boldsymbol{\alpha})$ is given in (4.2). The answer depends on the chosen parametric form of $\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)$. Under a generalized linear model with $\pi_{i}=E\left(R_{i} \mid \mathbf{x}_{i}\right)=g\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\alpha}\right)$, where $g(\cdot)$ is the so-called (monotone increasing) inverse link function, we have

$$
\mathbf{H}(\boldsymbol{\alpha})=\frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{G}(\boldsymbol{\alpha})=-\sum_{i \in \mathcal{S}_{A}} \frac{k\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\alpha}\right)}{\left\{g\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\alpha}\right)\right\}^{2}} \mathbf{x}_{i} \mathbf{x}_{i}^{\prime}
$$

where $k(t)=d g(t) / d t>0$. The matrix $\mathbf{H}(\boldsymbol{\alpha})$ is negative definite, as long as the data matrix $\left\{\mathbf{x}_{i}, i \in \mathcal{S}_{A}\right\}$ is of full rank, and the usual Newton-Raphson iterative procedures for solving $\mathbf{G}(\boldsymbol{\alpha})=\mathbf{0}$ is guaranteed to converge.

The calibrated IPW estimator can also be constructed when a linear regression model is not appropriate but there are sufficient grounds to use a nonlinear model in the form of $E_{\xi}\left(y_{i} \mid \mathbf{x}_{i}\right)=m\left(\mathbf{x}_{i}, \boldsymbol{\beta}\right)$ with a known function $m(\cdot, \cdot)$. For instance, if $y_{i}$ is a binary variable, then $m\left(\mathbf{x}_{i}, \boldsymbol{\beta}\right)$ no longer has a linear form but may be chosen as the inverse logit function. Let $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$ obtained by using suitable estimation method and the non-probability sample dataset $\left\{\left(y_{i}, \mathbf{x}_{i}\right), i \in \mathcal{S}_{A}\right\}$. Let $\hat{m}_{i}=m\left(\mathbf{x}_{i}, \hat{\boldsymbol{\beta}}\right)$. The calibrated propensity scores are computed as $\hat{\pi}_{i}^{C}=\left(w_{i}^{C}\right)^{-1}$, where the calibrated weights $w_{i}^{C}$ are obtained in two steps:
(1) Compute the initial propensity scores $\hat{\pi}_{i}^{o}=\pi\left(\mathbf{x}_{i}, \hat{\boldsymbol{\alpha}}\right), i \in \mathcal{S}_{A}$, where $\hat{\boldsymbol{\alpha}}$ is the solution to the pseudo score equations from (2.2). Let $w_{i}^{o}=\left(\hat{\pi}_{i}^{o}\right)^{-1}, i \in \mathcal{S}_{A}$.
(2) Obtain the model-calibrated weights $w_{i}^{C}, i \in \mathcal{S}_{A}$ by minimizing the distance measure $D=$ $\sum_{i \in \mathcal{S}_{A}}\left\{w_{i}^{c}-w_{i}^{o}\right\}^{2} / w_{i}^{o}$ subject to constraints

$$
\begin{equation*}
\sum_{i \in \mathcal{S}_{A}} w_{i}^{C}=\hat{N}^{B} \text { and } \sum_{i \in \mathcal{S}_{A}} w_{i}^{C} \hat{m}_{i}=\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \hat{m}_{i} . \tag{4.3}
\end{equation*}
$$

The constraints used in (4.3) follow the general model-calibration procedures of Wu and Sitter (2001).
The calibrated IPW approach uses the outcome regression model to mitigate the potential bias of the IPW estimator with deterministic undercoverage. When a linear outcome regression model can be justified, the calibration step does not involve estimation of the model parameters $\boldsymbol{\beta}$ and hence leads to a more robust estimator than the model-based prediction estimator. The model-calibrated IPW estimator under a nonlinear outcome regression model requires the estimator $\hat{\boldsymbol{\beta}}$ which is obtained by fitting the model with the nonprobability sample. There is a risk of extrapolation in computing $\hat{m}_{i}=m\left(\mathbf{x}_{i}, \hat{\boldsymbol{\beta}}\right)$ for $i \in \mathcal{S}_{B}$. See Section 4.2 for further discussion.

### 4.2 Model-based prediction approach

A common scenario for deterministic undercoverage is that units lacking of certain features have no access to the non-probability survey, and the features are reflected in values of certain auxiliary variables. In practice, the first step for analyzing a non-probability survey dataset is to check the (unweighted) empirical marginal distributions of auxiliary variables which are potentially related to survey participation, and compare them to the weighted sample distributions of the variables using the reference probability sample. In particular, the observed range (or the support) of each auxiliary variable from the non-probability sample should be compared to those from the probability sample.

Model-based prediction approach through mass imputation relies on a conditional model of $y$ given $\mathbf{x}$. While the conditional moment structure $E_{\xi}\left(y_{i} \mid \mathbf{x}_{i}\right)=m\left(\mathbf{x}_{i}, \boldsymbol{\beta}\right)$, which is assumed for the target population, most likely holds for any samples, there are two problematic consequences with fitting the model using a sample which has a limited range in the observed auxiliary variables. The first is unreliable estimation of the model parameters with inflated variances for the estimators. The second is the danger of extrapolation in using the fitted model for prediction. These observations have been sufficiently documented in the existing literature on regression modelling and analysis. Tan (2007) expressed concerns on extrapolation in using a fitted outcome regression model with a biased sample in the construction of doubly robust estimators for missing data analysis and causal inference.

If the non-probability sample includes all the important auxiliary variables which are required for characterizing the participation behaviour and the outcome regression, and if the observed ranges of the auxiliary variables are similar to those from the reference probability sample, a model-based prediction estimator may be preferred over the IPW estimator in the presence of deterministic undercoverage. The calibrated IPW estimator discussed in Section 4.1 is especially attractive under a linear regression model
since estimation of the model parameters $\boldsymbol{\beta}$ is not needed and therefore the two issues with model-based prediction estimators, namely, the inflated variances for parameter estimates and the danger of extrapolation, become non-issues.

### 4.3 The split population approach

The conceptually defined two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ play a central role in the discussion of deterministic undercoverage. The non-probability sample $\mathcal{S}_{A}$ can be viewed as coming from $\mathcal{U}_{0}$ and satisfying the positivity assumption. It is tempting to develop tools to separate units in the reference probability sample $\mathcal{S}_{B}$ that belong to $\mathcal{U}_{0}$ or $\mathcal{U}_{1}$, and to further develop strategies for dealing with the split population.

The accessibility function $\Phi(\mathbf{x})$ introduced in Section 3.2 is a useful tool for the task. Suppose that $\Phi(\mathbf{x})$ is a convex function of $\mathbf{x}$ and $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ are defined in (3.1) with an unknown threshold $c$ that separates units from the two subpopulations. Let $\mathcal{H}_{k}$ be the convex hull generated by $\left\{\mathbf{x}_{i}: i \in \mathcal{U}_{k}\right\}$ for $k=0$, 1 . It follows that $\Phi(\mathbf{x})>c$ if $\mathbf{x} \in \mathcal{H}_{0}$ and $\Phi(\mathbf{x}) \leq c$ if $\mathbf{x} \in \mathcal{H}_{1}$. There are no overlaps between $\mathcal{H}_{0}$ and $\mathcal{H}_{1}$.

Let $\mathcal{H}_{A}$ be the convex hull formed by $\left\{\mathbf{x}_{i}: i \in \mathcal{S}_{A}\right\}$. We have $\mathcal{H}_{A} \subseteq \mathcal{H}_{0}$ and the difference between the two becomes negligible when $n_{A}$ is large. Similarly, the convex hull $\mathcal{H}_{B}$ formed by $\left\{\mathbf{x}_{i}: i \in \mathcal{S}_{B}\right\}$ approximates $\mathcal{H}_{0} \cup \mathcal{H}_{1}$ when $n_{B}$ is large since $\mathcal{S}_{B}$ represents the entire target population $\mathcal{U}$. The two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ can be identified through a split among units in the reference probability sample $\mathcal{S}_{B}=$ $\mathcal{S}_{B, 0} \cup \mathcal{S}_{B, 1}$, where

$$
\mathcal{S}_{B, 0}=\left\{j \mid j \in \mathcal{S}_{B} \text { and } \mathbf{x}_{j} \in \mathcal{H}_{A}\right\}
$$

and $\mathcal{S}_{B, 1}=\mathcal{S}_{B} \backslash \mathcal{S}_{B, 0}$. Note that verifying $\mathbf{x}_{j} \in \mathcal{H}_{A}$ is equivalent to checking if there exists a sequence of constants $a_{i} \geq 0$ for $i \in \mathcal{S}_{A}$ such that

$$
\sum_{i \in \mathcal{S}_{A}} a_{i}=1 \text { and } \sum_{i \in \mathcal{S}_{A}} a_{i} \mathbf{x}_{i}=\mathbf{x}_{j} .
$$

It can be done with existing computational packages. The sizes $N_{0}$ and $N_{1}$ of the two subpopulations $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ can be estimated by

$$
\hat{N}_{k}^{B}=\sum_{i \in \mathcal{S}_{B, k}} d_{i}^{B}, \quad k=0,1,
$$

which satisfy $\hat{N}_{0}^{B}+\hat{N}_{1}^{B}=\hat{N}^{B}$.
Kim and Rao (2018) described an idea on splitting the population using a modified nearest neighbour method. They defined $\mathcal{S}_{B, 0}$ as the set of units in $\mathcal{S}_{B}$ which have a "close neighbour" in $\mathcal{S}_{A}$. More formally, they define

$$
\mathcal{S}_{B, 0}=\left\{j \mid j \in \mathcal{S}_{B} \text { and } \min _{i \in \mathcal{S}_{A}}\left|\mathbf{x}_{i}-\mathbf{x}_{j}\right|<\epsilon\right\},
$$

where $\epsilon>0$ is a pre-specified tolerance measuring similarities in $\mathbf{x}$ among units. Choosing a value for $\epsilon$, however, is difficult in practice and the idea has not been developed further in the literature.

### 4.4 Estimation for the split population

Estimation for the split population involves separate treatments for $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$. Note that $\mu_{y}=W_{0} \mu_{y 0}+$ $W_{1} \mu_{y 1}$. Estimation procedures developed under the assumptions A1-A4 can be applied directly for the estimation of $\mu_{y 0}$ by treating $\mathcal{S}_{B, 0}$ as the reference probability sample. Let $\hat{\mu}_{y}=\hat{W}_{0} \hat{\mu}_{y 0}+\hat{W}_{1} \hat{\mu}_{y 1}$, where $\hat{W}_{k}=\hat{N}_{k}^{B} / \hat{N}^{B}$ for $k=0,1$. The severity of the deterministic undercoverage from using $\hat{\mu}_{y 0}$ as an estimator for $\mu_{y}$ is partially reflected by the value of $\hat{W}_{1}$. When $\hat{W}_{1}$ is small as compared to $\hat{W}_{0}$, we may ignore the issue with undercoverage and proceed with estimation under the assumption that $\pi_{i}^{A}>0$ for all $i$.

It is apparent that valid estimation of $\mu_{y 1}$ requires additional information on $y$ since the only relevant data in the two samples $\mathcal{S}_{A}$ and $\mathcal{S}_{B}$ on the subpopulation $\mathcal{U}_{1}$ are the auxiliary information $\left\{\mathbf{x}_{i}, i \in \mathcal{S}_{B, 1}\right\}$ from the split reference probability sample. In the absence of any additional information on $y$ for units in $\mathcal{U}_{1}$, we propose a hybrid estimator of $\mu_{y}$ as follows. We first estimate the propensity scores under the assumption that $\pi_{i}^{A}>0$ for $i \in \mathcal{U}_{0}$. Let $\hat{\pi}_{i 0}^{A}=\pi\left(\mathbf{x}_{i}, \hat{\boldsymbol{\alpha}}_{0}\right)$ under a parametric propensity score model $\pi_{i}^{A}=\pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right)$, where $\hat{\boldsymbol{\alpha}}_{0}$ is the pseudo maximum likelihood estimator of $\boldsymbol{\alpha}$. If a logistic regression model is used, then $\hat{\boldsymbol{\alpha}}_{0}$ is the solution to

$$
\sum_{i \in \mathcal{S}_{A}} \mathbf{x}_{i}-\sum_{i \in \mathcal{S}_{B, 0}} d_{i}^{B} \pi\left(\mathbf{x}_{i}, \boldsymbol{\alpha}\right) \mathbf{x}_{i}=\mathbf{0} .
$$

A calibration-based estimator of $\boldsymbol{\alpha}$ in the form of (4.2), with $\mathcal{S}_{B}$ replaced by $\mathcal{S}_{B, 0}$, can also be used. In this latter case we have $\hat{N}_{0}^{A}=\sum_{i \in \mathcal{S}_{A}}\left(\hat{\pi}_{i 0}^{A}\right)^{-1}=\hat{N}_{0}^{B}$ if $\mathbf{x}$ contains 1 as the first component. Let $\hat{m}_{i}=m\left(\mathbf{x}_{i}, \hat{\boldsymbol{\beta}}\right)$, where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$ obtained by using suitable estimation method and the non-probability sample dataset $\left\{\left(y_{i}, \mathbf{x}_{i}\right), i \in \mathcal{S}_{A}\right\}$ under the assumed outcome regression model $E_{\xi}\left(y_{i} \mid \mathbf{x}_{i}\right)=m\left(\mathbf{x}_{i}, \boldsymbol{\beta}\right)$. The doubly robust estimator of $\mu_{y 0}$ is computed as

$$
\begin{equation*}
\hat{\mu}_{y 0, \mathrm{DR}}=\frac{1}{\hat{N}_{0}^{B}} \sum_{i \in \mathcal{S}_{A}} \frac{y_{i}-\hat{m}_{i}}{\hat{\pi}_{i 0}^{A}}+\frac{1}{\hat{N}_{0}^{B}} \sum_{i \in \mathcal{S}_{B, 0}} d_{i}^{B} \hat{m}_{i} . \tag{4.4}
\end{equation*}
$$

Note that we used $\hat{N}_{0}^{B}$ instead of $\hat{N}_{0}^{A}$ in the first term. It leads to a simplified form of the hybrid estimator given below. Let

$$
\begin{equation*}
\hat{\mu}_{y 1, \text { REG }}=\frac{1}{\hat{N}_{1}^{B}} \sum_{i \in \mathcal{S}_{B, 1}} d_{i}^{B} \hat{m}_{i} \tag{4.5}
\end{equation*}
$$

be the model-based prediction estimator for $\mu_{y 1}$. A hybrid estimator of $\mu_{y}=W_{0} \mu_{y 0}+W_{1} \mu_{y 1}$ is constructed by using the two estimators given in (4.4) and (4.5):

$$
\begin{equation*}
\hat{\mu}_{y \mathrm{HYB}}=\hat{W}_{0} \hat{\mu}_{y 0, \mathrm{DR}}+\hat{W}_{1} \hat{\mu}_{y 1, \mathrm{REG}}=\frac{1}{\hat{N}^{B}} \sum_{i \in \mathcal{S}_{A}} \frac{y_{i}-\hat{m}_{i}}{\hat{\pi}_{i 0}^{A}}+\frac{1}{\hat{N}^{B}} \sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \hat{m}_{i} . \tag{4.6}
\end{equation*}
$$

The form of $\hat{\mu}_{y H Y B}$ is similar to the doubly robust estimator $\hat{\mu}_{y \mathrm{DR}}$ given in (2.5), with the major difference of estimating the propensity scores through the split population.

The hybrid estimator does not have the double robustness interpretation and has the risk of extrapolation in estimating $\mu_{y 1}$ using the model-based prediction estimator $\hat{\mu}_{y 1, \text { REG }}$. Suppose that the split of $\mathcal{S}_{B}$ into $\mathcal{S}_{B, 0}$ and $\mathcal{S}_{B, 1}$ using the convex hull formulation on $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ is done correctly, and the propensity score model on $\pi_{i}^{A}>0$ for $i \in \mathcal{U}_{0}$ is correctly specified, then the hybrid estimator $\hat{\mu}_{y \mathrm{HYB}}$ has the asymptotic expression

$$
\hat{\mu}_{y \mathrm{HYB}}=W_{0} \mu_{y 0}+W_{1} \bar{m}_{1}^{*}+o_{p}(1),
$$

where $\bar{m}_{1}^{*}=N_{1}^{-1} \sum_{i \in u_{1}} m\left(\mathbf{x}_{i}, \boldsymbol{\beta}^{*}\right)$ and $\boldsymbol{\beta}^{*}$ satisfies $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*}=O_{p}\left(n_{A}^{-1 / 2}\right)$, regardless of the correctness of the outcome regression model (Chen, 2020). The potential bias of the hybrid estimator depends largely on the model-based prediction estimator $\hat{\mu}_{y 1 \text {, REG }}$ for estimating $\mu_{y 1}$, and the estimator has the advantage of the doubly robust estimator $\hat{\mu}_{y 0, \mathrm{DR}}$ for estimating $\mu_{y 0}$.

Under ideal situations where it is possible to take a small subsample from $\mathcal{S}_{B, 1}$ and obtain measurements on $y$ for the selected units, a rigorous development on estimation methods for $\mu_{y 1}$, and consequently for $\mu_{y}$, can be carried out without much difficulties. Let $\left\{\left(y_{i}, d_{2 i}^{B}\right), i \in \mathcal{S}_{B, 1}^{(2)}\right\}$ be the additional dataset where $\mathcal{S}_{B, 1}^{(2)}$ is a subsample from $\mathcal{S}_{B, 1}$ and $d_{2 i}^{B}$ are the sampling weights for the subsample conditional on the given $\mathcal{S}_{B, 1}$. Let $\tilde{\boldsymbol{\beta}}$ be the estimated parameters for the outcome regression model using the combined dataset $\left\{\left(y_{i}, \mathbf{x}_{i}\right), i \in \mathcal{S}_{A} \cup \mathcal{S}_{B, 1}^{(2)}\right\}$. Let $\tilde{m}_{i}=m\left(\mathbf{x}_{i}, \tilde{\boldsymbol{\beta}}\right)$. A model-assisted difference estimator (Wu and Sitter, 2001) for the subpopulation mean $\mu_{y 1}$ can be constructed as

$$
\begin{equation*}
\hat{\mu}_{y l, \mathrm{SS}}=\frac{1}{\hat{N}_{1}^{B}} \sum_{i \in \mathcal{E}_{B, 1}^{S},} d_{2 i}^{B} d_{i}^{B}\left(y_{i}-\tilde{m}_{i}\right)+\frac{1}{\hat{N}_{1}^{B}} \sum_{i \in \hat{S}_{\beta, 1}} d_{i}^{B} \tilde{m}_{i}, \tag{4.7}
\end{equation*}
$$

where the subscript "SS" indicates "subsample". This estimator is approximately unbiased for $\mu_{y 1}$ under the survey designs for $\mathcal{S}_{B}$ and $\mathcal{S}_{B, 1}^{(2)}$ regardless the correctness of the outcome regression model. The final estimator of $\mu_{y}$ can then be computed as $\hat{\mu}_{y \mathrm{sS}}=\hat{W}_{0} \hat{\mu}_{y 0, \mathrm{DR}}+\hat{W}_{1} \hat{\mu}_{y 1, \mathrm{Ss}}$. The estimator $\hat{\mu}_{y \mathrm{sS}}$ is doubly robust and is given by

$$
\begin{equation*}
\hat{\mu}_{y \mathrm{SS}}=\frac{1}{\hat{N}^{B}}\left\{\sum_{i \in \mathcal{S}_{A}} \frac{y_{i}-\tilde{m}_{i}}{\hat{\pi}_{i 0}^{A}}+\sum_{i \in S_{B, 1}^{\left(d^{2}\right.}} d_{2 i}^{B} d_{i}^{B}\left(y_{i}-\tilde{m}_{i}\right)+\sum_{i \in \mathcal{S}_{B}} d_{i}^{B} \tilde{m}_{i}\right\} . \tag{4.8}
\end{equation*}
$$

## 5. Simulation studies

We evaluate the finite sample performances of several estimation strategies with deterministic undercoverage. Additional simulation results under stochastic undercoverage can be found in Chen (2020). We consider a finite population of size $N=20,000$, with three auxiliary variables $x_{1}, x_{2}$ and $x_{3}$. Independent copies of $\left(x_{i 1}, x_{i 2}, x_{i 3}\right)$ are generated from $x_{i 1} \sim N(0,1), x_{i 2} \sim \operatorname{Exp}(1)$, and $x_{i 3} \sim \operatorname{Bernoulli}(0.5)$. The response variable $y_{i}$ follows the regression model,

$$
\begin{equation*}
y_{i}=3+x_{i 1}+x_{i 2}+x_{i 3}-\eta x_{i 1}^{2}+\sigma \varepsilon_{i}, \quad i=1,2, \ldots, N, \tag{5.1}
\end{equation*}
$$

where $\eta$ is the coefficient for the high order term $x_{i 1}^{2}$ with values representing the degree of departure from the standard linear regression model. The error terms $\varepsilon_{i}$ are generated from $N(0,1)$ and the value of $\sigma$ is chosen to control the correlation coefficient $\rho$ between the response $y_{i}$ and the linear predictor $3+x_{i 1}+$ $x_{i 2}+x_{i 3}-\eta x_{i 1}^{2}$. The simulation results reported in this section correspond to $\rho=0.5$. The parameter of interest is the finite population mean $\mu_{y}$.

The accessibility function $\Phi(\mathbf{x})$ is specified through an inverse logit function involving the three auxiliary variables,

$$
\log \left\{\frac{\Phi\left(\mathbf{x}_{i}\right)}{1-\Phi\left(\mathbf{x}_{i}\right)}\right\}=1-0.6 x_{i 1}+0.5 x_{i 2}+0.8 x_{i 3}, \quad i=1,2, \ldots, N .
$$

Note that the inverse logit function is not a convex function. However, it can be shown that the function is convex within the subspace $\{\mathbf{x}: \Phi(\mathbf{x})<0.5\}$, which is sufficient for our proposed split population approach with the choices $\tau=0.00,0.20$ and 0.40 used in the simulation. Let $Q(\tau)$ be the $100 \tau^{\text {th }}$ sample quantile of $\left\{\Phi\left(\mathbf{x}_{1}\right), \Phi\left(\mathbf{x}_{2}\right), \ldots, \Phi\left(\mathbf{x}_{N}\right)\right\}$. The deterministic split of the population is decided by the measure on accessibilities. Let $\mathcal{U}=\{1,2, \ldots, N\}$ and

$$
\mathcal{U}_{0}=\left\{i \mid i \in \mathcal{U} \text { and } \Phi\left(\mathbf{x}_{i}\right) \geq Q(\tau)\right\}, \quad \mathcal{U}_{1}=\left\{i \mid i \in \mathcal{U} \quad \text { and } \Phi\left(\mathbf{x}_{i}\right)<Q(\tau)\right\} .
$$

We set $\pi_{i}^{A}=0$ if $i \in \mathcal{U}_{1}$, and the size of the subpopulation $\mathcal{U}_{1}$ is given by $N_{1}=\tau N$ (i.e., $W_{1}=\tau$ ). The size of the subpopulation $\mathcal{U}_{0}$ is given by $N_{0}=N-N_{1}$.

The true propensity scores $\pi_{i}^{A}$ for $i \in \mathcal{U}_{0}$ are generated from a logistic regression model,

$$
\log \left(\frac{\pi_{i}^{A}}{1-\pi_{i}^{A}}\right)=\theta+0.3 x_{i 1}-0.3 x_{i 2}+0.5 x_{i 3},
$$

where the intercept $\theta$ is chosen such that $\sum_{i=1}^{N_{0}} \pi_{i}^{A}=n_{A}$, where $n_{A}$ is the planned size of the non-probability survey sample $\mathcal{S}_{A}$. We use the Poisson sampling method with inclusion probabilities $\pi_{i}^{A}$ to select units for $\mathcal{S}_{A}$ from the subpopulation $\mathcal{U}_{0}$. The actual sample size of $\mathcal{S}_{A}$ varies from sample to sample with $n_{A}$ as the expected size.

The reference probability sample $\mathcal{S}_{B}$ with a fixed sample size $n_{B}$ is drawn from the entire finite population $\mathcal{U}$ by the randomized systematic PPS sampling method; see Section 4.4.2 of Wu and Thompson (2020) for further detail. The inclusion probabilities $\pi_{i}^{B}$ are proportional to $z_{i}=c+x_{i 2}$, where the constant $c$ is chosen to control the variation of the survey weights such that $\max _{i \in \mathcal{U}} z_{i} / \min _{i \in U} z_{i}=50$.

We consider six estimators of the population mean $\mu_{y}$ discussed in Sections 2 and 4, plus the naive estimator of the sample mean of the non-probability sample, and evaluate their performances through repeated simulation samples:
(1) $\hat{\mu}_{y A}$, the naive estimator of the sample mean of the non-probability sample $\mathcal{S}_{A}$;
(2) $\hat{\mu}_{y \text { IPW }}$, the IPW estimator given in (2.3);
(3) $\hat{\mu}_{y \mathrm{MI}}$, the model-based prediction (mass imputation) estimator given in (2.4);
(4) $\hat{\mu}_{y \mathrm{DR}}$, the doubly robust estimator given in (2.5);
(5) $\hat{\mu}_{\mathrm{yPW}}^{C}$, the calibrated IPW estimator described in Section 4.1 ;
(6) $\hat{\mu}_{y \mathrm{HYB}}$, the hybrid estimator given in (4.6) introduced in Section 4.4;
(7) $\hat{\mu}_{y S S}$, the estimator specified in (4.8) using a subsample $\mathcal{S}_{B, 1}^{(2)}$ from $\mathcal{S}_{B, 1}$.

For each iteration of the simulation with samples $\mathcal{S}_{A}$ and $\mathcal{S}_{B}$, the working propensity score model for the estimators $\hat{\mu}_{y \mathrm{PPW}}$ and $\hat{\mu}_{y \mathrm{DR}}$ is chosen as $\log \left\{\pi_{i}^{A} /\left(1-\pi_{i}^{A}\right)\right\}=\alpha_{0}+\alpha_{1} x_{i 1}+\alpha_{2} x_{i 2}+\alpha_{3} x_{i 3}$, the working outcome regression model for $\hat{\mu}_{y \mathrm{MI}}$ and $\hat{\mu}_{y \mathrm{DR}}$ as well as $\hat{\mu}_{y \mathrm{HYB}}$ uses $y_{i}=\beta_{0}+\beta_{1} x_{i 1}+\beta_{2} x_{i 2}+\beta_{3} x_{i 3}+\epsilon_{i}$. The splitting of $\mathcal{S}_{B}$ into $\mathcal{S}_{B, 0}$ and $\mathcal{S}_{B, 1}$ is done through the convex hull method. The rates of correct identification of units belonging to $\mathcal{U}_{0}$ and $\mathcal{U}_{1}$ are higher than $97 \%$ for all the settings used in the simulation. The subsample $\mathcal{S}_{B, 1}^{(2)}$ is selected from $\mathcal{S}_{B, 1}$ using simple random sampling without replacement and the sampling fraction is fixed at $20 \%$. Noting that for $n_{B}=500$ used in the simulation, the size of the subsample is around 20 for $\tau=W_{1}=0.2$ and 40 for $\tau=0.4$.

The amount of misspecification of the working regression model is reflected by the value of $\eta$ used in the true model (5.1) for generating the finite population $\left\{\left(y_{i}, x_{i 1}, x_{i 2}, x_{i 3}\right), i=1,2, \ldots, N\right\}$. The deterministic undercoverage is represented by the value of $\tau=W_{1}$. We consider $3 \times 3=9$ different settings for the simulation, with the true values of the population and the subpopulation means $\left(\mu_{y}, \mu_{y 0}, \mu_{y 1}\right)$ given in Table 5.1. The setting $(\eta=0.0, \tau=0.0)$ represents the ideal situation of no model misspecifications and no issues with undercoverage.

Table 5.1
Population and Subpopulation Means $\left(\mu_{y}, \mu_{y, 1}, \mu_{y, 0}\right)$.

|  | $\tau=\mathbf{0 . 0}$ | $\tau=\mathbf{0 . 2}$ | $\tau=\mathbf{0 . 4}$ |
| :---: | :---: | :---: | :---: |
| $\eta=0.0$ | $(4.53, \mathrm{NA}, \mathrm{NA})$ | $(4.53,4.49,4.72)$ | $(4.53,4.53,4.52)$ |
| $\eta=0.5$ | $(4.03, \mathrm{NA}, \mathrm{NA})$ | $(4.03,4.06,3.91)$ | $(4.03,4.06,3.97)$ |
| $\eta=1.0$ | $(3.52, \mathrm{NA}, \mathrm{NA})$ | $(3.52,3.63,3.11)$ | $(3.52,3.59,3.42)$ |

The performance of an estimator $\hat{\mu}_{y}$ is measured by the simulated Relative Bias ( $\mathrm{RB} \%$, in percentage) and the simulated Mean Squared Error (MSE), which are computed as

$$
\mathrm{RB} \%=100\left(\frac{1}{B} \sum_{b=1}^{B} \frac{\hat{\mu}_{y}^{(b)}-\mu_{y}}{\mu_{y}}\right), \mathrm{MSE}=\frac{1}{B} \sum_{b=1}^{B}\left(\hat{\mu}_{y}^{(b)}-\mu_{y}\right)^{2}
$$

where $\hat{\mu}_{y}^{(b)}$ is the estimator $\hat{\mu}_{y}$ computed from the $b^{\text {th }}$ simulation samples. Results for $n_{A}=1,000$ and $n_{B}=500$ based on $B=5,000$ simulation runs are presented in Table 5.2. The values of MSE are multiplied by 100 .

Table 5.2
Simulated RB\% and MSE ( $\times 10^{2}$ ) of Seven Estimators of $\mu_{y}$.

|  | Estimator | $\tau=0$ |  | $\tau=0.2$ |  | $\tau=0.4$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | RB\% | MSE | RB\% | MSE | RB\% | MSE |
| $\bar{\eta}=0.0$ | $\hat{\mu}_{y A}$ | 17.23 | 61.90 | 18.47 | 70.99 | 21.79 | 98.48 |
|  | $\hat{\mu}_{\text {ylPW }}$ | -0.09 | 1.67 | 0.00 | 1.39 | 0.66 | 1.64 |
|  | $\hat{\mu}_{\text {yMI }}$ | -0.03 | 1.55 | -0.04 | 1.45 | 0.36 | 1.70 |
|  | $\hat{\mu}_{y \mathrm{ld}}$ | -0.01 | 1.59 | 0.00 | 1.47 | 0.39 | 1.75 |
|  | $\hat{\mu}_{\text {ylPW }}^{C}$ | 0.00 | 1.58 | -0.01 | 1.49 | 0.49 | 2.37 |
|  | $\hat{\mu}_{\text {yHYB }}$ | 0.00 | 1.59 | 0.04 | 1.48 | 0.42 | 1.72 |
|  | $\hat{\mu}_{y s \mathrm{~S}}$ | NA | NA | -0.03 | 4.83 | 0.08 | 7.14 |
| $\bar{\eta}=0.5$ | $\hat{\mu}_{y A}$ | 16.91 | 47.40 | 23.56 | 91.03 | 27.82 | 126.59 |
|  | $\hat{\mu}_{\text {ylpw }}$ | -0.13 | 2.19 | 3.44 | 3.60 | 4.16 | 4.77 |
|  | $\hat{\mu}_{y \mathrm{MI}}$ | 2.82 | 2.88 | 3.71 | 3.81 | 4.83 | 5.75 |
|  | $\hat{\mu}_{y \mathrm{DR}}$ | -0.10 | 2.02 | 3.39 | 3.57 | 4.71 | 5.73 |
|  | $\hat{\mu}_{\text {ylPW }}^{C}$ | 0.01 | 1.91 | 3.85 | 4.10 | 5.79 | 8.11 |
|  | $\hat{\mu}_{\text {yHYB }}$ | -0.02 | 2.03 | 2.14 | 2.57 | 3.66 | 4.31 |
|  | $\hat{\mu}_{y s \mathrm{~S}}$ | NA | NA | -1.10 | 6.61 | -0.65 | 9.39 |
| $\bar{\eta}=1.0$ | $\hat{\mu}_{y A}$ | 16.53 | 35.44 | 30.17 | 114.40 | 35.68 | 159.43 |
|  | $\hat{\mu}_{\text {ylPW }}$ | -0.19 | 3.34 | 7.87 | 10.08 | 8.70 | 12.19 |
|  | $\hat{\mu}_{y \mathrm{MI}}$ | 6.48 | 7.58 | 8.53 | 11.14 | 10.63 | 16.81 |
|  | $\hat{\mu}_{y \mathrm{DR}}$ | -0.23 | 3.35 | 7.76 | 9.84 | 10.35 | 16.39 |
|  | $\hat{\mu}_{\text {ylPW }}^{C}$ | 0.03 | 2.83 | 8.82 | 11.99 | 12.69 | 23.80 |
|  | $\hat{\mu}_{\text {yHYB }}$ | -0.04 | 3.45 | 4.85 | 5.67 | 7.91 | 10.96 |
|  | $\hat{\mu}_{y s \mathrm{~S}}$ | NA | NA | -2.49 | 11.79 | -1.60 | 16.17 |

The simulation results can be summarized as follows. The naive estimator $\hat{\mu}_{y A}$ using the sample mean from the non-probability sample is biased under all the settings and is not included in any further comparisons with other six estimators. (1) All five estimators (the sixth estimator $\hat{\mu}_{\mathrm{ySS}}$ using a subsample is not applicable) under the setting of no model misspecification and no undercoverage (i.e., $\eta=0.0$ and $\tau=0.0$ ) perform well with no biases and similar MSEs; (2) Without issues of undercoverage (i.e., $\tau=0.0$ ), the model-based prediction estimator $\hat{\mu}_{y \mathrm{MI}}$ starts to show biases as the outcome regression model is misspecified (e.g., $\mathrm{RB} \%=6.48$ for $\eta=1.0$ ), while the other four estimators show no biases with similar small MSEs, including the hybrid estimator $\hat{\mu}_{y H Y B}$ under the split population approach; (3) The estimators $\hat{\mu}_{y \mathrm{MI}}, \hat{\mu}_{y \mathrm{DR}}, \hat{\mu}_{y \mathrm{HYB}}$ using the correctly specified outcome regression model (i.e., $\eta=0.0$ ) show no biases and similar small MSEs in the presence of undercoverage (i.e., $\tau=0.2$ or 0.4 ). The calibrated IPW estimator $\hat{\mu}_{\text {yIPW }}^{C}$, which requires a linear outcome regression model to justify, also shows no biases with undercoverage; (4) When the outcome regression model is misspecified (i.e., $\eta=0.5$ or 1.0 ) and there is an undercoverage problem (i.e., $\tau=0.2$ or 0.4 ), all five estimators (excluding the last one $\hat{\mu}_{y s \mathrm{~S}}$ ), which rely on the correctness of one of the two working models, demonstrate clear biases and deteriorated MSEs; (5)

The estimator $\hat{\mu}_{y \mathrm{SS}}$, which uses additional information on $y$ through a subsample from $\mathcal{S}_{B, 1}$, shows negligible biases for all scenarios but the values of MSE are larger than several other estimators. Part of the reasons is the very small size of the subsample $\mathcal{S}_{B, 1}^{(2)}$ used in the simulation since the estimator $\hat{\mu}_{y s \mathrm{~S}}$ involves the key component $\hat{\mu}_{\mathrm{y} 1, \mathrm{SS}}$ given in (4.7), and the latter has variance depending on the size of $\mathcal{S}_{B, 1}^{(2)}$.

## 6. Concluding remarks

The undercoverage problem with non-probability survey samples is closely attached to issues with modelling on propensity scores, and parametric models usually fail without the positivity assumption. In practice, most non-probability samples do not represent the entire target population, rendering propensity score based weighting methods invalid under such scenarios. Model-based prediction approach is sensitive to model specification, and the quality of the model on the response variable $y$ depends largely on the auxiliary variables which are available in both the non-probability sample and the reference probability sample. If the analyst is confident with the prediction model, the model-based prediction approach can be reliable in dealing with undercoverage problems when the support (i.e., the range) of each auxiliary variable in the non-probability sample matches the one from the probability sample. Otherwise there is a risk of extrapolation leading to biased estimation. From the theoretical view point, the deterministic undercoverage is a consequence of the violation of the positivity assumption $\mathbf{A 2}$. It leads to issues with fitting the outcome regression model using the non-probability sample data since $E(y \mid \mathbf{x}, R=1)=E(y \mid \mathbf{x})$ implicitly requires $P(R=1) \neq 0$ even if A1 holds. This is why the calibrated IPW estimator may have some advantages under a linear outcome regression model since the estimation of the model parameters $\boldsymbol{\beta}$ is not required.

The undercoverage problem is intrinsically related to the sample selection and participation mechanism for non-probability samples, which can be further complicated by the so-called non-ignorable selection bias. Dealing with non-ignorable selection bias for non-probability samples is itself an active research topic and has been investigated in several recent publications; see, for instance, Andridge, West, Little, Boonstra and Alvarado-Leiton (2019), Boonstra, Little, West, Andridge and Alvarado-Leiton (2021), and West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton (2021), among others. Sensitive analysis and quantitative measures on selection bias developed in these papers can be valuable tools for dealing with undercoverage problems.

The split population approach has been used in survey sampling to analyze and combine data from different sources; see, for instance, Zhang (2019) for further discuss and related references. Our proposed convex hull formulation in splitting the population into two subpopulations shows some potential in dealing with undercoverage problems, but a complete removal of biases in estimation after the split requires additional information on one of the subpopulations. The modified nearest neighbour method of Kim and Rao (2018) for splitting the target population seems to be a promising idea and may deserve some conscious efforts in future research.

Another important topic which is not addressed in the paper is on variance estimation under the strategies discussed in the paper. We are currently undertaking a separate research project on variance estimation and we hope to report our progresses in the near future.

The literature on missing data and causal inference includes methodological developments in dealing with the impact of very small but positive estimated propensity scores on the estimation of the main parameters through inverse probability weighting. Some of these developments may be useful for addressing undercoverage problems with non-probability samples, such as the stable weights approach of Zubizarreta (2015). It is hoped that discussions presented in this paper will add insights to the growing field of data integration and combining data from multiple sources.

## Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and a collaborative research team grant from the Canadian Statistical Sciences Institute. Part of the materials presented in this paper was taken from a chapter of the first author's doctoral dissertation completed at the University of Waterloo under the joint supervision of the other two co-authors.

## References

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

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

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

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

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

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

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

Kim, J.K., and Rao, J.N.K. (2018). Data integration for big data analysis in finite population inference. Paper presented at the 2018 Annual Meeting of the Statistical Society of Canada.

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

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

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

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

Tan, Z. (2007). Comment on "Understanding OR, PS and DR". Statistical Science, 22, 560-568.

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

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

Wisniowski, A., Sakshaug, J.W., Ruiz, D.A.P. and Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. Journal of Survey Statistics and Methodology, 8, 120-147.

Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, 48, 2, 283311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002eng.pdf.

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

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

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

Zubizarreta, J.R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110, 910-922.

# Bayesian small area models under inequality constraints with benchmarking and double shrinkage 

Balgobin Nandram, Nathan B. Cruze and Andreea L. Erciulescu ${ }^{1}$


#### Abstract

We present a novel methodology to benchmark county-level estimates of crop area totals to a preset state total subject to inequality constraints and random variances in the Fay-Herriot model. For planted area of the National Agricultural Statistics Service (NASS), an agency of the United States Department of Agriculture (USDA), it is necessary to incorporate the constraint that the estimated totals, derived from survey and other auxiliary data, are no smaller than administrative planted area totals prerecorded by other USDA agencies except NASS. These administrative totals are treated as fixed and known, and this additional coherence requirement adds to the complexity of benchmarking the county-level estimates. A fully Bayesian analysis of the Fay-Herriot model offers an appealing way to incorporate the inequality and benchmarking constraints, and to quantify the resulting uncertainties, but sampling from the posterior densities involves difficult integration, and reasonable approximations must be made. First, we describe a single-shrinkage model, shrinking the means while the variances are assumed known. Second, we extend this model to accommodate double shrinkage, borrowing strength across means and variances. This extended model has two sources of extra variation, but because we are shrinking both means and variances, it is expected that this second model should perform better in terms of goodness of fit (reliability) and possibly precision. The computations are challenging for both models, which are applied to simulated data sets with properties resembling the Illinois corn crop.


Key Words: Devroye method; Fay-Herriot model; Grid method; Hierarchical Bayesian model; Metropolis sampler.

## 1. Introduction

For many problems in official statistics, it is necessary to incorporate constraints in model-based inference. For example, in small area estimation, there may be constraints on the model estimates, which are to be benchmarked to a target. These may be known lower (or upper) bounds for county estimates, which should "add up" to the state estimate, obtained earlier. One practical example is the estimation of planted acres for counties within states, with a state estimate obtained earlier, when there are survey data and administrative data that can provide lower bounds to the county estimates, which are required to add up to the state estimate. While we focus on an application in agriculture, we develop a methodology to solve the problem in which small area estimates are needed to satisfy certain lower bounds and these estimates are further benchmarked to an estimate at a higher level via the top down approach.

In the United States, official county-level estimates of crop yield, total production, and total acreage published by National Agricultural Statistics Service (NASS), an agency of the United States Department of Agriculture (USDA) are important. These official estimates may determine the amount of payments to be made to farmers and ranchers enrolled in several programs administered by other USDA agencies including the Farm Service Agency (FSA) and the Risk Management Agency (RMA). Accordingly, NASS strives to improve the accuracy, reliability, and coverage of its official crop county estimates. As described

[^8]in a report titled Improving Crop County Estimates by Integrating Multiple Data Sources (National Academies of Sciences, Engineering, and Medicine, 2017), one way to do so is to use defensible models that include multiple sources of variability and other auxiliary data. The report highlighted many of the challenges faced by NASS and emphasized the role model-based inference can play in the publication of official county estimates. The findings of the report were further discussed in Cruze, Erciulescu, Nandram, Barboza and Young (2019), and the authors identified coherence of crop area estimates with known, sameyear administrative acreage totals as a significant need for the NASS crops county estimates program.

Constraints on estimates may enter in the form of order or shape restrictions (e.g., Nandram, Sedransk and Smith, 1997; Silvapulle and Sen, 2005; Chen and Nandram, 2022) or in the form of inequality constraints (Sen and Silvapulle, 2002). The latter type of restriction is of particular interest as it relates to the coherence of tabulated crop estimates in the presence of available administrative data curated by USDA. Benchmarking estimates for smaller geographic domains to those of larger geographic areas is one common form of equality constraint encountered in official statistics. For example, several past NASS studies have achieved this by ratio adjustment (raking) made after model output analysis (e.g., Erciulescu, Cruze and Nandram, 2018, 2019, 2020); see also Steorts, Schmid and Tzavidis (2020) and the references therein for an informative review on benchmarking. While the emphasis of the present work is methodological, we note the recent NASS-authored case study and companion paper (Chen, Nandram and Cruze, 2022) on the constrained planted area problem, single shrinkage model. Also, we note that in the current paper our main contributions are on the inequality constraints; see also NASS's RDD Research Report, Number RDD-2202 (Nandram, Cruze, Erciulescu and Chen, 2022).

Non-probability data are not devoid of errors. First, it is understood that while participation in agricultural support programs is popular in the United States, the voluntary enrollment in FSA and RMA programs contributes to potential under-coverage (a downward bias) in these administrative acreage totals. Moreover, rates of participation in these support programs may differ each year, by commodity crop, by state, or even more locally within state. Other nonsampling errors, however, are believed to be minimized through FSA and RMA quality controls. For example, farmers certify their enrolled acreages with FSA agents on geolocated field boundaries, and farmers are subject to penalties for falsifying their reports. With these properties in mind, the available administrative totals are viewed by NASS and USDA as informative lower bounds and publication of coherent tabular data on planted area requires: 1) that county acreage totals sum to the state acreage totals that are published prior to the release of county estimates, and 2) that official county-level planted area estimates honor the lower bound constraint in each county.

Additionally, we consider possible gains from double shrinkage by borrowing strength from means and variances simultaneously. Both frequentist and Bayesian model-based estimation techniques for the sampling variances have been considered in the literature for the area-level models. For example, see Wang and Fuller (2003); You and Chapman (2006); Gonzalez-Manteiga, Lombardia, Molina, Morales and SantaMaria (2010); Maiti, Ren and Sinha (2014); and Dass, Maiti, Ren and Sinha (2012). Recently, Erciulescu, Cruze and Nandram (2019) incorporated double shrinkage in estimates of unconstrained harvested area totals.

Let $\hat{\theta}_{i}, i=1, \ldots, \ell$, denote the observed direct estimates of total acreage for $\ell$ counties, and $s_{i}^{2}, i=$ $1, \ldots, \ell$, denote the corresponding observed variances for the $\ell$ counties. The Fay-Herriot (area-level) model (Fay and Herriot, 1979; see also, Rao and Molina, 2015) is a standard model in small area estimation for the $\hat{\theta}_{i}$, where,

$$
\begin{equation*}
\hat{\theta}_{i} \mid \theta_{i} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\theta_{i}, s_{i}^{2}\right), i=1, \ldots, \ell, \tag{1.1}
\end{equation*}
$$

and at the second stage,

$$
\begin{equation*}
\theta_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{\mathbf{i}}^{\prime} \boldsymbol{\beta}, \delta^{2}\right), i=1, \ldots, \ell \tag{1.2}
\end{equation*}
$$

where $\mathbf{x}_{i}$ is a $p$-vector of covariates with an intercept and $\boldsymbol{\beta}$ is a $p$-vector of regression coefficients. In a full Bayesian analysis of this model, prior distributions of model parameters are assumed; a priori we take $\pi\left(\boldsymbol{\beta}, \delta^{2}\right)=\pi(\boldsymbol{\beta}) \pi\left(\delta^{2}\right)$, where $\pi\left(\delta^{2}\right)$ is proper but $\pi(\boldsymbol{\beta})=1$ is improper.

Procedurally, NASS state estimates of planted area (denote these state targets by the scalar $a$ ) are determined and published prior to the publication of county-level estimates. Nandram, Erciulescu and Cruze (2019) developed a full Bayesian Fay-Herriot model incorporating the benchmarking constraint $\sum_{i=1}^{\ell} \theta_{i}=a$ directly into the model. This was achieved by deleting the last area to accommodate the benchmarking constraint. They empirically showed that, in practice, it does not really matter much which area is deleted in order to incorporate the benchmarking constraint. However, it is more convenient in the current paper to use an alternative approach, which does not use deletion.

We now want to refine this model to accommodate benchmarking and inequality constraints on the $\theta_{i}$. In addition to the benchmarking constraint, we need to add the county-specific inequality constraints

$$
\begin{equation*}
\theta_{i} \geq c_{i}, i=1, \ldots, \ell, \tag{1.3}
\end{equation*}
$$

where the $c_{i}$ are fixed, known quantities that represent administrative values provided by FSA or RMA. (In practice, when both data sources are present, the larger of the two is used to establish the lower bound, $c_{i}$.) In NASS planted acres data, some of the direct estimates of planted area totals may be more than one or two standard errors below their corresponding $c_{i}$, thereby creating some difficulties for the model estimates to be larger than the $c_{i}$. It is worth noting that $a=\sum_{i=1}^{\ell} \theta_{i} \geq \sum_{i=1}^{\ell} c_{i} \equiv c$. That is, the estimation processes that generate state targets also respect the available administrative totals at state level, however, the benchmarking constraint can create additional difficulties when the target is only slightly larger than $c$, i.e., as $\frac{c}{a} \rightarrow 1$ from below. We need to add the inequality constraints to the Fay-Herriot model specified in (1.1), (1.2) and the priors to get the joint posterior density of $\theta_{i}, i=1, \ldots, \ell$. In order to incorporate the inequality constraints into the Bayesian Fay-Herriot model, we propose the following simplification. In departure from Nandram, Erciulescu and Cruze (2019), we incorporate the inequality constraints directly while only partially incorporating the benchmarking constraint into the Bayesian Fay-Herriot model. That is, we will incorporate the constraints, $c_{i} \leq \theta_{i}, i=1, \ldots, \ell$, together with the restriction that $\sum_{i=1}^{\ell} \theta_{i}<a$ into the model.

When the latter inequality is enforced, a raking of model estimates to the state total $a$ in an output analysis will still satisfy all individual county inequality constraints. Incorporating double shrinkage into the inequality-constrained model entails additional computational considerations. Therefore, our key contributions are to provide small area estimates, which are subjected to inequality constraints, benchmarked to a target, and we describe a single shrinkage model (sample variances fixed) and two double shrinkage models (sample variances random).

In this paper, we discuss a novel methodology to solve these dual problems by modifying the Bayesian Fay-Herriot model described in Nandram, Erciulescu and Cruze (2019) to accommodate both benchmarking and inequality constraints into the Bayesian area-level models of Equations (1.1) and (1.2). Additionally, we extend the model to accommodate double shrinkage of means and variances. In Section 2, we introduce the methodology for single-shrinkage model in the presence of inequality constrained totals. In Section 3, we describe the methodology for the double-shrinkage model, gamma regression, and the log-linear model is discussed in Appendix B; again double-shrinkage models incorporate inequality constrained totals. Special emphasis is given to the computation that facilitates these approaches. In Section 4, as confidentiality of USDA survey and administrative data is a concern, simulated data sets with properties resembling those of the Illinois corn crop are generated and used to fit and assess these models. We offer concluding remarks in Section 5, noting that constrained acreage methodologies were successfully incorporated in NASS official statistics beginning with the 2020 crop year.

## 2. Methodology under the single shrinkage model

In this section, we develop the methodologies and computational strategies to incorporate inequality constraints and benchmarking procedures into the Bayesian area-level models of Equations (1.1) and (1.2). This provides the single shrinkage model in which the sampling variances are assumed fixed and known.

Our strategy is to use the composition rule (i.e., multiplication rule of probability) to draw samples from the posterior density $\pi\left(\boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^{2}\right)$ and then to draw samples from $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \delta^{2}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^{2}\right)$. Both of these problems are difficult. In this section, we have used the shrinkage prior for $\delta^{2}$ (i.e., $\pi\left(\delta^{2}\right)=1 /\left(1+\delta^{2}\right)^{2}$, $\delta^{2}>0$ ) to avoid impropriety of the posterior density. Letting $\phi=1 /\left(1+\delta^{2}\right)$, then $\phi \sim \operatorname{Beta}(1,1)$ (i.e., uniform). Note that if $x$ has a half Cauchy density, then the density of $\sqrt{\delta^{2}}$ after the transformation $x=\sqrt{\delta^{2}}$ is the Cauchy-based prior, $\pi\left(\delta^{2}\right)=\frac{1}{\pi \sqrt{\delta^{2}\left(1+\delta^{2}\right)}}$, which translates to $\phi \sim \operatorname{Beta}(0.5,0.5)$. In addition, both densities are in the Snedecor's $f$ distribution form, where the first density is a $f(2,2)$ and the Cauchy version is a $f(1,1)$; the $f(2,2)$ is mathematically a bit more convenient when we transform to $(0,1)$.

Let $V=\left\{\boldsymbol{\theta}: \theta_{i} \geq c_{i}, i=1, \ldots, \ell, \sum_{i=1}^{\ell} \theta_{i}<a\right\}$. Here, this conditional posterior density, $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \delta^{2}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^{2}\right)$, is subject to the inequality constraint and the constraint $\sum_{i=1}^{\ell} \theta_{i}<a$, where $a$ is the benchmarking target. Note that the inequality is strict because with the equality, one of the $\theta_{i}$ becomes redundant. This redundancy has to be taken into consideration when the model is fit (a much more difficult problem), but with the inequality constraint we do not need to do so (a much easier problem). That is, we need to draw $\theta_{1}, \ldots, \theta_{\ell}$ subject to the constraints $\theta_{i} \geq c_{i}, i=1, \ldots, \ell$ and $\sum_{i=1}^{\ell} \theta_{i}<a$. Note again that the benchmarking
constraint is only partially included in the Fay-Herriot model. We will use a Gibbs sampler to carry out this sampling procedure, and the benchmarking constraint will be fully incorporated in an output analysis from the Gibbs sampler using a raking procedure.

The joint prior density is

$$
\begin{equation*}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^{2}\right)=\pi\left(\boldsymbol{\beta}, \delta^{2}\right) \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{\mathbf{i}}^{\prime} \boldsymbol{\beta}\right)\right\} / \delta d \boldsymbol{\theta}}, \boldsymbol{\theta} \in V, \tag{2.1}
\end{equation*}
$$

where $\phi(\cdot)$ is the standard normal density. Indeed, this is a very awkward joint prior density with the normalization constant a function of ( $\boldsymbol{\beta}, \delta^{2}$ ). Then, using Bayes' theorem, the joint posterior density is

$$
\begin{equation*}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) \propto \pi\left(\boldsymbol{\beta}, \delta^{2}\right) \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}\left[\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\hat{\theta}_{i}\right) / s_{i}\right\}\right], \boldsymbol{\theta} \in V . \tag{2.2}
\end{equation*}
$$

It is difficult to use Markov chain Monte Carlo methods to efficiently draw samples from $\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$ in (2.2).

We now show how to draw samples from $\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$ using numerical integration, the Gibbs sampler and the Metropolis sampler. [Note that in the discussion below, apart from $\sum_{i=1}^{\ell} \theta_{i}<a$, it does not matter whether we use "less than or equal" symbols because the $\theta_{i}$ are continuous random variables.]

We first show how to draw the $\theta_{i}$ using the Gibbs sampler. For the constraints, we have $c_{i} \leq \theta_{i}, i=$ $1, \ldots, \ell$, and $\sum_{i=1}^{\ell} \theta_{i}<a$. This means that $\sum_{i=1}^{\ell} c_{i}<\sum_{i=1}^{\ell} \theta_{i}<a$, and so $\max \left(c_{i}, \sum_{j=1}^{\ell} c_{j}-\sum_{j=1, j \neq i}^{\ell} \theta_{j}\right)<\theta_{i}<$ $a-\sum_{j=1, j \neq i}^{\ell} \theta_{j}, i=1, \ldots, \ell$. Therefore, the support of the conditional posterior density of $\theta_{i}$ given $\boldsymbol{\theta}_{(i)}=$ $\left(\theta_{1}, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_{\ell}\right)^{\prime}$, is

$$
\max \left(c_{i}, \sum_{j=1}^{\ell} c_{j}-\sum_{j=1, j \neq i}^{\ell} \theta_{j}\right)<\theta_{i}<a-\sum_{j=1, j \neq i}^{\ell} \theta_{j}, i=1, \ldots, \ell .
$$

It is easy to show that the conditional posterior density is

$$
\begin{gather*}
\theta_{i} \mid \boldsymbol{\theta}_{(i)}, \boldsymbol{\beta}, \delta^{2}, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2} \sim \operatorname{Normal}\left\{\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta},\left(1-\lambda_{i}\right) \delta^{2}\right\}, \lambda_{i}=\delta^{2} /\left(\delta^{2}+s_{i}^{2}\right), \\
u_{i}=\max \left(c_{i}, \sum_{j=1}^{\ell} c_{j}-\sum_{j=1, j \neq i}^{\ell} \theta_{j}\right)<\theta_{i}<a-\sum_{j=1, j \neq i}^{\ell} \theta_{j}=v_{i}, i=1, \ldots, \ell . \tag{2.3}
\end{gather*}
$$

Now, we want to draw $\theta_{i}$ subject to the constraint, $u_{i} \leq \theta_{i} \leq v_{i}$. To sample $X \sim \operatorname{Normal}\left(\mu, \sigma^{2}\right)$, $a \leq X \leq b$, we have the following result (see Devroye, 1986),

$$
X=\mu+\sigma \Phi^{-1}\left\{(1-U) \Phi\left(\frac{a-\mu}{\sigma}\right)+U \Phi\left(\frac{b-\mu}{\sigma}\right)\right\}
$$

where $U \sim \operatorname{Uniform}(0,1)$ and $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are respectively the cdf and the inverse cdf of the standard normal density. We use the Gibbs sampler to draw a sample $\boldsymbol{\theta}$ in (2.3). This is obtained by drawing $u_{i} \leq \theta_{i} \leq v_{i}, i=1, \ldots, n$, each in turn.

The final step is to rake up $\theta_{1}, \ldots, \theta_{\ell}$ to the target $a$ for each iterate. So that the final iterates are

$$
\tilde{\theta}_{i}=\frac{a}{\sum_{j=1}^{\ell} \theta_{j}} \theta_{i}, i=1, \ldots, \ell
$$

and posterior inference can be made about $\theta_{1}, \ldots, \theta_{\ell}$ using these raked vectors of iterates. It is now clear why $\sum_{i=1}^{\ell} \theta_{i}<a$. Note again that this is a straight forward output analysis from the Gibbs sampler.

We next show how to draw samples from $\pi\left(\boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$ using numerical integration and the Metropolis sampler. The joint posterior density of ( $\boldsymbol{\beta}, \delta^{2}$ ) is

$$
\pi\left(\boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) \propto \pi\left(\boldsymbol{\beta}, \delta^{2}\right) \frac{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} \phi\left\{\left(\theta_{i}-\hat{\theta}_{i}\right) / s_{i}\right\} d \boldsymbol{\theta}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}
$$

which, by completing the squares, can be simplified to

$$
\begin{equation*}
\pi\left(\boldsymbol{\beta}, \delta^{2} \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) \propto \pi\left(\boldsymbol{\beta}, \delta^{2}\right)\left[\prod_{i=1}^{\ell} \phi\left\{\left(\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \sqrt{\delta^{2} / \lambda_{i}}\right\}\right] R\left(\boldsymbol{\beta}, \delta^{2}\right) \tag{2.4}
\end{equation*}
$$

with

$$
R\left(\boldsymbol{\beta}, \delta^{2}\right)=\frac{\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mu_{i}\right) / \tau_{i}\right\} d \boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}
$$

where $\mu_{i}=\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}$ and $\tau_{i}^{2}=\left(1-\lambda_{i}\right) \delta^{2}, i=1, \ldots, \ell$. We will use the Metropolis sampler to fit (2.4). There are two key issues, which are to construct an efficient proposal density and to compute the ratio, $R\left(\boldsymbol{\beta}, \delta^{2}\right)$, of the two integrals in (2.4).

First, we consider how to construct a proposal density. We have samples of ( $\boldsymbol{\beta}, \delta^{2}$ ) from the Fay-Herriot model. We can now transform $\delta^{2}$ to $\beta_{p+1}=\log \left(\delta^{2}\right)$ and add it as the last component to get a new vector $\boldsymbol{\beta}$ with $p+1$ components. Now fit a multivariate normal density to the samples, $\boldsymbol{\beta} \sim \operatorname{Normal}\left(\hat{\boldsymbol{\beta}}, \sigma^{2} \hat{\Sigma}\right)$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$ are the posterior mean and covariance matrix of the samples from the Fay-Herriot model, and $\eta / \sigma^{2} \sim \operatorname{Gamma}(\eta / 2,1 / 2)$ to complete the $(p+1)$-variate Student's $t$ density on $\eta$ degrees of freedom, where $\eta$ is a tuning constant.

Second, we describe how to estimate the ratio of the integrals in (2.4). Let $\tilde{V}=\left\{\boldsymbol{\theta}: c_{i}<\theta_{i}<\infty\right.$, $i=1, \ldots, \ell\}$; we have actually selected an upper bound for each $\theta_{i}$. Note that $V \subset \tilde{V}$, and perhaps $\tilde{V}$ is not much bigger than $V$. Let $I(\boldsymbol{\theta} \in V)=1$ if $\boldsymbol{\theta} \in V$ and $I(\boldsymbol{\theta} \in V)=0$ otherwise. Then,

$$
R\left(\boldsymbol{\beta}, \delta^{2}\right)=\frac{\int_{\boldsymbol{\theta} \in \tilde{V}} I(\boldsymbol{\theta} \in V) \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mu_{i}\right) / \tau_{i}\right\} d \boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \tilde{V}} I(\boldsymbol{\theta} \in V) \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}
$$

Now, $R\left(\boldsymbol{\beta}, \delta^{2}\right)$ can be calculated using Monte Carlo methods. As an importance function, we use the conditional posterior densities of the $\theta_{i}, i=1, \ldots, \ell$, constrained on $\tilde{V}$. That is,

$$
\begin{equation*}
\theta_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mu_{i}, \tau_{i}^{2}\right), c_{i}<\theta_{i}<\infty, i=1, \ldots, \ell . \tag{2.5}
\end{equation*}
$$

It is now easy to draw samples $\boldsymbol{\theta}^{(h)}, h=1, \ldots, M$, in (2.5), where $M \approx 1,000$ or so; see Devroye (1986). Then, a Monte Carlo estimator of $R\left(\boldsymbol{\beta}, \delta^{2}\right)$ is

$$
\widehat{R\left(\boldsymbol{\beta}, \delta^{2}\right)}=\frac{\sum_{h=1}^{M} I\left(\boldsymbol{\theta}^{(h)} \in V\right)}{\sum_{h=1}^{M} I\left(\boldsymbol{\theta}^{(h)} \in V\right)\left[\prod_{i=1}^{\ell} \frac{\phi\left\{\left(\theta_{i}^{(h)}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\phi\left\{\left(\theta_{i}^{(h)}-\mu_{i}\right) / \tau_{i}\right\}}\right]} .
$$

Note that for each $h$, once $\theta_{i}^{(h)}, i=1, \ldots, \ell$, are drawn from the proposal density, we simply need to check that $\sum_{i=1}^{\ell} \theta_{i}^{(h)}<a$. However, it is possible that this Monte Carlo estimator does not exist, and this clearly occurs when $\boldsymbol{\theta}^{(h)} \notin V, h=1, \ldots, M$ (all $M$ ), and in this case we use the modified estimator,

$$
\widehat{R_{m}\left(\boldsymbol{\beta}, \delta^{2}\right)}=\left[\frac{1}{M} \sum_{h=1}^{M} \prod_{i=1}^{\ell} \frac{\phi\left\{\left(\theta_{i}^{(h)}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\phi\left\{\left(\theta_{i}^{(h)}-\mu_{i}\right) / \tau_{i}\right\}}\right]^{-1}
$$

That is, we simply replace $V$ by $\tilde{V}$ to form an approximation in the case that the Monte Carlo estimator might not exist. In either case, we have drawn the $\theta_{i}$ as in (2.5), where $\theta_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mu_{i}, \tau_{i}^{2}\right)$, $c_{i}<\theta_{i}<\infty, i=1, \ldots, \ell$. It is possible for some of the $\boldsymbol{\theta}^{(h)}$ to be in $V$, and in this case if the number of $\boldsymbol{\theta}^{(h)} \in V$ is at least $M / 2$, we use the former estimator.

Our procedure gives us 1,000 samples from the posterior density of $\left(\boldsymbol{\beta}, \delta^{2}\right)$ using the Metropolis sampler. Then the more important samples of $\theta_{1}, \ldots, \theta_{\ell}$ are obtained using the Gibbs sampler. For each of the 1,000 iterates of $\left(\boldsymbol{\beta}, \delta^{2}\right)$ from the Metropolis sampler, we run the Gibbs sampler to say, 100 iterations or so, and pick the last set of $\theta_{1}, \ldots, \theta_{\ell}$. This is the so-called Gibbs-within-Metropolis sampler. This is not too expensive and it is reasonably efficient; we have seen similar difficulties in some of our projects (e.g., Nandram and Choi, 2010; Chen, Nandram and Cruze, 2022).

In this method, it is not really necessary to monitor the Gibbs sampler for convergence because we need only one value but a "burn-in" is required.

## 3. Methodology under the double shrinkage models

Two double shrinkage models are introduced, where we model both the sample variances and the means. The inequality constraints are also included. Here borrowing of strength occurs via both the means and the variances. For the specification of variances, the first uses a gamma regression model and the second uses a log-linear model. In Section 3, we model the sample variances using gamma regression; Section 3.1 describes the method and Section 3.2 describes the computation; further computations are shown in Appendix A. In Appendix B, we describe the second double shrinkage model for the sample variances using
the log-linear model. Even a full Bayesian treatment of the log-linear model offers remarkable computational advantages relative to the gamma regression model.

We discuss the reasons for two double shrinkage models. The computations in both models are difficult. We prefer the gamma model because it is more accurate within MCMC standards. Unfortunately, the computations are too time-consuming and it is not operational at NASS. We thought that if we move to a log-linear model, we can make some mathematical approximations, which will allow the double-shrinkage procedure to be operationalized at NASS and many other government agencies. Within the log-linear model, we made two approximations, which allow the computations to go very fast (in seconds) with reasonable accuracy. It is mathematically more difficult to make approximations within the gamma model, but some researchers might still prefer it.

### 3.1 Gamma regression model

For $\ell$ areas, we have the survey estimates $\hat{\theta}_{i}$, their standard errors $s_{i}$, and the sample sizes $n_{i} \geq 2$ (sample sizes must be at least 2). We start with a convenient model that builds upon our work on the FayHerriot model. We assume that

$$
\begin{gathered}
\hat{\theta}_{i} \mid \theta_{i}, \sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\theta_{i}, \sigma_{i}^{2}\right), i=1, \ldots, \ell \\
\frac{\left(n_{i}-1\right) s_{i}^{2}}{\sigma_{i}^{2}} \left\lvert\, \sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Gamma}\left(\frac{n_{i}-1}{2}, \frac{1}{2}\right)\right., i=1, \ldots, \ell
\end{gathered}
$$

where $X \sim \operatorname{Gamma}(a, b)$ means that $f(x)=b^{a} x^{a-1} e^{-b x} / \Gamma(a), x \geq 0$. Note that, given $\theta_{i}$ and $\sigma_{i}^{2}$, we are assuming $\hat{\theta}_{i}$ and $s_{i}^{2}$ are independent. Under the first assumption, the $\theta_{i}$ and $\sigma_{i}^{2}$ are not estimable, but the first and second assumptions together make $\theta_{i}$ and $\sigma_{i}^{2}$ estimable. Here, the chi-square assumption is reasonable, but the degrees of freedom may be a little smaller than the original sample size because it should be the effective sample size. The effective sample size is not normally presented at NASS, and in fact, it is the number of reports with positive responses that is presented. So we have used the original sample size; see Erciulescu, Cruze and Nandram (2019) for a similar model without the inequality constraint, of course.

A priori, we assume that

$$
\begin{aligned}
& \theta_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}, \delta^{2}\right), i=1, \ldots, \ell \\
& \sigma_{i}^{-2} \mid \alpha, \gamma \sim \operatorname{ind} \\
& \sim \operatorname{Gamma}\left(\frac{\alpha}{2}, \frac{\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma}}{2}\right), i=1, \ldots, \ell
\end{aligned}
$$

These assumptions on $\theta_{i}$ and $\sigma_{i}^{2}$ provide double shrinkage (shrinking both means and variances). Here, we have assumed that the two sets of covariates are the same, but they can, of course, be different.

It is worth noting that the prior for $\sigma_{i}^{2}$ is conjugate providing some simplicity in the computations; see Nandram and Erhardt (2004) for similar specifications for the corresponding binomial and Poisson models. Our prior for the hyperparameters is

$$
\pi\left(\boldsymbol{\beta}, \delta^{2}, \gamma, \alpha\right) \propto \frac{1}{\left(1+\delta^{2}\right)^{2}} \frac{1}{(1+\alpha)^{2}}, \delta^{2}, \alpha \geq 0 .
$$

That is, flat priors are assumed for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, shrinkage priors (proper) are assumed on $\delta^{2}$ and $\alpha$, and all parameters are independent. Note that $\delta^{2}$ and $\alpha$ are nonnegative, and so we prefer to use a shrinkage prior. At this point, there are virtually no mathematical, computational or scientific benefits using other noninformative priors for $\alpha$.

In our model, we include the inequality constraint, $\theta_{i}>c_{i}, i=1, \ldots, \ell, \sum_{i=1}^{\ell} \theta_{i}<a$, where $a$ is the target. Note again that we only partially include the benchmarking constraint. It is convenient that this is the same region as for single shrinkage model, $V=\left\{\boldsymbol{\theta}: \theta_{i} \geq c_{i}, i=1, \ldots, \ell, \sum_{i=1}^{\ell} \theta_{i}<a\right\}$. Therefore, the prior densities for the $\theta_{i}$ remain the same,

$$
\pi\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \delta^{2}\right)=\pi\left(\boldsymbol{\beta}, \delta^{2}\right) \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}, \boldsymbol{\theta} \in V,
$$

where $\phi(\cdot)$ is the standard normal density. It is convenient to define $\Omega=\left(\boldsymbol{\beta}, \delta^{2}, \gamma, \alpha\right)$. Then, the joint prior density is

$$
\begin{align*}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \Omega\right)= & \pi(\Omega) \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\}}{\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}}  \tag{3.1}\\
& \times \prod_{i=1}^{\ell}\left\{\left(\alpha e^{\left.\left.-\mathbf{x}_{i}^{\prime} / 2\right)^{\alpha / 2}\left(1 / \sigma_{i}^{2}\right)^{\alpha / 2+1} e^{-\left(\alpha e^{\left.-x_{i} / / 2 \sigma_{i}^{2}\right)}\right.} / \Gamma(\alpha / 2)\right\}, \boldsymbol{\theta} \in V .}\right.\right.
\end{align*}
$$

By independence, the joint density of $\left(\hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$, is

$$
\begin{align*}
& f\left(\hat{\boldsymbol{\theta}}, \mathbf{S}^{2} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \Omega\right)= \\
& \quad \prod_{i=1}^{\ell}\left\{\frac{1}{\sigma_{i}} \phi\left\{\frac{\hat{\theta}_{i}-\theta_{i}}{\sigma_{i}}\right\} \times \prod_{i=1}^{\ell}\left\{\left\{\left[\left(n_{i}-1\right) / 2 \sigma_{i}^{2}\right]^{\left(n_{i}-1\right) / 2}\left(s_{i}^{2}\right)^{\left(n_{i}-1\right) / 2-1} e^{-\left(n_{i}-1\right) s_{i}^{2} / 2 \sigma_{i}^{2}}\right\} / \Gamma\left\{\left(n_{i}-1\right) / 2\right\}\right\}\right\} . \tag{3.2}
\end{align*}
$$

Finally, using Bayes' theorem, the joint posterior density is proportional to the product of (3.1) and (3.2) and it can be shown to be

$$
\begin{align*}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \Omega \mid \hat{\boldsymbol{\theta}}, \mathbf{S}^{2}\right) & \propto \pi\left(\boldsymbol{\beta}, \delta^{2}, \boldsymbol{\gamma}, \alpha\right) \frac{1}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{\mathbf{i}}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}} \\
& \times \prod_{i=1}^{\ell}\left\{\left(\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma} / 2\right)^{\alpha / 2}\left(1 / \sigma_{i}^{2}\right)^{\alpha / 2+1} e^{-\left(\alpha e^{-x i t / 2} / 2 \sigma_{i}^{2}\right)} / \Gamma(\alpha / 2)\right\} \\
& \times \prod_{i=1}^{\ell}\left\{\frac{1}{\left.\sqrt{\left(1-\lambda_{i}\right) \delta^{2}} \phi\left(\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right) \frac{1}{\sqrt{\delta^{2} / \lambda_{i}}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\}}\right.  \tag{3.3}\\
& \times \prod_{i=1}^{\ell}\left\{\left[\left(n_{i}-1\right) / 2 \sigma_{i}^{2}\right]^{\left(n_{i}-1\right) / 2} e^{-\left(n_{i}-1\right) s_{i}^{2} / 2 \sigma_{i}^{2}}\right\}, \boldsymbol{\theta} \in V,
\end{align*}
$$

where $\lambda_{i}=\delta^{2} /\left(\delta^{2}+\sigma_{i}^{2}\right), i=1, \ldots, \ell$.
It now follows from (3.3) that the conditional posterior densities of the $\theta_{i}$ are

$$
\begin{equation*}
\theta_{i} \mid \boldsymbol{\sigma}^{2}, \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left\{\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta},\left(1-\lambda_{i}\right) \delta^{2}\right\}, i=1, \ldots, \ell, \boldsymbol{\theta} \in V \tag{3.4}
\end{equation*}
$$

Now, one can integrate out the $\theta_{i}$ from (3.3) to get the joint conditional posterior density of $\boldsymbol{\sigma}^{2}$,

$$
\begin{align*}
\pi\left(\boldsymbol{\sigma}^{2} \mid \Omega, \hat{\boldsymbol{\theta}}, \mathbf{S}^{2}\right) & \propto \int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta} \\
& \times \prod_{i=1}^{\ell}\left\{\sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\} \prod_{i=1}^{\ell}\left\{\left(1 / \sigma_{i}^{2}\right)^{\left(n_{i}+\alpha-1\right) / 2+1} e^{-\left\{\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-x_{i}^{\prime}}\right\} / 2 \sigma_{i}^{2}}\right\} . \tag{3.5}
\end{align*}
$$

Note that the term, $\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}$, is not a function of the $\sigma_{i}^{2}$ and has been eliminated together with other such terms.

Now, one can integrate out the $\sigma_{i}^{2}$ from (3.3) to get the joint posterior density of $\Omega$,

$$
\begin{align*}
\pi\left(\Omega \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) & \propto \pi\left(\boldsymbol{\beta}, \delta^{2}, \boldsymbol{\gamma}, \alpha\right) \prod_{i=1}^{\ell}\left\{\frac{\Gamma(\alpha / 2)}{\left(\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma} / 2\right)^{\alpha / 2}} \frac{\Gamma\left(n_{i}+\alpha-1\right) / 2}{\left\{\left(\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma}\right) / 2\right\}^{\left(n_{i}+\alpha-1\right) / 2}}\right\} \\
& \times \frac{1}{\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right) / \delta\right\} d \boldsymbol{\theta}} \int_{\sigma^{2}}\left[\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta}\right.  \tag{3.6}\\
& \left.\times \prod_{i=1}^{\ell}\left\{\frac{1}{\sqrt{\delta^{2} / \lambda_{i}}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right) \mathrm{IG}_{\sigma_{i}^{2}}\left(a_{i}, b_{i}\right)\right\}\right] d \boldsymbol{\sigma}^{2},
\end{align*}
$$

where $a_{i}=\left(n_{i}+\alpha-1\right) / 2$ and $b_{i}=\left\{\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-\mathrm{x}_{i}^{\prime} \gamma}\right\} / 2$ and $\mathrm{IG}_{c}(a, b)$ is the inverse gamma density, which is given by $f(c)=b^{a}\left(\frac{1}{c}\right)^{a+1} e^{-b / c} / \Gamma(a), c>0$.

### 3.2 Computation for the gamma regression model

Our strategy is to draw samples from the joint posterior density of $\Omega$ in (3.6). This is a difficult task, but once this is accomplished, we can use the multiplication rule to draw samples of the $\sigma_{i}^{2}$ from (3.5) and then the $\theta_{i}$ from (3.4). This strategy is useful if there are a large number of counties; the state of Texas has 254 counties. We draw the $\theta_{i}$ in the same manner as described in Section 2. It is more difficult to draw samples of $\sigma_{i}^{2}$. We describe how to draw samples from $\Omega$ in (3.6). The basic strategy has two key steps.

First, we fit the double shrinkage model without the inequality constraints and the benchmarking. This gives an approximate sample of size $M=1,000$ iterates from the posterior density of $\Omega$ that we obtained using a Metropolis sampler. The details of this first step are given in Appendix A.

Second, we convert this approximate sample to a sample from the posterior density with the inequality constraint and the benchmarking. We use the $M$ iterates from the first step to construct a multivariate Student's $t$ density for $\left(\boldsymbol{\beta}, \log \left(\delta^{2}\right), \gamma, \log (\alpha)\right)$. At each of the iterate obtained from the first step, we run a Metropolis sampler with the multivariate Student's $t$ density 100 times and picked the last one; see Nandram and Choi (2010) for a similar procedure. In this divide-and-conquer manner, we minimize the
chance of the Metropolis sampler getting stuck. We want the Metropolis sampler to move from the starting value at least once; no other monitoring is necessary; if it does not move at least once, we discard this run. It is good that this procedure gives a sample of $M$ independent iterates of $\Omega$. However, this step in timeconsuming and for the current simulated data it took roughly sixteen hours.

Now, we describe how to use the accept-reject algorithm to draw samples of $\sigma_{i}^{2}$. We can rewrite (3.5) as

$$
\begin{align*}
& \pi\left(\boldsymbol{\sigma}^{2} \mid \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) \propto \int_{\theta \in \tilde{V}} \prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta} \\
& \quad \times \prod_{i=1}^{\ell}\left\{\sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\} \int_{\theta \in \tilde{V}} I(\boldsymbol{\theta} \in V) \frac{\prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}{\int_{\theta \in \tilde{V}} \prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta}} d \boldsymbol{\theta}  \tag{3.7}\\
& \quad \times \prod_{i=1}^{\ell}\left\{\left(1 / \sigma_{i}^{2}\right)^{\left(n_{i}+\alpha-1\right) / 2+1} e^{-\left\{\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-x_{i}}\right\} / 2 \sigma_{i}^{2}}\right\},
\end{align*}
$$

where $\tilde{V} \supseteq V$ and $\tilde{V}$ is a larger rectangular set.
Note that the first and third terms in (3.7) are probabilities. It is also true that the second term in (3.7) is a probability because

$$
\prod_{i=1}^{\ell}\left\{\sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\} \leq\left\{\frac{1}{\sqrt{2 \pi}}\right\}^{\ell} .
$$

Therefore, we can use an accept-reject sampler to draw the $\sigma_{i}^{2}$.
Note that, by construction, the first term in (3.7) is a product over $i=1, \ldots, \ell$. This is also true for the second term. So if we ignore the third term, we can independently draw $\sigma_{i}^{2} \sim \operatorname{ind} \operatorname{IG}\left(a_{i}, b_{i}\right), i=1, \ldots, \ell$ (unrestricted distributions) and take it with probability,

$$
\int_{\theta \in \tilde{V}} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta} \times\left\{\sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\}
$$

to complete the accept-reject algorithm. It is possible that there are several rejections before an acceptance, but this rarely happens. If there are 25 rejections, we simply draw the $\sigma_{i}^{2}$ from their unrestricted distributions, $\sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{IG}\left(a_{i}, b_{i}\right), i=1, \ldots, \ell$.

The remaining question then is how to calculate

$$
C=\int_{\boldsymbol{\theta} \in \tilde{V}} I(\boldsymbol{\theta} \in V) \frac{\prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}{\int_{\boldsymbol{\theta} \in \tilde{V}} \prod_{i=1}^{\ell} \phi\left\{\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right\} / \sqrt{\left(1-\lambda_{i}\right) \delta^{2}} d \boldsymbol{\theta}} d \boldsymbol{\theta} .
$$

A Monte Carlo estimator of $C$ is

$$
\hat{C}=\frac{1}{M} \sum_{h=1}^{M} I\left(\boldsymbol{\theta}^{(h)} \in V\right)
$$

where

$$
\theta_{i}^{(h)} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left\{\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta},\left(1-\lambda_{i}\right) \delta^{2}\right\}, c_{i}<\theta_{i}^{(h)}<\infty, h=1, \ldots, M=1,000, i=1, \ldots, \ell .
$$

However, the term, $\frac{1}{M} \sum_{h=1}^{M} I\left(\boldsymbol{\theta}^{(h)} \in V\right)$, is difficult to incorporate into the accept-reject sampler. We have overcome the difficulty in the following manner. We have computed $\hat{C}$ and found that more than $60 \%$ of the $\hat{C}$ leads to acceptance of all the $\sigma_{i}^{2}, i=1, \ldots, \ell$. When the $\sigma_{i}^{2}$ are not accepted, we draw samples from their unrestricted distributions, $\sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{IG}\left(a_{i}, b_{i}\right), i=1, \ldots, \ell$.

## 4. Comparisons using simulated examples

We compare our models using simulated examples. We are not performing a simulation study, where replication is important because the models are already complicated. We use the coefficient of variation as a measure of reliability for the comparisons. We also show graphically how the observations in the simulated data violate the lower bound constraints and how this problem is corrected by our models.

Both NASS survey data and USDA administrative acreage data are subject to confidentiality protections, therefore, we describe a means of simulating data with similarity to Illinois corn crop data that have been used extensively in recent NASS studies on crop county estimates and we use it to show the key features of our benchmarking procedure with inequality constraints. As a practical matter, participation in farm support programs can vary by crop and by state. Some of the survey estimates may already satisfy the lower bound constraint, i.e., some $\hat{\theta}_{i}>c_{i}$, so that the lower bound constraints imposed on model estimates for these areas may be loose or non-binding restrictions in those counties. However, in states with high rates of enrollment in farm support programs, like the corn crop in Illinois, administrative totals may capture large parts of the population, so that direct estimates, subject to sampling error, fall below to the administrative totals in many counties. The model estimates of the counties must be constrained by the lower bounds and the benchmarking target as well.

In Section 4.1, we describe several simulated data sets. In Section 4.2, we present results under the single shrinkage model with the inequality constraints. In Section 4.3, we present results under the double shrinkage model for the gamma regression model and the log-linear model, again with the inequality constraints. At the same time, we have compared these models with the direct estimates (DE), the estimates from the Bayesian Fay-Herriot model (ME), without benchmarking or inequality constraints, and the Bayesian Fay-Herriot model with random benchmarking (MERB) at both the county level and at the level of agricultural statistic districts (discussed below).

It is worth noting that all computations were performed on a machine with CentOS (version 6.10) operative system using an Intel CPU Xeon E5-2690 at 2.90 GHz having 16 logical cores, 128GB of RAM, and the software was compiled with ifort version 11.1.

### 4.1 Description of the simulated data sets

Nandram, Erciulescu and Cruze (2019) simulated a data set similar to the one in Battese, Harter and Fuller (1988); see also Toto and Nandram (2010) and Nandram, Toto and Choi (2011). These data are on planted acres of corn and soybeans for 37 segments with 12 counties in the state of Iowa and there are two covariates. (Like Illinois, Iowa is a large corn producing state in the United States.) A Bayesian version of the small area model of Battese, Harter and Fuller (1988) is described in Toto and Nandram (2010); see also Molina, Nandram and Rao (2014).

By simulating from these data, we can create a data set with as many areas we please. In particular Illinois has $\ell=102$ counties grouped in 9 smaller-than-state regions called Agricultural Statistics Districts (ASDs). The data are processed to obtain the survey estimates and standard errors. In our simulated data, based on the actual sizes of the ASDs, we have taken the first set of counties to be in the first ASD, the second set to be in the second ASD and so on so that the first 12 counties correspond to the first ASD, the next 11 correspond to the second, and the remaining ASDs have $9,11,7,13,15,12,12$ counties, respectively. In the process of simulating acreage data, we also added a random effect for each ASD. The sample sizes within the counties are chosen uniformly in $(2,74)$, a realistic range of sample sizes across the state comparable to actual Illinois corn data reported during the 2014 crop year (Erciulescu, Cruze and Nandram, 2018, 2019). Additionally, county-level coefficients of variation $\mathrm{CV}_{i}$ will be simulated uniformly from within the range of $(0.08,0.93)$; these extremes are comparable to values reported in Erciulescu, Cruze and Nandram (2020) in reference to the 2015 crop year. Given simulated survey estimates and coefficients of variation, computed standard errors are obtained $\hat{\sigma}_{i}=\mathrm{CV}_{i} \times \hat{\theta}_{i}$. Thus, we have a data set with the survey estimates, $\hat{\theta}_{i}$, survey standard error, $\hat{\sigma}_{i}$ and sample sizes, $n_{i}$ for the $i^{\text {th }}$ county, $i=1, \ldots, \ell$.

The last piece to be simulated is the data corresponding to the administrative acreage values, i.e., lower bounds, $c_{i}$. For simplicity, we call these the FSA values throughout the simulation. In order to reflect the relationship between the FSA values and the survey estimates for Illinois, we assume the following equation holds,

$$
c_{i}=\hat{\theta}_{i}+U_{i} \times \hat{\theta}_{i}, i=1, \ldots, \ell,
$$

where $U_{i} \stackrel{\text { iid }}{\sim}$ Uniform $(-s, s)$ and $s$ is taken to be a suitable value (e.g., $s=0.10$ ). However, the key problem is how to set the benchmarking target. In the real problem, we will know the target, but the target has to be larger than the sum of the lower bounds. Therefore, it is sensible to take the target to be $a=c / d$, where $c \equiv \sum_{i=1}^{\ell} c_{i}$ and specify $0<d<1$. The completeness of the administrative data relative to the state total can vary by state and crop, but in Illinois, this value will often be close to 1 .

### 4.2 Results under the single shrinkage model

In applying the methodology for an inequality-constrained model with fixed variances developed in Section 2, we specify a plausible value of $d=0.99$ indicating the simulated administrative data embody $99 \%$ of the state-level planted area total for corn in Illinois. In this first instance, we restrict the range of
coefficients of variation to $(0.05,0.25)$. Figure 4.1 shows the simulated survey estimates of $\hat{\theta}_{i}$ versus the FSA values $c_{i}$ (top panel) and the posterior mean of $\theta_{i}$ versus the FSA values $c_{i}$ under the Bayesian Fay-Herriot model with inequality constraint and benchmarking, not including double shrinkage (call this model MFSA-NDS). In the top panel, we can see many points are above or below the $45^{\circ}$ straight line through the origin. (This resembles a realistic pattern shown in Figure 4.4 of Erciulescu, Cruze and Nandram (2020), as applied to the 2015 Illinois corn crop.) Where the survey estimates for many counties are below their corresponding FSA values, all points in the bottom are immediately above the $45^{\circ}$ straight line through the origin, indicating that all MFSA-NDS estimates are no smaller than their corresponding FSA values. Moreover, the sum of the 102 MFSA is equal to the state total, satisfying the benchmarking requirement by raking to the state target.

In Table 4.1, we present results for Illinois simulated data. We compare the results with our new model that incorporates the inequality constraints (FSA values are lower bounds of the model estimates). Specifically, we compare estimates from DE, ME and MERB and the single shrinkage Bayesian Fay-Herriot model with inequality constraint and benchmarking (MFSA-NDS).

The minimum, median and maximum posterior coefficients of variation (expressed as percents, \%) are smaller than the other two models (ME, MERB), even more so for the direct estimates (DE). Of course, as expected, the coefficients of variation for the ASDs are smaller than those for the counties; there is one exception ( 5.13 versus 5.31 in Table 4.1, but this is a minor difference). It is worth noting that county minimum and ASD minimum are not comparable as the county with the minimum CV is not necessarily nested in the ASD with the minimum CV. We note that, as expected, the coefficients of variation are in decreasing order (DE, ME, MERB, MFSA), and modeling appears beneficial, but more importantly we can accommodate the FSA values in our model (MFSA) and provide much smaller coefficients of variation.

Table 4.1
Coefficients of variation (\%) for Illinois simulated data for 102 counties and 9 Agricultural Statistical Districts, fixed variances.

| Level | Statistic | DE | ME | MERB | MFSA-NDS |
| :---: | :---: | :---: | :---: | :---: | :---: |
| County | min | 5.13 | 4.76 | 4.79 | 0.57 |
|  | median | 15.57 | 10.67 | 10.58 | 0.97 |
|  | max | 24.93 | 15.80 | 15.34 | 5.22 |
| ASD | min | 5.31 | 2.54 | 2.39 | 0.24 |
|  | median | 10.60 | 3.25 | 3.01 | 0.30 |
|  | max | 14.81 | 3.92 | 3.51 | 0.42 |

Notes: MFSA is the new benchmarking model with FSA values as lower bounds for the model estimates, CV(0.05-0.25) and $d=0.99$. ASD = Agricultural Statistics Districts; CV = Coefficient of variation; DE = Direct estimates; FSA = Farm Service Agency; ME = Bayesian Fay-Herriot model; MERB = Bayesian Fay-Herriot model with random benchmarking; MFSA = Bayesian Fay-Herriot model with inequality constraint and benchmarking; NDS $=$ Not including double shrinkage.

Figure 4.1 Plots of survey estimates (top panel) and posterior means (bottom panel) under MFSA-NDS for $\theta$ versus FSA values for Illinois and the simulated data, not double shrinkage, CV(0.05-0.25) and $d=0.99$.


Notes: $\mathrm{CV}=$ Coefficient of variation; FSA = Farm Service Agency; MFSA = Bayesian Fay-Herriot model with inequality constraint and benchmarking; NDS $=$ Not including double shrinkage.

### 4.3 Results under the double shrinkage model

Fitting the double shrinkage model with the inequality constraints of Section 3 and denoting these estimates as MFSA-DS, we fit the model to the data already generated for Section 4.2. That is, data for which the simulated $\mathrm{CV}_{i} \in(0.05,0.25)$ and $d=0.99$. Summaries of the coefficients in variation for the MFSA-DS are given in Table 4.2, with the first four columns duplicated from Table 4.1. We notice a small difference between the double shrinkage model and the single shrinkage model. Over counties the maximum CV under the double shrinkage model is a bit smaller than the one under the MFSA-NDS models, $3.58 \%$ versus $5.22 \%$ for the fixed-variances case, but over ASDs (aggregates of counties within) there are smaller differences between the two approaches.

The top panel in Figure 4.2 once again plots the survey estimates versus FSA values (identical to top panel, Figure 4.1), and the lower panel is a plot of the posterior means versus the FSA values under the double shrinkage model with benchmarking and inequality constraints. The lower panel of Figure 4.2 is only slightly different from that of Figure 4.1, in part because the value $d=0.99$ implies that there is little slack between the state target and the total of administrative data summed over all counties in the state.

Table 4.2
Coefficients of variation (\%) for Illinois simulated data for 102 counties and 9 Agricultural Statistical Districts, double shrinkage, gamma model.

| Level | Statistic | DE | ME | MERB | MFSA-NDS | MFSA-DS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| County | min | 5.13 | 4.76 | 4.79 | 0.57 | 0.55 |
|  | median | 15.57 | 10.67 | 10.58 | 0.97 | 1.01 |
|  | max | 24.93 | 15.80 | 15.34 | 5.22 | 3.58 |
| ASD | min | 5.31 | 2.54 | 2.39 | 0.24 | 0.26 |
|  | median | 10.60 | 3.25 | 3.01 | 0.30 | 0.34 |
|  | max | 14.81 | 3.92 | 3.51 | 0.42 | 0.41 |

Notes: MFSA is the new benchmarking model with FSA values as lower bounds for the model estimates. MFSA-DS refers to the double shrinkage model with benchmarking and inequality constraint, $\mathrm{CV}(0.05-0.25)$ and $d=0.99$.
ASD = Agricultural Statistics Districts; CV = Coefficient of variation; DE = Direct estimates; DS = Double shrinkage; FSA = Farm Service Agency; ME = Bayesian Fay-Herriot model; MERB = Bayesian Fay-Herriot model with random benchmarking; MFSA = Bayesian FayHerriot model with inequality constraint and benchmarking; NDS $=$ Not including double shrinkage.

Figure 4.2 Plots of survey estimates (top panel) and posterior means (bottom panel) under MFSA-DS for $\theta$ versus FSA values for Illinois and the simulated data, double shrinkage, gamma regression, $\mathrm{CV}(0.05-0.25)$ and $d=0.99$.


Notes: $\mathrm{CV}=$ Coefficient of variation; DS = Double shrinkage; FSA = Farm Service Agency; MFSA = Bayesian Fay-Herriot model with inequality constraint and benchmarking.

For the purposes of demonstrating the log-linear model, a second data set with slightly different features has been generated. Namely, we specify lower coverage of the FSA values $(d=0.95)$ and allow a higher range of values of survey coefficients of variation, ( $0.08,0.93$ ), comparable to the actual survey coefficients of variation observed during the 2015 crop year. We present summaries of the CVs in Table 4.3. Again we notice a small difference between the log-linear double shrinkage model and the single shrinkage model at the ASD level. Differences in coefficients of variation at the county level are minimal for the lower half of all counties, but the maximum county CV obtained from the double shrinkage model (23.94\%) is substantially smaller than the maximum CV obtained under the single shrinkage model (fixed-variances case) $(44.92 \%)$. Of course, as expected, the coefficients of variation for the ASDs are smaller than those for the counties; there is one exception for DE ( 8.57 versus 18.90 in Table 4.3). Again, it is worth noting that county minimum and ASD minimum are not comparable as the county with the minimum CV is not necessarily nested in the ASD with the minimum CV. Yet the models correct this problem.

In its upper panel, Figure 4.3 depicts the new simulated survey estimates versus their corresponding FSA values, while the lower panel shows the posterior means of the log-linear MFSA-DS model versus the corresponding FSA values. In contrast to the $d=0.99$ data set of the previous sections, the present $d=$ 0.95 data set represents a looser lower-bound constraint. Accordingly, the resulting county acreage estimates, which also sum to the state total, are all visibly above the $45^{\circ}$ line. For comparison, the MFSADS estimates obtained under gamma regression are plotted in the lower panel of Figure 4.4. The two approaches to double shrinkage yield similar (not identical) point estimates given the same state target and administrative lower bound constraints.

In contrast to the computationally expensive gamma regression which required in excess of 16 hours of run time, results of the log-linear model were obtained in a matter of minutes, and additional opportunities to speed up the process may be possible through approximate Bayesian computation described in Appendix B.

## Table 4.3

Coefficients of variation (\%) for Illinois simulated data for 102 counties and 9 Agricultural Statistical Districts, double shrinkage, log-linear model.

| Level | Statistic | DE | ME | MERB | MFSA-NDS | MFSA-DS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| County | min | 8.57 | 7.73 | 7.90 | 2.57 | 2.54 |
|  | median | 52.90 | 17.83 | 17.16 | 4.56 | 4.92 |
|  | max | 92.70 | 24.25 | 24.82 | 44.92 | 23.94 |
| ASD | min | 18.90 | 5.11 | 3.78 | 1.17 | 1.11 |
|  | median | 37.70 | 6.15 | 4.71 | 1.83 | 1.43 |
|  | max | 52.10 | 7.19 | 5.81 | 2.65 | 1.63 |

[^9]Figure 4.3 Plots of survey estimates (top panel) and posterior means (bottom panel) under MFSA for $\boldsymbol{\theta}$ versus FSA values for Illinois and the simulated data, double shrinkage, log-linear model; CV(0.08-0.93) and $d=0.95$.


Notes: $\mathrm{CV}=$ Coefficient of variation; FSA $=$ Farm Service Agency; MFSA $=$ Bayesian Fay-Herriot model with inequality constraint and benchmarking.

Figure 4.4 Plots of survey estimates (top panel) and posterior means (bottom panel) under MFSA-DS for $\boldsymbol{\theta}$ versus FSA values for Illinois and the simulated data, double shrinkage, gamma regression, $\mathrm{CV}(0.08-0.93)$ and $d=0.95$.



Notes: $\mathrm{CV}=$ Coefficient of variation; DS = Double shrinkage; FSA = Farm Service Agency; MFSA = Bayesian Fay-Herriot model with inequality constraint and benchmarking.

## 5. Concluding remarks

Beginning with the 2020 crop year, NASS successfully converted its county-estimates data product into a system model-based estimates of planted area, harvested area, total production, and yield per harvested acre. It is true that our methods can be applied to yield directly; only a small adjustment is needed in the benchmarking in the output analysis. The official estimates for 13 different commodity crops grown nationwide now include a benchmarking of county estimates to predetermined state targets, and lower bound constraints on planted area. Motivated by the needs of the NASS crop estimation program to produce coherent published tables across all parameters and with respect available administrative data, we have shown how to incorporate the area-specific inequality constraints and benchmarking into the Fay-Herriot model. Single shrinkage model and double shrinkage models are available. Because there are difficulties in performing full Metropolis samplers, we overcame these computational difficulties by making additional reasonable approximations in the double shrinkage model.

It is possible to extend the hierarchical Bayesian model so that all the constraints are actually included in it. That is, $\boldsymbol{\theta}$ is in $V=\left\{\boldsymbol{\theta}: c_{i} \leq \theta_{i}, \sum_{i=1}^{n} \theta_{i}=a\right\}$, where $a$ is the benchmarking target and $c_{1}, \ldots, c_{n}$ are the FSA values. So that the hierarchical Bayesian model (i.e., extended version of the Bayesian Fay-Herriot model) has $\boldsymbol{\theta} \in V$. We have attempted to do so for the simplest model, the Bayesian Fay-Herriot model, but the problem is extremely difficult. It requires the computation of orthant probabilities (e.g., Ridgway, 2016; Geweke, 1991; Genz, 1992) at each step of a Markov chain Monte Carlo sampler. There are no such problems mentioned in Rao and Molina (2015), although they have used the raking procedure for benchmarking only, not the inequality constraints, where the $\theta_{i}>c_{i}$, the FSA problem.

Nevertheless, incorporating the total constraint into the hierarchical Bayesian model will be beneficial because it will help protect against model failure so prominent in small area estimation, and one needs to be careful with this. Toto and Nandram (2010), Nandram and Sayit (2011) and Nandram, Toto and Choi (2011), Nandram, Erciulescu and Cruze (2019) and Janicki and Vesper (2017) were able to incorporate a much simpler constraint (i.e., $\sum_{i=1}^{n} \theta_{i}=a$ ) in a complete Bayesian analysis. But as is evident, it is much more difficult to incorporate the constraint $\boldsymbol{\theta} \in V$, and it is a problem we would like to solve in the future. We can add random effects on both means and variances to accommodate sub-areas (counties within ASDs). However, the computations are difficult and approximations beyond those based on Markov chain Monte Carlo methods need to be considered. Currently, we are doing research in this area.

In Appendix C, we have comments on generalizion. It is possible to avoid the inequality constraint using a logarithmic transformation, but this method looses generality or it makes unnecessary approximation. Our solution remains strong for both the single shrinkage model and the double shrinkage model.

## Appendix

## A. Double-shrinkage model fitting - Gamma regression

Dropping the inequality constraint of the double shrinkage model (see (3.3)), the joint posterior density is

$$
\begin{align*}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \Omega \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) & \propto \pi\left(\boldsymbol{\beta}, \delta^{2}, \boldsymbol{\gamma}, \alpha\right) \prod_{i=1}^{\ell}\left\{\left(\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma} / 2\right)^{\alpha / 2}\left(1 / \sigma_{i}^{2}\right)^{\alpha / 2+1} e^{-\left(\alpha e^{\left.-x_{i} / 2 \sigma_{i}^{2}\right)} / \Gamma(\alpha / 2)\right\}}\right. \\
& \times \prod_{i=1}^{\ell}\left\{\frac{1}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}} \phi\left(\frac{\theta_{i}-\left(\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta}\right)}{\sqrt{\left(1-\lambda_{i}\right) \delta^{2}}}\right) \frac{1}{\sqrt{\delta^{2} / \lambda_{i}}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\right\}  \tag{A.1}\\
& \times \prod_{i=1}^{\ell}\left\{\left[\left(n_{i}-1\right) / 2 \sigma_{i}^{2}\right]^{\left(n_{i}-1\right) s_{i}^{2} / 2} e^{-\left(n_{i}-1\right) / 2 \sigma_{i}^{2}}\right\},
\end{align*}
$$

where $\lambda_{i}=\delta^{2} /\left(\delta^{2}+\sigma_{i}^{2}\right), i=1, \ldots, \ell$. Conditional on $\Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}$, it is clear that $\left(\theta_{i}, \sigma_{i}^{2}\right)$ are independent over $i=1, \ldots, \ell$. This is the key difference between the double-shrinkage model with and without the inequality constraints.

Our strategy is to first sample the posterior density $\pi\left(\Omega \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$. Once this is done, we draw samples from the joint conditional posterior density of $\pi\left(\boldsymbol{\sigma}^{2} \mid \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$. Then, finally we obtain the required samples from $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{\sigma}^{2}, \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$. Thus, after draws are obtained for $\Omega$, we use the multiplication rule to get the $\sigma_{i}^{2}$ and $\theta_{i}$ (i.e., $\Omega$, the $\sigma_{i}^{2}$ and $\theta_{i}$ are drawn simultaneously).

It follows from (A.1) that the conditional on $\boldsymbol{\sigma}^{2}, \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}$, the $\theta_{i}$ are independent and

$$
\begin{equation*}
\theta_{i} \mid \boldsymbol{\sigma}^{2}, \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left\{\lambda_{i} \hat{\theta}_{i}+\left(1-\lambda_{i}\right) \mathbf{x}_{i}^{\prime} \boldsymbol{\beta},\left(1-\lambda_{i}\right) \delta^{2}\right\}, i=1, \ldots, \ell . \tag{A.2}
\end{equation*}
$$

Conditional on $\Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}$, the $\sigma_{i}^{2}$ are independent. Therefore, integrating out the $\theta_{i}$ from (A.1), we have the conditional posterior density of $\sigma_{i}^{2}$ is

$$
\begin{equation*}
\pi\left(\sigma_{i}^{2} \mid \Omega, \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) \propto \sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right)\left[\left(1 / \sigma_{i}^{2}\right)^{\left(n_{i}+\alpha-1\right) / 2+1} e^{-\left\{\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-x_{i}}\right\} / 2 \sigma_{i}^{2}}\right], \tag{A.3}
\end{equation*}
$$

$i=1, \ldots, \ell$. Note unnecessary constants are dropped (e.g., parameters conditioned on).
Now, one can integrate out the $\theta_{i}$ and $\sigma_{i}^{2}$ from (A.1) to get the joint posterior density of $\Omega$,

$$
\begin{align*}
\pi\left(\Omega \mid \hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right) & \propto \pi\left(\boldsymbol{\beta}, \delta^{2}, \boldsymbol{\gamma}, \alpha\right) \prod_{i=1}^{\ell}\left\{\frac{\Gamma(\alpha / 2)}{\left(\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma} / 2\right)^{\alpha / 2}} \frac{\Gamma\left(\left(n_{i}+\alpha-1\right) / 2\right)}{\left\{\left(\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-\mathbf{x}_{i}^{\prime} \gamma}\right) / 2\right\}^{\left(n_{i}+\alpha-1\right) / 2}}\right\}  \tag{A.4}\\
& \times \prod_{i=1}^{\ell}\left\{\int_{0}^{\infty} \frac{1}{\sqrt{\delta^{2} / \lambda_{i}}} \phi\left(\frac{\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}}{\sqrt{\delta^{2} / \lambda_{i}}}\right) \mathrm{IG}_{\sigma_{i}^{2}}\left(a_{i}, b_{i}\right) d \sigma_{i}^{2}\right\},
\end{align*}
$$

where $a_{i}=\left(n_{i}+\alpha-1\right) / 2$ and $b_{i}=\left\{\left(n_{i}-1\right) s_{i}^{2}+\alpha e^{-\mathrm{x}^{\prime} ; \gamma}\right\} / 2$. Here, $\mathrm{IG}_{x}(a, b)$ is the inverse gamma density and is given by $f(x)=b^{a}\left(\frac{1}{x}\right)^{a+1} e^{-b / x} / \Gamma(a), x>0$.

It is easy to sample the $\sigma_{i}^{2}$ in (A.3) using the accept-reject sampler; simply draw $\sigma_{i}^{2} \mid \Omega, \hat{\boldsymbol{\theta}}$, $\mathbf{S}^{2} \sim \operatorname{IG}\left(a_{i}, b_{i}\right)$ and take it with probability $\sqrt{\lambda_{i}} \phi\left(\frac{\hat{\theta}_{i}-x_{i} \beta}{\delta^{2} / \lambda_{i}}\right)$. Then, clearly the $\theta_{i}$ are easy to draw from (A.2). The main problem now is how to sample the joint posterior density of $\Omega$ in (A.4). We will use the Metropolis sampler to do so.

Once we obtain a sample from (A.4), we convert it to a sample from (3.6), our main objective. This is accommodated by another Metropolis sampler that we execute in a novel manner. We prefer to use proposal densities that will provide independent chains. This is obtained by taking draws from a multivariate Student's $t$ density (to be constructed). We will not use a long run because with a Metropolis sampler, the chain tends to get stuck a long time, introducing long-range dependence to the sample, thereby giving poor mixing that is inefficient. Instead we run several chains, say $M=1,000$ chains. Each chain is run with a random start from an approximate density for 100 iterates, and the last one is taken. Only minor monitoring is needed to ensure reasonable jumping rates. If the chain does not move from the initial random start, it is not used in the final sample. In the end, we get a random sample of $M$ iterates from the required density in (A.4).

We describe how to obtain samples from the posterior density of $\Omega=\left(\boldsymbol{\beta}, \delta^{2}, \boldsymbol{\gamma}, \alpha\right)$. There are three steps. The first step obtains a sample of $M$ starting values, the second step is to obtain a proposal density for Metropolis sampler at each of the starting values, and the third step is to make a short run of 100 iterates of each of the Metropolis samplers in second step.

First, we integrate out the $\theta_{i}$ and we replace the $\sigma_{i}^{2}$ by $s_{i}^{2}, i=1, \ldots, \ell$. Given $\hat{\boldsymbol{\theta}}, \mathbf{s}^{2}$, then $\left(\boldsymbol{\beta}, \delta^{2}\right)$ and ( $\gamma, \alpha$ ) are independent; so they can be sampled separately to get $M=1,000$ independent starts. We have obtained these $M$ starts using simple approximations.

Second, at each start, we run a Gibbs sampler to get $\sigma_{i}^{2}, i=1, \ldots, \ell$, and $\Omega$. This is done by drawing the $\sigma_{i}^{2}$ from their exact conditional posterior densities using rejection sampling. Then, given $\boldsymbol{\sigma}^{2}, \mathbf{s}^{2},\left(\boldsymbol{\beta}, \delta^{2}\right)$ and $(\gamma, \alpha)$ are again independent, and draws from their respective joint posterior densities are taken in a similar manner. It is worth noting that given $\delta^{2}$, the distribution of $\boldsymbol{\beta}$ is multivariate normal and $\boldsymbol{\beta}$ can be integrated out to get the conditional posterior density of $\delta^{2}$ that can be sampled using a grid. However, this is not the case for $(\gamma, \alpha)$ because the conditional posterior density of $\gamma$ given $\alpha$ is nonstandard (i.e., not multivariate normal). Thus, we approximate the posterior density of $\gamma$ using a multivariate normal density, and with this approximation, sampling of $(\boldsymbol{\gamma}, \alpha)$ takes place in the same manner as for $\left(\boldsymbol{\beta}, \delta^{2}\right)$.

Third, we run the second step 1,100 times with a "burn-in" of 100 runs and we use the $M=1,000$ samples to construct a multivariate Student's $t$ density for $\Omega_{a}=\left(\boldsymbol{\beta}, \log \left(\delta^{2}\right), \gamma, \log (\alpha)\right)$, which we use as a proposal density in a Metropolis sampler to sample the exact posterior density. This is performed 100 times and the last iterate is selected. Each random start contributes to the sample of $M=1,000$ iterates of $\Omega_{a}$ or $\Omega$ from the posterior density under the double shrinkage model without the inequality constraint and the benchmarking.

To complete the entire procedure, for each $\Omega_{a}$, we sample $\sigma_{i}^{2}$ from their conditional posterior densities using rejection sampling to access the posterior densities more efficiently. Then, more importantly, the $\theta_{i}$ are drawn from their conditional posterior densities (normal is this case). The entire procedure took roughly four hours, and the jumping rates are mostly larger than $5 \%$.

## B. Double shrinkage model fitting - Log-linear model

We describe the double shrinkage log-linear model and show how to fit. The main purpose is to show that there are additional gains in computational speed using approximate Bayesian computation.

Our model is similar to the one in Section 3, where assuming that $\hat{\theta}_{i}$ and $s_{i}^{2}$ are pairwise independent,

$$
\begin{gathered}
\hat{\theta}_{i} \mid \theta_{i}, \sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\theta_{i}, \sigma_{i}^{2}\right), i=1, \ldots, \ell, \\
\frac{\left(n_{i}-1\right) s_{i}^{2}}{\sigma_{i}^{2}} \left\lvert\, \sigma_{i}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Gamma}\left(\frac{n_{i}-1}{2}, \frac{1}{2}\right)\right., i=1, \ldots, \ell
\end{gathered}
$$

However, a priori, we assume that

$$
\theta_{i} \mid \boldsymbol{\beta}_{1}, \delta_{1}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{\mathbf{i}}^{\prime} \boldsymbol{\beta}_{1}, \delta_{1}^{2}\right), i=1, \ldots, \ell, \boldsymbol{\theta} \in V,
$$

with the log-linear model on the $\sigma_{i}^{2}$,

$$
\ln \left(\sigma_{i}^{2}\right) \mid \boldsymbol{\beta}_{2}, \delta_{2}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{2}, \delta_{2}^{2}\right), i=1, \ldots, \ell
$$

where we also assume that $\theta_{i}$ and $\sigma_{i}^{2}$ are pairwise independent. Note that we also have the restriction $\boldsymbol{\theta} \in V$. Because we will use an approximate Gibbs sampler to fit the model, we assume that $\pi\left(\boldsymbol{\beta}_{1}, \boldsymbol{\beta}_{2}, \delta_{1}^{2}\right.$, $\left.\delta_{2}^{2}\right) \propto \frac{1}{\delta_{1}^{2}} \frac{1}{\delta_{2}^{2}}$ (i.e., posterior propriety is not an issue provided that the design matrix is full rank).

Then, letting $D=\left(\hat{\boldsymbol{\theta}}, \mathbf{s}^{2}\right)$, the joint posterior density of $\boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \boldsymbol{\beta}_{1}, \delta_{1}^{2}, \boldsymbol{\beta}_{2}, \delta_{2}^{2}$ is given by

$$
\begin{aligned}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^{2}, \boldsymbol{\beta}_{1}, \delta_{1}^{2}, \boldsymbol{\beta}_{2}, \delta_{2}^{2} \mid D\right) & \propto \prod_{i=1}^{\ell}\left\{\frac{1}{\sqrt{2 \pi \sigma_{i}^{2}}} e^{-\left(\hat{\theta}_{i}-\theta_{i}\right)^{2} / 2 \sigma_{i}^{2}}\right\} \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\} d \boldsymbol{\theta}} \\
& \times \frac{1}{\delta_{1}^{2}} \frac{1}{\delta_{2}^{2}} \prod_{i=1}^{\ell}\left\{\left(\frac{n_{i}-1}{\sigma_{i}^{2}}\right)^{\left(n_{i}-1\right) / 2} e^{-\left(n_{i}-1\right) s_{i}^{2} / 2 \sigma_{i}^{2}} \frac{1}{\sqrt{2 \pi \delta_{2}^{2}}} e^{-\left(\ln \left(\sigma_{i}^{2}\right)-x_{i} \boldsymbol{\beta}_{2}\right)^{2} / 2 \delta_{2}^{2}}\right\}, \boldsymbol{\theta} \in V .
\end{aligned}
$$

Our strategy in the computation is to sample the exact conditional posterior density of $\theta_{i}, i=1, \ldots, \ell$, and $\sigma_{i}^{2}, i=1, \ldots, \ell$. However, we want to replace the conditional posterior densities of $\boldsymbol{\beta}_{1}, \delta_{1}^{2}$ and $\boldsymbol{\beta}_{2}, \delta_{2}^{2}$ by approximate posterior densities. The main issue now is how to do this latter task.

We consider the two simpler models for $\hat{\theta}_{i}$ and $s_{i}^{2}, i=1, \ldots, \ell$. These are

$$
\hat{\theta}_{i} \mid \boldsymbol{\beta}_{1}, \delta_{1}^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}, \delta_{1}^{2}\right), i=1, \ldots, \ell, \pi\left(\boldsymbol{\beta}_{1}, \delta_{1}^{2}\right) \propto 1 / \delta_{1}^{2},
$$

and

$$
\ln \left(s_{i}^{2}\right) \mid \boldsymbol{\beta}_{2}, \delta_{2}^{2} \sim \operatorname{ind} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{2}, \delta_{2}^{2}\right), i=1, \ldots, \ell, \pi\left(\boldsymbol{\beta}_{2}, \delta_{2}^{2}\right) \propto 1 / \delta_{2}^{2} .
$$

Note that in the full model, we simply replace the $\theta_{i}$ by $\hat{\theta}_{i}$ and $\sigma_{i}^{2}$ by $s_{i}^{2}$. Here, the posterior densities of $\left(\boldsymbol{\beta}_{1}, \delta_{1}^{2}\right)$ and $\left(\boldsymbol{\beta}_{2}, \delta_{2}^{2}\right)$, which are independent, have simple forms. Letting $X$ denote the $n \times p$ design matrix, then

$$
\boldsymbol{\beta}_{1}\left|\hat{\boldsymbol{\theta}}, \delta_{1}^{2} \sim \operatorname{Normal}\left\{\hat{\beta}_{1},\left(X^{\prime} X\right)^{-1} \delta_{1}^{2}\right\}, \delta_{1}^{2}\right| \hat{\boldsymbol{\theta}} \sim \mathrm{IG}\left\{\frac{n-p}{2}, \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \hat{\beta}_{1}\right)^{2}}{2}\right\},
$$

where $\hat{\beta}_{1}=\left(X^{\prime} X\right)^{-1} X^{\prime} \hat{\boldsymbol{\theta}}$. Therefore, the posterior density of $\boldsymbol{\beta}_{1}$ is a multivariate Student's $t$ density, and, in this case, it is easy to draw samples of $\boldsymbol{\beta}_{1}$ and $\delta_{1}^{2}$. In addition, letting $z_{i}=\ln \left(s_{i}^{2}\right), i=1, \ldots, \ell$, then

$$
\boldsymbol{\beta}_{2}\left|\mathbf{z}, \delta_{2}^{2} \sim \operatorname{Normal}\left\{\hat{\beta}_{2},\left(X^{\prime} X\right)^{-1} \delta_{2}^{2}\right\}, \delta_{2}^{2}\right| \mathbf{z} \sim \operatorname{IG}\left\{\frac{n-p}{2}, \frac{\sum_{i=1}^{n}\left(z_{i}-\mathbf{x}_{i}^{\prime} \hat{\beta}_{2}\right)^{2}}{2}\right\},
$$

where $\hat{\beta}_{2}=\left(X^{\prime} X\right)^{-1} X^{\prime} \mathbf{z}$. Again, the posterior density of $\boldsymbol{\beta}_{2}$ is a multivariate Student's $t$ density, and it is easy to draw samples of $\boldsymbol{\beta}_{2}$ and $\delta_{2}^{2}$. Our approximate Gibbs sampler runs by taking these posterior densities as the conditional posterior densities. We need to do so because the computation is difficult and timeconsuming.

The joint density of $\left(\theta_{i}, \sigma_{i}^{2}\right), i=1, \ldots, \ell$, is

$$
\begin{aligned}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^{2} \mid \boldsymbol{\beta}_{1}, \delta_{1}^{2}, \boldsymbol{\beta}_{2}, \delta_{2}^{2}, D\right) & \propto \prod_{i=1}^{\ell}\left\{\frac{1}{\sqrt{2 \pi \sigma_{i}^{2}}} e^{-\left(\hat{\theta_{i}}-\theta_{i}\right)^{2} / 2 \sigma_{i}^{2}}\right\} \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\}}{\int_{\theta \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\} d \boldsymbol{\theta}} \\
& \times \prod_{i=1}^{\ell}\left\{\left(\frac{n_{i}-1}{\sigma_{i}^{2}}\right)^{\left(n_{i}-1\right) / 2} e^{-\left(n_{i}-1\right) s_{i}^{2} / 2 \sigma_{i}^{2}} \frac{1}{\sqrt{2 \pi \delta_{2}^{2}}} e^{-\left(\ln \left(\sigma_{i}^{2}\right)-x_{i}^{\prime} \boldsymbol{\beta}_{2}\right)^{2} / 2 \delta_{2}^{2}}\right\}, \boldsymbol{\theta} \in V .
\end{aligned}
$$

Additional difficulties in the computation reside in this joint conditional posterior density. Observe that because $\boldsymbol{\theta} \in V$, the $\theta_{i}$ are not independent, the $\sigma_{i}^{2}$ are not independent and $\theta_{i}$ and $\sigma_{i}^{2}$ are not pairwise independent. However, note that the $\sigma_{i}^{2}$ are independent in their joint conditional posterior density, but the $\theta_{i}$ are not independent in their joint conditional posterior density. The $\sigma_{i}^{2}$ are drawn using the grid method with range $\left(\frac{1}{10} s_{i}^{2}, 10 s_{i}^{2}\right)$, fairly wide, and the $\theta_{i}$ are drawn using Devroye's method.

For the Gibbs sampler, we used 2,500 iterates as a burn-in and took every third iterate to get a random sample of 1,000 iterates. We found that the Geweke tests for all the $\theta_{i}$ and the $\sigma_{i}^{2}$ are not significant and the effective sample sizes are all near the actual sample size of 1,000 (mostly all of them are 1,000 ). Therefore, we have an efficient Gibbs sampler and amazingly the computation took less than 20 seconds.

Next, we describe a slightly different computational method from the one described above. However, we just need to say how to draw samples from the conditional posterior densities of $\left(\boldsymbol{\beta}_{1}, \delta_{1}^{2}\right)$ and $\left(\boldsymbol{\beta}_{2}, \delta_{2}^{2}\right)$.

The conditional posterior density of $\left(\boldsymbol{\beta}_{2}, \delta_{2}^{2}\right)$ is straight forward (i.e., we simply need to replace $s_{i}^{2}$ by $\left.\sigma_{i}^{2}\right)$. So that, letting $z_{i}=\ln \left(\sigma_{i}^{2}\right)$,

$$
\boldsymbol{\beta}_{2}\left|\mathbf{z}, \delta_{2}^{2} \sim \operatorname{Normal}\left\{\hat{\beta}_{2},\left(X^{\prime} X\right)^{-1} \delta_{2}^{2}\right\}, \delta_{2}^{2}\right| \mathbf{z} \sim \operatorname{IG}\left\{\frac{n-p}{2}, \frac{\sum_{i=1}^{n}\left(z_{i}-\mathbf{x}_{i}^{\prime} \hat{\beta}_{2}\right)^{2}}{2}\right\}
$$

It is more difficult to sample the conditional posterior density of $\left(\boldsymbol{\beta}_{1}, \delta_{1}^{2}\right)$,

$$
\pi\left(\boldsymbol{\beta}_{1}, \delta_{1}^{2}\left|\boldsymbol{\theta}, \boldsymbol{\sigma}^{2} \boldsymbol{\beta}_{2}, \delta_{2}^{2},\right| D\right) \propto \frac{1}{\delta_{1}^{2}} \frac{\prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\}}{\int_{\boldsymbol{\theta} \in V} \prod_{i=1}^{\ell} \phi\left\{\left(\theta_{i}-\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}_{1}\right) / \delta_{1}\right\} d \boldsymbol{\theta}}
$$

We started by using the Metropolis sampler. After we have used two different proposal densities, we found long-range dependence with low jumping rates, so we abandoned the Metropolis sampler. We decided to use grid samplers as follows. We fit the simpler model, where letting $X$ denote the $n \times p$ design matrix,

$$
\boldsymbol{\beta}_{1}\left|\hat{\boldsymbol{\theta}}, \delta_{1}^{2} \sim \operatorname{Normal}\left\{\hat{\beta}_{1},\left(X^{\prime} X\right)^{-1} \delta_{1}^{2}\right\}, \delta_{1}^{2}\right| \hat{\boldsymbol{\theta}} \sim \operatorname{IG}\left\{\frac{n-p}{2}, \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{i}-\mathbf{x}_{i}^{\prime} \hat{\beta}_{1}\right)^{2}}{2}\right\}
$$

with $\hat{\boldsymbol{\beta}}_{1}=\left(X^{\prime} X\right)^{-1} X^{\prime} \hat{\boldsymbol{\theta}}$. Therefore, we can now sample $\boldsymbol{\beta}_{1}$ and $\delta_{1}^{2}$ using the multiplication rule. Then, we find the posterior means (PM) and standard deviations (PSD) of each component of $\boldsymbol{\beta}_{1}$ and $\delta_{1}^{2}$; we choose their supports to be $\mathrm{PM} \pm 6^{*} \mathrm{PSD}$ with the lower bound for $\delta_{1}^{2}$ being $\max \left(0, \mathrm{PM}-6^{*} \mathrm{PSD}\right)$. [Almost the entire support of a unimodal density is within this range; actually we have found the procedure to be nonsensitive to the choice of 6 to inference about the $\left.\theta_{1}\right]$. We now run the grid method within the Gibbs sampler to draw $\boldsymbol{\beta}_{1}$ and $\delta_{1}^{2}$ with the supports mentioned above for $\boldsymbol{\beta}$ and $\delta_{1}^{2}$.

For the Gibbs sampler, we used 3,500 iterates as a burn-in and took every fourth iterate to get a random sample of 1,000 iterates. We found that the Geweke tests for all the $\theta_{i}$ and the $\sigma_{i}^{2}$ are not significant and the effective sample sizes are all near the actual sample size of 1,000 (mostly all of them are 1,000 ). Therefore, we have an efficient Gibbs sampler and amazingly the computation took less than 40 seconds. This is double the time (still fast) for the approximate Gibbs sampler above.

## C. Discussions on generalization

We show that the problem is more ubiquitous than we have stated in this paper. Then, we discuss issues with standard solutions using the logarithmic transformation. Recall that our problem is to provide estimates subjected to the lower bound inequality constraints and an equality benchmarking constraint. We discuss mainly the inequality constraint.

The Fay-Herriot model is

$$
\begin{gathered}
\hat{\theta}_{i} \mid \theta_{i} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\theta_{i}, s_{i}^{2}\right) \\
\theta_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}, \delta^{2}\right), i=1, \ldots, \ell
\end{gathered}
$$

with prior $\pi\left(\boldsymbol{\beta}, \delta^{2}\right)$. This is subjected to the inequality constraint, $\theta_{i} \geq c_{i}, i=1, \ldots, \ell$, and the benchmarking constraint, $\sum_{i=1}^{\ell} \theta_{i}=a$, where $a$ is the target. Letting $\hat{\phi}_{i}=\hat{\theta}_{i}-c_{i}, i=1, \ldots, \ell$, and $c=\sum_{i=1}^{\ell} c_{i}$. Then,

$$
\begin{gather*}
\hat{\phi}_{i} \mid \phi_{i} \sim \operatorname{ind} \operatorname{Normal}\left(\phi_{i}, s_{i}^{2}\right)  \tag{C.1}\\
\phi_{i} \mid \boldsymbol{\beta}, \delta^{2} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\mathbf{x}_{i}^{\prime} \boldsymbol{\beta}, \delta^{2}\right), \phi_{i} \geq 0, i=1, \ldots, \ell, \tag{C.2}
\end{gather*}
$$

with $\sum_{i=1}^{\ell} \phi_{i}=a-c$; note that there is a change in the regression coefficients. Therefore, we have a general problem with positivity constraints and a benchmarking constraint, and the problem is not specific to agriculture. The solution of problem remains the same as we have done in this paper, but we can use the logarithmic transformation to avoid the positivity constraint.

There are two ways to proceed without the positivity constraints.
a) Transform the $\hat{\phi}_{i}$, replacing $\hat{\phi}_{i}$ by $\log \left(\hat{\phi}_{i}\right)$ in (C.1). Note that some of the $\hat{\phi}_{i}$ can be negative, thereby loosing some generality. For the case when they are positive, we can approximate the means and the variances of the normal distribution in (C.1) using a first-order Taylor's series approximation. That is, $\log \left(\hat{\phi}_{i}\right) \left\lvert\, \phi_{i} \stackrel{\text { ind }}{\sim} \operatorname{Normal}\left(\log \left(\phi_{i}\right), \frac{s_{i}^{2}}{\phi_{i}^{2}}\right)\right.$. One can proceed in (C.2) with either a log-normal regression or another distribution for positive size data (e.g., gamma regression).
b) Transform the $\phi_{i}$, replacing $\phi_{i}$ by $e^{\phi_{i}}$ in (C.1). This introduces non-conjugacy with (C.2), thereby creating difficulties in computation.

Note again that benchmarking is done in an output analysis as we have done in this paper, and both single shrinkage models and double shrinkage models can be done. When the logarithmic transformation is used, back transformation to the original $\phi_{i}$ is problematic (e.g., Manandhar and Nandram, 2021). However, the methodology in this paper provides our front line solution.

## Acknowledgements and disclaimers

Dr. Cruze's and Dr. Erciulescu's contributions to this research were made during their tenure in the Research and Development Division at USDA's NASS. The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA or US Government determination or policy. This research was supported by the U.S. Department of Agriculture, National Agriculture Statistics Service. Balgobin Nandram was supported by a grant from the Simons Foundation (\#353953, Balgobin Nandram). The authors also thank Dr. Linda Young for her encouragement and the two reviewers for their informative comments and assistance.

## References

Battese, G.E., Harter, R. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), 28-36.

Chen, L., Nandram, B. and Cruze, N.B. (2022). Hierarchical Bayesian model with inequality constraints for US county estimates. Journal of Official Statistics, 38(3), 709-732.

Chen, X., and Nandram, B. (2022). Bayesian inference for multinomial data from small areas incorporating uncertainty about order restriction. Survey Methodology, 48, 1, 145-176. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022001/article/00004-eng.pdf.

Cruze, N.B., Erciulescu, A.L., Nandram, B., Barboza, W.J. and Young, L.J. (2019). Producing official county-level agricultural estimates in the United States: Needs and challenges. Statistical Science, 34(2), 301-316.

Dass, S.C., Maiti, T., Ren, H. and Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. Survey Methodology, 38, 2, 173-187. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012002/article/11756-eng.pdf.

Devroye, L. (1986). Non-Uniform Random Variate Generation, New York: Springer-Verlag.

Erciulescu, A.L., Cruze, N.B. and Nandram, B. (2018). Benchmarking a triplet of official estimates. Environmental and Ecological Statistics, 25, 523-547.

Erciulescu, A.L., Cruze, N.B. and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. Journal of the Royal Statistical Society, Series A, 182(1), 283-303.

Erciulescu, A.L., Cruze, N.B. and Nandram, B. (2020). Statistical challenges in combining survey and auxiliary data to produce official statistics. Journal of Official Statistics, 36(1), 63-88.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366), 269-277.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics, 1, 141-149.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student- $t$ distributions subject to linear constraints and the evaluation of constraint probabilities. Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, 23, 571-578.

Gonzalez-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and SantaMaria, L. (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. Statistical Modelling, 10, 215-239.

Janicki, R., and Vesper, A. (2017). Benchmarking techniques for reconciling Bayesian small area models at distinct geographical levels. Statistical Methods and Applications, 26, 557-581.

Maiti, T., Ren, H. and Sinha, S. (2014). Prediction error of small area predictors shrinking both means and variances. Scandinavian Journal of Statistics, 41, 775-790.

Manandhar, B., and Nandram, B. (2021). Hierarchical Bayesian models for continuous and positively skewed data from small areas. Communications in Statistics - Theory and Methods, 50(4), 944-962.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. The Annals of Applied Statistics, 8(2), 852-885.

Nandram, B., and Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. Journal of the American Statistical Association, 105, 120135.

Nandram, B., Cruze, N.B., Erciulescu, A.L. and Chen, L. (2022). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. RDD Research Report, Number RDD-22-02, National Agricultural Statistics Service, USDA, 1-41.

Nandram, B., Erciulescu, A.L. and Cruze, N.B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. Survey Methodology, 45, 2, 365-390. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2019002/article/00004-eng.pdf.

Nandram, B., and Erhardt, E.B. (2004). Fitting Bayesian two-stage generalized linear models using random samples via the SIR algorithm. Sankhyā: The Indian Journal of Statistics, 66(4), 733-755.

Nandram, B., and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. Survey Methodology, 37, 2, 137-152. Paper avilable at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11603-eng.pdf.

Nandram, B., Sedransk, J. and Smith, S.J. (1997). Order restricted Bayesian estimation of the age composition of a population of Atlantic cod. Journal of the American Statistical Association, 92, 33-40.

Nandram, B., Toto, M.C. and Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. Journal of Statistical Computation and Simulation, 81, 1593-1608.

National Academies of Sciences, Engineering, and Medicine (2017). Improving Crop Estimates by Integrating Multiple Data Sources, The National Academies Press, Washington, DC. https://doi.org/ 10.17226/24892.

Rao, J.N.K., and Molina, I. (2015). Small Area Estimation, Wiley Series in Survey Methodology.

Ridgway, J. (2016). Computation of Gaussian orthant probabilities in high dimension. Statistics and Computing, 26, 899-916.

Sen, P.K., and Silvapulle, M.J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. Journal of Statistical Planning and Inference, 107, 3-43.

Silvapulle, M.J., and Sen, P.K. (2005). Constrained Statistical Inference: Inequality, Order, and Shape Restrictions, New York: John Wiley \& Sons, Inc.

Steorts, R.C., Schmid, T. and Tzavidis, N. (2020). Smoothing and benchmarking for small area estimation. International Statistical Review, 88(3), 580-598.

Toto, M.C.S., and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. Journal of Statistical Planning and Inference, 140, 2963-2979.

Wang, J., and Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. Journal of the American Statistical Association, 98, 716-723.

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.

# Small area prediction of general small area parameters for unit-level count data 

Emily Berg ${ }^{1}$


#### Abstract

We investigate small area prediction of general parameters based on two models for unit-level counts. We construct predictors of parameters, such as quartiles, that may be nonlinear functions of the model response variable. We first develop a procedure to construct empirical best predictors and mean square error estimators of general parameters under a unit-level gamma-Poisson model. We then use a sampling importance resampling algorithm to develop predictors for a generalized linear mixed model (GLMM) with a Poisson response distribution. We compare the two models through simulation and an analysis of data from the Iowa Seat-Belt Use Survey.


Key Words: Poisson; Bootstrap; Small area estimation.

## 1. Introduction

Small area estimation is the problem of constructing estimators for domains where sample sizes are too small to support reliable direct estimators. The standard approach to small area estimation is to use modelbased estimators instead of direct estimators. Model-based estimators garner efficiency gains for small area estimation through restrictions that different areas share common distributional properties and through the incorporation of population-level auxiliary information. Extensive reviews of small area models are available in Rao and Molina (2015), Jiang and Lahiri (2006), and Pfeffermann (2013). More recent reviews include Ghosh (2020) and Molina, Corral and Nguyen (2022). The small area literature has focused heavily on the situation in which the parameter of interest is a small area mean. Many small area parameters are not simple means but are nonlinear functions of the model response variable. Molina and Rao (2010) develops a simulation-based procedure for constructing predictors of small area parameters that may be nonlinear functions of the model response variable. We refer to the types of parameters of interest in Molina and Rao (2010) as "general parameters". Molina, Nandram and Rao (2014), Hobza, Marhuenda and Morales (2020), Rojas-Perilla, Pannier, Schmid and Tzavidis (2020), Marhuenda, Molina, Morales and Rao (2017) and Guadarrama, Molina and Rao (2018) extend Molina and Rao (2010) to Bayesian inference, generalized linear mixed models, data transformations, two-fold models, and complex sampling. We develop predictors of general small area parameters for unit-level count data.

The two primary small area models for count data are (1) the gamma-Poisson model and (2) the Poisson generalized linear mixed model (GLMM). In the context of the area-level model, Reluga, Lombardía and Sperlich (2021) and Boubeta, Lombardía and Morales (2016) develop small area prediction procedures for the gamma-Poisson model and the Poisson GLMM, respectively. We focus on unit-level models. Tzavidis, Ranalli, Salvati, Dreassi and Chambers (2015) develops an M-quantile based procedure for prediction of
small area means for unit-level count data. As demonstrated in Tzavidis et al. (2015), this procedure is less efficient than model-based methods if the model assumptions hold. Jiang and Lahiri (2006) develops an empirical best predictor of the mean for a unit-level Poisson GLMM. Berg (2022) develops empirical best predictors of the mean under a unit-level gamma-Poisson model. We refer to Berg (2022) for a more complete review of unit-level models and area-level models for count data. Jiang and Lahiri (2006), Tzavidis et al. (2015), and Berg (2022) emphasize prediction of means. We develop procedures that are applicable to nonlinear parameters, such as quantiles.

We propose empirical best predictors of general parameters under two unit-level models. The first is a unit-level gamma-Poisson model. The second is a unit-level Poisson GLMM. We establish a common notation that we use for both models. Let $i=1, \ldots, D$ index the small areas, and let $j=1, \ldots, N_{i}$ index the units in the population for small area $i$. Let $y_{i j}$ be the observed count for unit $j$ in small area $i$, where $y_{i j} \in\{0,1,2, \ldots\}$. Let $\mathbf{x}_{i j}=\left(x_{i j 1}, \ldots, x_{i j p}\right)^{\prime}$ be a vector of covariates that does not include an intercept. We consider prediction of a general parameter defined as

$$
\begin{equation*}
\theta_{i}=\theta_{i}\left(\mathbf{y}_{i}\right)=Q\left(y_{i 1}, \ldots, y_{i N_{i}}\right), \tag{1.1}
\end{equation*}
$$

where $Q(\cdot)$ is a real-valued, measurable function, and $\mathbf{y}_{i}=\left(y_{i 1}, \ldots, y_{i N_{i}}\right)^{\prime}$. Common choices of $Q(\cdot)$ are the finite population mean or quantile. Molina and Rao (2010) gives several examples of the function $Q(\cdot)$. Assume $y_{i j}$ is observed only for the elements in the sample. The covariate $\mathbf{x}_{i j}$ is required for every element of the population. In this probabilistic framework, the population $U$ is partitioned into two parts as $U=A \cup R$, where $A$ is the index set of the sample and $R$ is the index set of the non-sampled elements. We partition $A$ and $R$ as $A=\cup_{i=1}^{D} A_{i}$ and $R=\cup_{i=1}^{D} R_{i}$, where $A_{i}$ is the index set of sampled elements for area $i$, and $R_{i}$ is the index set of non-sampled elements in area $i$. Without loss of generality, it can be assumed that $A_{i}=\left\{1, \ldots, n_{i}\right\}$, and $R_{i}=\left\{n_{i}+1, \ldots, N_{i}\right\}$. With this convention, $\mathbf{y}_{i}=\left(\mathbf{y}_{i s}^{\prime}, \mathbf{y}_{i r}^{\prime}\right)^{\prime}$, where $\mathbf{y}_{i s}=\left(y_{i 1}, \ldots, y_{i n_{i}}\right)^{\prime}$, and $\mathbf{y}_{i r}=\left(y_{i n_{i}+1}, \ldots, y_{i N_{i}}\right)^{\prime}$.

We compare the predictors of $\theta_{i}$ and corresponding MSE estimators based on the gamma-Poisson model to the predictors and MSE estimators based on the Poisson GLMM through simulation. We simulate data from both the Poisson GLMM and the gamma-Poisson model. For each simulation model, we calculate the predictors and MSE estimators based on both the Poisson GLMM and the gamma-Poisson model. This allows us to evaluate the properties of the procedures for situations where the model is correctly specified and under model misspecification.

We illustrate the methods using a subset of data collected in the 2018 Iowa Seat-Belt Use survey. Berg (2022) constructs predictors of county-level means using this same data set. We extend the analysis of Berg (2022) to include predictors of the median and the inter-quartile range (IQR), using the gamma-Poisson model as well as the Poisson GLMM. The data analysis is somewhat contrived to suit our interest in count data. The actual parameters of interest are proportions of belted occupants. In Berg (2023a), we conduct a more extensive analysis of the data that is geared toward the practical needs of the seat-belt use survey. That analysis motivated our interest in developing methodology for small area models for counts. The analysis
in this paper allows us to effectively illustrate the proposed methodology. The methods that we propose are of practical interest beyond the illustrative application. Count data appear frequently in the small area estimation literature (Tzavidis et al., 2015). Applications often benefit from estimates of parameters that are more general than the mean (Molina and Rao, 2010; Hobza et al., 2020). The method that we propose uniquely provides estimates of a broad range of parameters for count data.

Our primary contribution is the development of predictors of nonlinear small area parameters, such as those considered in Molina and Rao (2010), for a situation in which the response variable is a count. The models in this paper are not new, but to our knowledge, the proposed prediction algorithms are novel. Our development of predictors of nonlinear parameters builds on work in Hobza et al. (2020). The methods of Hobza et al. (2020) apply to a general GLMM specification, but their simulations and data analysis focus on the gamma response distribution. We provide detailed steps to construct predictors of nonlinear parameters for count data. Also, Hobza et al. (2020) restricts attention to additive parameters of the form $N_{i}^{-1} \sum_{j=1}^{N_{i}} q\left(y_{i j}\right)$ for a specified function $q$. The class of nonlinear parameters that we define in (1.1) is more general than the class of additive parameters discussed in Hobza et al. (2020). The class of parameters that we define is broad enough to encompass quantiles. It also includes other parameters of practical importance, such as the small area skewness and kurtotsis. The method of Hobza et al. (2020) is not immediately applicable to estimation of non-additive parameters, such as quantiles, the skewness, and the kurtosis. While the mean provides an indication of the central tendency within a small area, estimates of the quantiles and higher moments provide a more complete picture of the distribution of the characteristic at the small area level.

Our approach has two limitations which are important to assert. First, we assume that the sample design is noninformative for the specified model. If the sampling weights are correlated with the model response variable, after conditioning on model covariates then the design is informative. The methods that we propose will render biased inferences under informative sampling. In ongoing work, Berg and Eideh (2023) extend the proposed methods to a complex sample design. We refer the reader to Parker, Janicki and Holan (2019) for a comprehensive review of small area estimation under informative sampling. A second limitation is that we require the covariate for every unit in the population. In many applications, it may be difficult to satisfy this assumption. If only area-level covariates are available, then area-level models may be preferable.

The rest of this manuscript is organized as follows. In Section 2, we develop empirical best predictors of nonlinear parameters for the unit-level gamma-Poisson model. In Section 3, we develop empirical best predictors of nonlinear parameters for a GLMM with a Poisson response distribution. In Section 4, we compare the two procedures through simulation. In Section 5, we apply both procedures to the seat-belt survey data. We conclude with a discussion highlighting the strengths and weaknesses of the two models Section 6.

## 2. Unit-level gamma Poisson model and predictor

We define the unit-level gamma-Poisson model as in Berg (2022). Assume

$$
\begin{equation*}
y_{i j} \mid \mu_{i j} \stackrel{\text { ind }}{\sim} \operatorname{Poisson}\left(\mu_{i j}\right), \quad i=1, \ldots, D ; j=1, \ldots, N_{i}, \tag{2.1}
\end{equation*}
$$

where $\mu_{i j}=\lambda_{i j} u_{i}, u_{i} \sim \operatorname{ind} \operatorname{Gamma}(\alpha, \beta), \lambda_{i j}=g\left(\mathbf{x}_{i j}^{\prime} \gamma\right), \gamma=\left(\gamma_{1}, \ldots, \gamma_{p}\right)^{\prime}$ is a fixed vector of regression coefficients, and $g(\cdot)$ is a specified link function. We let $g\left(\mathbf{x}_{i j}^{\prime} \gamma\right)=\exp \left(\mathbf{x}_{i j}^{\prime} \gamma\right)$. The notation $\operatorname{Gamma}(a, b)$ denotes a gamma distribution with shape parameter $a$ and rate parameter $b$ such that $E\left[u_{i}\right]=\alpha / \beta$.

We develop an empirical best predictor of $\theta_{i}$. As in Berg (2022), the assumptions of the model (2.1) imply that the conditional distribution of $u_{i}$ given the observed data satisfies,

$$
\begin{equation*}
u_{i} \mid \mathbf{y}_{i s} \stackrel{\text { ind }}{\sim} \operatorname{Gamma}\left(y_{i .}+\alpha, \beta+\lambda_{i .}\right), \tag{2.2}
\end{equation*}
$$

where $y_{i .}=\sum_{j=1}^{n_{i}} y_{i j}$ and $\lambda_{i .}=\sum_{j=1}^{n_{i}} \lambda_{i j}$. Note that (2.2) holds exactly for any sample size and does not require approximations. The conditional distribution (2.2) is the crux of the development of the empirical best predictors. The conditional distribution (2.2) depends on the unknown $\alpha, \beta$, and $\gamma$. To operate with the conditional distribution, we use the maximum likelihood estimators of these fixed parameters. As demonstrated in $\operatorname{Berg}(2022)$, the $\log$ likelihood for $\alpha, \beta, \gamma$ is of the form $\ell(\alpha, \beta, \gamma)=\sum_{i=1}^{D} \log \left(L_{i}(\alpha, \beta, \gamma)\right)$, where

$$
L_{i}(\alpha, \beta, \gamma)=\frac{\beta^{\alpha}\left[\prod_{j=1}^{n_{i}} \lambda_{i j}^{y_{i j}}\right]}{\Gamma(\alpha) \prod_{j=1}^{n_{i}} y_{i j}!} \frac{\Gamma\left(y_{i,}+\alpha\right)}{\left(\beta+\lambda_{i}\right)^{v_{i}+\alpha}} .
$$

Define the maximum likelihood estimator by

$$
\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}^{\prime}\right)^{\prime}=\operatorname{argmax}_{(\alpha, \beta, \gamma)} \ell(\alpha, \beta, \gamma) .
$$

Berg (2022) discusses the theoretical properties of the maximum likelihood estimator for the gammaPoisson model.

The known conditional distribution for $u_{i}$ in (2.2), combined with the maximum likelihood estimator, motivates a computationally simple procedure for predicting $\theta_{i}$. The procedure is an application of the general method of Molina and Rao (2010) to the gamma-Poisson model (2.1). The best predictor of $\theta_{i}$ under squared error loss is $\tilde{\theta}_{i}=E\left[\theta_{i} \mid \alpha, \beta, \gamma, \mathbf{y}_{s}\right]$. By mutual independence of $\mathbf{y}_{1 s}, \ldots, \mathbf{y}_{D s}$, the best predictor simplifies as $\tilde{\theta}_{i}=\tilde{\theta}_{i}\left(\alpha, \beta, \gamma, \mathbf{y}_{i s}\right)=E\left[\theta_{i} \mid \alpha, \beta, \gamma, \mathbf{y}_{i s}\right]$, where

$$
\tilde{\theta}_{i}=E\left[\theta_{i} \mid \alpha, \beta, \gamma, \mathbf{y}_{i s}\right]=\sum_{y_{y_{i+1}+1}=0}^{\infty} \cdots \sum_{y_{N_{i}=}=0}^{\infty} \theta_{i}\left(\mathbf{y}_{i}\right) f\left(\mathbf{y}_{i r} \mid \mathbf{y}_{i s} ; \alpha, \beta, \gamma\right),
$$

and

$$
f\left(\mathbf{y}_{i r} \mid \mathbf{y}_{i s} ; \alpha, \beta, \gamma\right)=\left(\frac{\prod_{i=n_{i}+1}^{N_{i}} \lambda_{i j}^{y_{i j}}}{\prod_{j=n_{i}+1}^{N_{i}} y_{i j}!}\right)\left(\frac{\Gamma\left(\sum_{j=1}^{N_{i}} y_{i j}+\alpha\right)}{\Gamma\left(\sum_{j=1}^{n_{i}} y_{i j}+\alpha\right)}\right)\left(\frac{\left(\beta+\lambda_{i}\right)^{y_{i}+\alpha}}{\left(\beta+\sum_{j=1}^{N_{i}} \lambda_{i j}\right)^{\sum_{j=1}^{N_{i}} v_{i j}+\alpha}}\right) .
$$

The notation $\tilde{\theta}_{i}\left(\alpha, \beta, \gamma, \mathbf{y}_{i s}\right)$ emphasizes dependence of the best predictor on the unknown $\alpha, \beta$, and $\gamma$. An empirical best predictor is obtained by substitution of the unknown parameters defining the best predictor with the maximum likelihood estimators. That is, the empirical best predictor is defined as

$$
\begin{equation*}
\hat{\theta}_{i}=E\left[\theta_{i} \mid \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \mathbf{y}_{i s}\right]=\sum_{y_{i,+1}=0}^{\infty} \cdots \sum_{y_{i_{i}}=0}^{\infty} \theta_{i}\left(\mathbf{y}_{i}\right) f\left(\mathbf{y}_{i r} \mid \mathbf{y}_{i s} ; \hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) . \tag{2.3}
\end{equation*}
$$

The infinite sum defining the empirical best predictor is analytically intractable. We define a Monte Carlo (MC) approximation for the empirical best predictor of $\theta_{i}$, as in Molina and Rao (2010).

We define $L$ simulated populations. For $\ell=1, \ldots, L$, we set $y_{i j}^{(\ell)}=y_{i j}$ for $j=1, \ldots, n_{i}$. For non-sampled elements, $j=n_{i}+1, \ldots, N_{i}$, we generate $y_{i j}^{(\ell)}$ as $y_{i j}^{(\ell)} \mid \mu_{i j}^{(\ell)} \sim \operatorname{Poisson}\left(\mu_{i j}^{(\ell)}\right)$, where

$$
\begin{aligned}
\mu_{i j}^{(\ell)} & =\exp \left(\mathbf{x}_{i j}^{\prime} \hat{\gamma}\right) u_{i}^{(\ell)}, \\
u_{i}^{(\ell)} & \sim \operatorname{Gamma}\left(y_{i .}+\hat{\alpha}, \hat{\beta}+\hat{\lambda}_{i .}\right),
\end{aligned}
$$

and $\hat{\lambda}_{i .}=\sum_{j=1}^{n_{i}} \exp \left(\mathbf{x}_{i j}^{\prime} \hat{\gamma}\right)$. An MC approximation for the empirical best predictor of $\theta_{i}$ is then

$$
\begin{equation*}
\hat{\theta}_{i, L}=\hat{\theta}_{i, L}\left(\hat{\gamma}, \hat{\alpha}, \hat{\beta}, \mathbf{y}_{i s}\right)=L^{-1} \sum_{\ell=1}^{L} \theta_{i}^{(\ell)}, \tag{2.4}
\end{equation*}
$$

where $\theta_{i}^{(\ell)}=Q\left(y_{i 1}^{(\ell)}, \ldots, y_{i N_{i}}^{(\ell)}\right)$. We express the predictor $\hat{\theta}_{i, L}$ as a function of $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$, and $\mathbf{y}_{i s}$ to emphasize dependence of the predictor on the parameter estimators and the observed counts for the area.

### 2.1 MSE Estimation

We use the bootstrap method of González-Manteiga, Lombardía, Molina, Morales and Santamaría (2007) and Molina and Rao (2010) to estimate the MSE of $\hat{\theta}_{i, L}$. For $b=1, \ldots, B$, we repeat the following steps:

1. For $i=1, \ldots, D$, generate $\left\{y_{i 1}^{*(b)}, \ldots, y_{i N_{i}}^{*(b)}\right\}$ from the model in (2.1) with parameters equal to the maximum likelihood estimate $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})^{\prime}$. Define a bootstrap version of the population parameter by $\theta_{i}^{*(b)}=Q\left(y_{i 1}^{*(b)}, \ldots, y_{i N_{i}}^{*(b)}\right)$. Let $\mathbf{y}_{i s}^{*(b)}=\left(y_{i 1}^{*(b)}, \ldots, y_{i i_{i}}^{*(b)}\right)^{\prime}$. The bootstrap sample is then $\left\{\mathbf{y}_{i s}^{*(b)}: i=\right.$ $1, \ldots, D\}$.
2. Use the bootstrap sample, $\left\{\mathbf{y}_{i s}^{*(b)}: i=1, \ldots, D\right\}$, to obtain a maximum likelihood estimator denoted as $\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)},\left(\hat{\gamma}^{*(b)}\right)^{\prime}\right)^{\prime}$. Specifically, $\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)},\left(\hat{\gamma}^{*(b)}\right)^{\prime}\right)^{\prime}$ satisfies

$$
\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)},\left(\hat{\gamma}^{*(b)}\right)^{\prime}\right)^{\prime}=\operatorname{argmax}_{(\alpha, \beta, \gamma)} \ell^{*(b)}(\alpha, \beta, \gamma)
$$

where

$$
\ell^{*(b)}(\alpha, \beta, \gamma)=\sum_{i=1}^{D} \log \left(L_{i}^{*(b)}(\alpha, \beta, \gamma)\right),
$$

$$
L_{i}^{*(b)}(\alpha, \beta, \gamma)=\frac{\beta^{\alpha}\left[\prod_{j=1}^{n_{i}} \lambda_{i j}^{y_{i j}^{z_{j}(b)}}\right]}{\Gamma(\alpha) \prod_{j=1}^{n_{i}} y_{i j}^{*(b)}!} \frac{\Gamma\left(y_{i .}^{*(b)}+\alpha\right)}{\left(\beta+\lambda_{i .}\right)^{v_{i}^{(b)}+\alpha},}
$$

and $y_{i .}^{*(b)}=\sum_{j=1}^{n_{i}} y_{i j}^{*(b)}$.
3. Construct the MC approximation for the empirical best predictor using the bootstrap sample and the bootstrap maximum likelihood estimator. Denote the bootstrap version of the empirical best predictor by $\hat{\theta}_{i, L}^{*(b)}=\hat{\theta}_{i}\left(\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}, \hat{\gamma}^{*(b)}, \mathbf{y}_{i s}^{*(b)}\right)$. We construct $\hat{\theta}_{i, L}^{*(b)}$ as follows. We define $L$ simulated populations. For $\ell=1, \ldots, L$, we set $y_{i j}^{(\ell)}=y_{i j}^{*(b)}$ for $j=1, \ldots, n_{i}$. For non-sampled elements, $j=n_{i}+1, \ldots, N_{i}$, we generate $y_{i j}^{\left(\iota^{*} b\right)}$ as $y_{i j}^{\left(e^{*}\right)} \mid \mu_{i j}^{\left(e^{*} b\right)} \sim \operatorname{Poisson}\left(\mu_{i j}^{\left(e^{*} b\right)}\right)$, where

$$
\begin{aligned}
& \mu_{i j *}^{\left(e^{*}\right)}=\exp \left(\mathbf{x}_{i j}^{\hat{\gamma}^{*}(b)}\right) u_{i}^{\left({ }^{* *}\right)}, \\
& u_{i}^{\left(e^{*}\right)} \sim \operatorname{Gamma}\left(y_{i .}^{*(b)}+\hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}+\hat{\lambda}_{i .}^{*(b)}\right),
\end{aligned}
$$

$y_{i .}^{*(b)}=\sum_{j=1}^{n_{i}} y_{i j}^{*(b)}$, and $\hat{\lambda}_{i .}^{*(b)}=\sum_{j=1}^{n_{i}} \exp \left(\mathbf{x}_{i j}^{\prime} \hat{\gamma}^{*(b)}\right)$. We then define $\hat{\theta}_{i, L}^{*(b)}=\hat{\theta}_{i, L}\left(\hat{\gamma}^{*(b)}, \hat{\alpha}^{*(b)}, \hat{\beta}^{*(b)}\right.$, $\left.\mathbf{y}_{i s}^{*(b)}\right)=L^{-1} \sum_{\ell=1}^{L} \theta_{i}^{\left(*^{*}\right)}$, where $\theta_{i}^{\left(\left(_{i}^{*}\right)\right)}=Q\left(y_{i 1}^{\left({ }^{(*)}\right)}, \ldots, y_{i N_{i}}^{\left(e^{*}\right)}\right)$.

Define a bootstrap estimator of the MSE by

$$
\begin{equation*}
\widehat{\mathrm{MSE}}_{i}=\frac{1}{B} \sum_{b=1}^{B}\left(\hat{\theta}_{i, L}^{*(b)}-\hat{\theta}_{i}^{*(b)}\right)^{2} . \tag{2.5}
\end{equation*}
$$

## 3. Poisson GLMM

We next define an empirical best prediction procedure for a GLMM with a Poisson response distribution. The Poisson GLMM assumes that

$$
\begin{gather*}
y_{i j} \mid \mu_{i j} \stackrel{\text { ind }}{\sim} \operatorname{Poisson}\left(\mu_{i j}\right), \quad i=1, \ldots, D ; j=1, \ldots, N_{i},  \tag{3.1}\\
h\left(\mu_{i j}\right)=\beta_{0}+\mathbf{x}_{i j}^{\prime} \boldsymbol{\beta}_{1}+b_{i},
\end{gather*}
$$

and $b_{i} \stackrel{\text { iid }}{\sim} N\left(0, \sigma_{b}^{2}\right)$ for $i=1, \ldots, D$. The function $h\left(\mu_{i j}\right)$ is a specified link function. We assume that $h\left(\mu_{i j}\right)=\log \left(\mu_{i j}\right)$.

We estimate the parameters of the Poisson GLMM using the method of Schall (1991). This method is applicable to a very general GLMM. The method of Schall (1991) is used in the context of binomial data in González-Manteiga et al. (2007). The supplementary material of Berg (2022) describes the steps of the Schall (1991) procedure for the specific Poisson GLMM. Because the method has the form of an iteratively reweighted least squares (IRLS) procedure, we refer to the algorithm of Schall (1991) as the IRLS algorithm. Let $\hat{\beta}_{0}, \hat{\boldsymbol{\beta}}_{1}, \hat{b}_{i}$ and $\hat{\sigma}_{b}^{2}$ be the estimators and predictors obtained upon completion of the IRLS algorithm.

Remark: The R function $g l m e r$ is a widely used alternative to the IRLS algorithm. We emphasize the IRLS algorithm in the main document because the IRLS algorithm is reproducible in programming languages other than R. We present results using glmer in Section 2 of the supplementary material (Berg, 2023b).

### 3.1 Empirical best predictor for Poisson GLMM

The best predictor of $\theta_{i}$ under squared error loss is

$$
E\left[\theta_{i} \mid \mathbf{y}_{i s} ; \beta_{0}, \boldsymbol{\beta}_{1}, \sigma_{b}^{2}\right]=\sum_{y_{y_{i+1}+1}=0}^{\infty} \cdots \sum_{y_{N_{i}}=0}^{\infty} \theta_{i}\left(\mathbf{y}_{i}\right) f\left(\mathbf{y}_{i r} \mid \mathbf{y}_{i s}\right),
$$

where

$$
\begin{aligned}
f\left(\mathbf{y}_{i r} \mid \mathbf{y}_{i s}\right) & =\int_{-\infty}^{\infty}\left[\prod_{j=n_{i}+1}^{N_{i}} \mu_{i j}^{v_{j}} \exp \left(-\mu_{i j}\right) / y_{i j}!\right] f\left(b_{i} \mid \mathbf{y}_{i s}\right) d b_{i}, \\
f\left(b_{i} \mid \mathbf{y}_{i s}\right) & =\frac{\left[\prod_{j=1}^{n_{i}} \mu_{i j}^{v_{i j}} \exp \left(-\mu_{i j}\right) / y_{i j}!\right] \phi\left(b_{i} / \sigma_{b}\right) / \sigma_{b}}{\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n_{i}} \mu_{i j}^{v_{i j}} \exp \left(-\mu_{i j}\right) / y_{i j}!\right] \phi\left(b_{i} / \sigma_{b}\right) / \sigma_{b} d b_{i}},
\end{aligned}
$$

and $\phi$ is the pdf of a standard normal distribution. The empirical best predictor is then $E\left[\theta_{i} \mid \mathbf{y}_{i s} ; \hat{\beta}_{0}, \hat{\boldsymbol{\beta}}_{1}, \hat{\sigma}_{b}^{2}\right]$.

The conditional distribution of $b_{i}$ given the data does not have a known form for the GLMM. We therefore require a Monte Carlo procedure to approximate this conditional distribution. We use a method called sampling importance resampling (SIR) to obtain a Monte Carlo approximation for the empirical best predictor of (1.1) under the assumptions of the model (3.1). The SIR algorithm is traditionally used to sample from posterior distributions in a Bayesian context (Smith and Gelfand, 1992). We use SIR for the purpose of obtaining an MC approximation for the empirical best predictor. The SIR algorithm involves simulating from a proposal distribution and then accepting a proposed value with probability proportional to the ratio of the target and proposal distributions. The SIR algorithm is a general algorithm, and the details of implementation depend on the context. We describe how we implement the SIR algorithm for the specific Poisson GLMM in steps 1-2 below.

For $l=1, \ldots, L$, repeat the following steps:

1. Generate $b_{i}^{(\ell, 1)}, \ldots, b_{i}^{(\ell, T)} \stackrel{\text { iid }}{\sim} N\left(\hat{b}_{i}, \hat{\sigma}_{b}^{2}\right)$ for $i=1, \ldots, D$.
2. Define

$$
p_{i}^{(\ell, t)}=\exp \left(\ell_{1 i}^{(\ell, t)}-\frac{1}{T} \sum_{t=1}^{T} \ell_{1 i}^{(\ell, t)}\right)\left\{\frac{\phi\left(b_{i}^{(\ell, t)} / \hat{\sigma}_{b}\right)}{\phi\left(\left(b_{i}^{(\ell, t)}-\hat{b}_{i}\right) / \hat{\sigma}_{b}\right)}\right\},
$$

where

$$
\ell_{1 i}^{(\ell, t)}=\log \left(\left\{\prod_{j=1}^{n_{i}} \frac{\mu_{i j}^{(\ell, t)} \exp \left(-\mu_{i j}^{(\ell, t)}\right)}{y_{i j}!}\right\}\right),
$$

and $\log \left(\mu_{i j}^{(\ell, t)}\right)=\hat{\boldsymbol{\beta}}_{0}+\mathbf{x}_{i j}^{\prime} \hat{\boldsymbol{\beta}}_{1}+b_{i}^{(\ell, t)}$. Set

$$
\tilde{p}_{i}^{(e, t)}=\frac{p_{i}^{(t, t)}}{\sum_{i=1}^{T} p_{i}^{(e, t)}} .
$$

Set $b_{i}^{(t)}=b_{i}^{(t, t)}$ with probability $\tilde{p}_{i}^{(e, t)}$ for $i=1, \ldots, D$.
3. Generate $y_{i j}^{(i)} \sim \operatorname{Poisson}\left(\mu_{i j}^{(i)}\right)$, where $\log \left(\mu_{i j}^{(i)}\right)=\hat{\boldsymbol{\beta}}_{0}+\mathbf{x}_{i j}^{\prime} \hat{\boldsymbol{\beta}}_{1}+b_{i}^{(i)}$ for $i=1, \ldots, D$, and $j=n_{i}+$ $1, \ldots, N_{i}$. Set $y_{i j}^{(i)}=y_{i j}$ for $j=1, \ldots, n_{i}$.
4. Define $\theta_{i}^{() \text {()LLMM }}=Q\left(y_{i 1}^{(\ell)}, \ldots, y_{i N_{i}}^{()}\right)$.

Finally, define the predictor of $\theta_{i}$ by

$$
\begin{equation*}
\hat{\theta}_{i}^{\text {GLMM }}=\hat{\theta}_{i}^{\text {CLMM }}\left(\hat{\beta}_{0}, \hat{\boldsymbol{\beta}}_{1}, \hat{\sigma}_{b}^{2}, \mathbf{y}_{i s}\right)=L^{-1} \sum_{\ell=1}^{L} \theta_{i}^{(1) \text { (iLMM }} \text {. } \tag{3.2}
\end{equation*}
$$

### 3.2 Bootstrap MSE estimator for Poisson GLMM

We use the parametric bootstrap for MSE estimation. The bootstrap procedure is essentially that of Molina and Rao (2010) and González-Manteiga et al. (2007), applied to the Poisson GLMM. For $b=$ $1, \ldots, B$, repeat the following steps:

1. Generate a bootstrap population from the model (3.1), with parameters equal to the estimated parameters. Specifically, for $i=1, \ldots, D$, and $j=1, \ldots, N_{i}$, generate

$$
y_{i j}^{(b)} \mid \mu_{i j}^{(b)} \sim \operatorname{Poisson}\left(\mu_{i j}^{(b)}\right),
$$

where $\log \left(\mu_{i j}^{(b)}\right)=\hat{\beta}_{0}+\mathbf{x}_{i j}^{\prime} \hat{\boldsymbol{\beta}}_{1}+b_{i}^{(b)}$, and $b_{1}^{(b)}, \ldots, b_{D}^{(b)} \sim N\left(0, \hat{\sigma}_{b}^{2}\right)$. Define the bootstrap version of the population parameter by $\theta_{i}^{*(b)}=Q\left(y_{i 1}^{*(b)}, \ldots, y_{i i_{i}}^{*(b)}\right)$.
2. Let $\mathbf{y}_{i s}^{*(b)}=\left(y_{i 1}^{*(b)}, \ldots, y_{i_{i}}^{*(b)}\right)$ denote the generated values for the index set in the sample. We call $\left\{\mathbf{y}_{1 s}^{*(b)}, \ldots, \mathbf{y}_{D_{s}}^{\xi_{s}^{(b)}}\right\}$ the bootstrap sample.
3. Apply the IRLS method of Schall (1991), described above, to the bootstrap sample, $\left\{\mathbf{y}_{1 s}^{*(b)}, \ldots\right.$, $\left.\mathbf{y}_{D_{s}(b)}^{p_{s}}\right\}$, to obtain bootstrap versions of the parameter estimates. Denote the estimates obtained from the bootstrap sample by $\left(\hat{\beta}_{0}^{*(b)},\left(\hat{\boldsymbol{\beta}}_{1}^{*(b)}\right)^{\prime}, \hat{\sigma}_{b}^{*(b)}\right)^{\prime}$.
4. Implement the procedure of Section 3.1 with the bootstrap sample and the bootstrap estimates $\left(\hat{\boldsymbol{\beta}}_{0}^{*(b)},\left(\hat{\boldsymbol{\beta}}_{1}^{*(t)}\right)^{\prime}, \hat{\boldsymbol{\theta}}_{b}^{*(b)}\right)^{\prime}$ to obtain a bootstrap version of the predictor. The bootstrap version of the predictor is $\hat{\theta}_{i}^{*(b) C L M M}=\hat{\theta}_{i}^{\text {GLMM }}\left(\hat{\beta}_{0}^{*(b)}, \hat{\boldsymbol{\beta}}_{1}^{*(b)}, \hat{\sigma}_{b}^{*(b) 2}, \mathbf{y}_{i s}^{*(b)}\right)$.

Define the bootstrap MSE estimator by

$$
\begin{equation*}
\widehat{\mathrm{MSE}}_{i}^{\mathrm{GLMM}}=\frac{1}{B} \sum_{b=1}^{B}\left(\hat{\theta}_{i}^{*(b) \text { CLMM }}-\theta_{i}^{*(b)}\right)^{2} . \tag{3.3}
\end{equation*}
$$

## 4. Simulations

The simulation study has primary and secondary objectives. The primary objectives of the simulation study are two-fold. The first is to evaluate the performance of the small area predictors based on the gammaPoisson model relative to the small area predictors based on the Poisson-GLMM. The second goal is to evaluate the quality of the proposed bootstrap MSE estimators. The two secondary goals of the simulation study are (1) to evaluate the computational time of the alternative procedures, and (2) to assess the choice of $T$ for the SIR algorithm. We present output related to the primary objectives of the simulation in this main document. We relegate further discussion of the secondary objectives of the simulation to the supplementary material (Berg, 2023b). We refer the reader to Berg (2022) for a study of the properties of the estimators of the fixed parameters.

We generate data from both the gamma-Poisson model defined in (2.1) and from the unit-level Poisson GLMM defined in (3.1). For each simulation model, we calculate predictors based on the gamma-Poisson model and the GLMM. This permits an evaluation of the procedures under model misspecification. We simulate a univariate covariate as $x_{i j} \sim N(0.5,1)$ for $i=1, \ldots, D$, and $j=1, \ldots, N_{i}$, where $N_{i}=100$ for $i=1, \ldots, D$, and $D=100$. The covariate is held fixed across simulation runs. We simulate a population from either the gamma-Poisson model (2.1) or the GLMM (3.1). For the gamma-Poisson model, we set $\gamma=1$ and $\beta=2$. We use two values of $\alpha$ for the gamma-Poisson model. We first use $\alpha=5$. We then generate a more skewed distribution by setting $\alpha=0.5$. For the GLMM, we set $\beta_{0}=0.5$ and $\beta_{1}=0.5$. We use 0.5 and 1.5 as the two values for $\sigma_{b}^{2}$ for the GLMM. The combination of two model forms (gammaPoisson and GLMM) with two values for each of $\sigma_{b}^{2}$ and $\alpha$ results in a total of four data generating models. We select a simple random sample from each area with a common sampling rate of $5 \%$. The use of a constant sampling rate is fairly unrealistic but is chosen intentionally for two reasons. The first is simplicity. The second is to construct a situation with sample sizes that are small enough to reflect the challenges in real small area problems. We construct predictors of three small area parameters of interest: the area mean, the area median, and the area inter-quartile range (IQR). The area IQR is defined as the difference between the 75 -percentile and the 25 -percentile for the area.

We construct the two main proposed predictors of each parameter. We use Gam-Pois to denote the empirical best predictor for the gamma-Poisson model. The Gam-Pois predictor is defined in (2.4), where it is denoted as $\hat{\theta}_{i, L}$. We use GLMM to denote the empirical best predictor for the Poisson-GLMM. The GLMM predictor is denoted $\hat{\theta}_{i}^{\text {GLMM }}$ where it is defined in (3.2). When implementing the GLMM procedure, we use $T=200$ for the Monte Carlo SIR algorithm. The choice of $T$ is discussed in the supplementary material (Berg, 2023b). For both the GLMM and Gam-Pois predictors, we use $L=1,000$. The choice of $L$ is based on a comparison of $L=100$ to $L=1,000$. The results for $L=100$ are presented in Section 5 of the supplementary material (Berg, 2023b).

We also compute a direct estimator of each parameter. The direct estimator of the mean is the sample mean for the area. The direct estimator of the median is the sample median. The direct estimator of the IQR is calculated as the difference between the sample 75 -percentile and the sample 25 -percentile for the area.

For the GLMM, we further define plug-in (PI) predictors. The PI predictor of the mean is defined as

$$
\begin{equation*}
\hat{\bar{y}}_{i}^{\mathrm{PI}}=\frac{1}{N_{i}} \sum_{j=1}^{N_{i}} \hat{y}_{i j}^{\mathrm{PI}}, \tag{4.1}
\end{equation*}
$$

where $\hat{y}_{i j}^{\mathrm{PI}}=y_{i j}$ for $j=1, \ldots, n_{i}, \quad \hat{y}_{i j}^{\mathrm{PI}}=\exp \left(\hat{\beta}_{0}+x_{i j} \hat{\beta}_{1}+\hat{b}_{i}\right)$ for $j=n_{i}+1, \ldots, N_{i}$, and $\left(\hat{\beta}_{0}, \hat{\beta}_{1}, \hat{b}_{i}\right)$ is obtained from the IRLS algorithm used for the GLMM model (3.1). We define a PI predictor of the median as the median of $\left\{\hat{y}_{i j}^{\mathrm{PI}}: j=1, \ldots, N_{i}\right\}$. We define the PI predictor of the IQR as the difference between the 75-percentile of $\left\{\hat{y}_{i j}^{\mathrm{PI}}: j=1, \ldots, N_{i}\right\}$ and the 25-percentile of $\left\{\hat{y}_{i j}^{\mathrm{PI}}: j=1, \ldots, N_{i}\right\}$. A PI predictor of the form (4.1) is compared to the M-quantile predictor in Tzavidis et al. (2015).

For the mean, we also calculate the closed form expression for the empirical best predictor based on the gamma-Poisson model. This predictor is defined in Berg (2022). We refer to the predictor of Berg (2022) as Gam-Pois-Alt in the tables below.

The procedures for the median and the IQR require calculating percentiles of a set of numbers. Many procedures to calculate percentiles exist. We calculate all percentiles using the default method in the $R$ function quantile.

### 4.1 Comparison of efficiency of alternative predictors

We compare the predictors using two criteria. To define the criteria, let $\hat{\theta}_{i}^{(m)}$ and $\theta_{i}^{(m)}$, respectively, denote a predictor of $\theta_{i}$ and corresponding population parameter obtained in MC simulation $m$, where $m=1, \ldots, M$. The first criterion is the average relative root mean square error defined by

$$
\begin{equation*}
\% \mathrm{RRMSE}=100 \frac{1}{D} \sum_{i=1}^{D} \mathrm{RMSE}_{i}, \tag{4.2}
\end{equation*}
$$

where $\operatorname{RMSE}_{i}=\sqrt{\operatorname{MSE}_{\mathrm{MC} i}}\left[\frac{1}{M} \sum_{m=1}^{M} \theta_{i}^{(m)}\right]^{-1}$, and

$$
\begin{equation*}
\mathrm{MSE}_{\mathrm{MC} i}=M^{-1} \sum_{m=1}^{M}\left(\hat{\theta}_{i}^{(m)}-\theta_{i}^{(m)}\right)^{2} . \tag{4.3}
\end{equation*}
$$

The second criterion is the percent average absolute relative bias defined by

$$
\begin{equation*}
\% \mathrm{RB}=100 \frac{1}{D} \sum_{i=1}^{D} \mathrm{RB}_{i} \tag{4.4}
\end{equation*}
$$

where $\mathrm{RB}_{i}=M^{-1}\left|\sum_{m=1}^{M}\left(\hat{\theta}_{i}^{(m)}-\theta_{i}^{(m)}\right)\right|\left[M^{-1} \sum_{m=1}^{M} \theta_{i}^{(m)}\right]^{-1}$. We report the $\%$ RRMSE and $\% \mathrm{RB}$. We use a Monte Carlo (MC) simulation size of $M=500$.

We first simulate data from the gamma-Poisson model defined in (2.1). Table 4.1 contains the \%RRMSE and $\% \mathrm{RB}$ of the alternative predictors when the data are generated from the model (2.1). The direct estimator is very inefficient for this simulation because the area sample size is only 5 . For this configuration, the gamma-Poisson model is the true model, so it is not surprising that the Gam-Pois-Alt predictor has the smallest \%RRMSE for the mean. Likewise, the Gam-Pois predictor is the most efficient predictor for the
median and IQR. Unlike the Gam-Pois-Alt predictor, the Gam-Pois predictor is constructed from $L$ simulated samples and is therefore subject to an extra layer of Monte Carlo variability. The difference between the Gam-Pois predictor and the Gam-Pois-Alt predictor shows the effect of the Monte Carlo error from the $L$ simulated samples on the efficiency of the Gam-Pois predictor. Even though the GLMM is misspecified, the loss of efficiency from using the GLMM for the IQR and the median is much less than the loss from using the direct estimator. The PI predictor is a poor predictor of the median and the IQR. For the median and the IQR, the \%RRMSE of the PI predictor exceeds the \%RRMSE of the GLMM and Gam-Pois predictors. The bias makes an important contribution to the MSE for the PI predictor of the median and the IQR. The bias is negligible for the Gam-Pois and GLMM predictors, indicating that for these predictors, the contribution from the variance to the MSE is more important than the contribution from the bias.

Second, we simulate data from the GLMM defined in (3.1). Table 4.2 contains the \%RRMSE and \%RB when the true model is the Poisson GLMM defined in (3.1). The direct estimator is inefficient, compared to the model-based predictors. The PI predictor is efficient for the mean when $\sigma_{b}^{2}=0.5$, and the GLMM predictor has the smallest $\%$ RRMSE when $\sigma_{b}^{2}=1.5$. The GLMM predictor has smaller $\%$ RRMSE than the Gam-Pois and Gam-Pois-Alt predictors. This is not surprising because the GLMM is the true model for the simulation used to construct Table 4.2. Even though the Gam-Pois model is incorrectly specified, the loss of efficiency from using the Gam-Pois predictor is much smaller than the loss of efficiency from using the direct estimator, relative to the predictors based on the GLMM. For the GLMM and Gam-Pois predictors, the $\%$ RB is negligible compared to the $\%$ RRMSE, indicating that the contribution from the variance to the overall MSE of the predictor is more important than the contribution from the bias. The PI predictor is less efficient than the Gam-Pois predictor or the GLMM predictor for the median and for the IQR. The PI predictor has a severe bias for predicting the median and the IQR.

The comparison of predictors leads to three main conclusions. First, the relative efficiencies of the predictors depend on the data generating model. If the gamma-Poisson model is true, then the Gam-PoisAlt predictor is most efficient for the mean, and the Gam-Pois predictor is most efficient for nonlinear parameters. When the GLMM is true, the PI/GLMM predictors are most efficient. Second, the Gam-Pois and GLMM predictors appear to have reasonable efficiency, even under model mis-specification. When the Gam-Pois model is correctly specified, the loss of efficiency from incorrect use of the GLMM predictor is slight. Similarly, the ratio of the \%RRMSE of the Gam-Pois predictor to the \%RRMSE of the GLMM predictor when the GLMM is true is usually about 1.01 . The loss of efficiency from use of the Gam-Pois predictor when the GLMM is true is greater than the loss from the use of the GLMM when the Gam-Pois is true, but not by much. The third conclusion concerns the properties of the PI predictor. The PI predictor is not an estimator of an optimal predictor but nonetheless has good properties for predicting the mean in our simulations. For predicting the median and the IQR, the PI predictor has a substantial enough bias that the PI predictor is less efficient than the GLMM or Gam-Pois predictor. Given our interest in a broad range of parameters, we prefer the GLMM predictor over the PI predictor.

Table 4.1
\%RB and \%RRMSE of alternative predictors when the true model is the gamma-Poisson model and $L=\mathbf{1 , 0 0 0}$.

| $\sigma_{b}^{2}$ | Mean |  | Med. |  | IQR |  |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | \%RRMSE | \%RB | \%RRMSE | \%RB | \%RRMSE | \%RB |
| Gam-Pois | 5.000 | 17.362 | 0.667 | 20.263 | 0.798 | 19.406 | 0.764 |
| Gam-Pois-Alt | 5.000 | 17.355 | 0.670 |  |  |  |  |
| GLMM 5.000 | 17.452 | 0.711 | 20.340 | 0.877 | 19.473 | 0.782 |  |
| PI | 5.000 | 17.450 | 0.718 | 20.837 | 3.209 | 23.004 | 10.676 |
| Direct | 5.000 | 60.966 | 2.026 | 79.232 | 15.558 | 78.220 | 16.629 |
| Gam-Pois | 0.500 | 55.451 | 1.882 | 109.729 | 3.929 | 64.552 | 2.145 |
| Gam-Pois-Alt | 0.500 | 55.431 | 1.888 |  |  |  |  |
| GLMM | 0.500 | 56.584 | 2.850 | 110.644 | 4.558 | 65.023 | 2.332 |
| PI | 0.500 | 56.508 | 2.222 | 121.632 | 45.899 | 79.661 | 26.730 |
| Direct | 0.500 | 106.654 | 3.906 | 237.870 | 38.205 | 127.712 | 17.089 |

Notes: Gam-Pois = gamma-Poisson; GLMM = Generalized linear mixed model; $\mathrm{IQR}=$ Inter-quartile range; PI = Plug-in; $R B=$ Relative biases; RRMSE $=$ Relative root mean square error.

Table 4.2
$\%$ RB and \%RRMSE of alternative predictors when the true model is the Poisson-GLMM model and $L=\mathbf{1 , 0 0 0}$.

|  | $\sigma_{b}^{2}$ | Mean |  | Med. |  | IQR |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | \%RRMSE | \%RB | \%RRMSE | \%RB | \%RRMSE | \%RB |
| Gam-Pois | 0.500 | 24.571 | 0.816 | 29.014 | 1.014 | 25.128 | 0.977 |
| Gam-Pois-Alt | 0.500 | 24.555 | 0.810 |  |  |  |  |
| GLMM | 0.500 | 24.249 | 0.847 | 28.688 | 1.036 | 24.900 | 0.967 |
| PI | 0.500 | 24.239 | 0.820 | 29.306 | 4.696 | 42.980 | 34.347 |
| Direct | 0.500 | 38.820 | 1.399 | 53.578 | 4.722 | 64.172 | 24.574 |
| Gam-Pois | 1.500 | 19.995 | 0.624 | 22.728 | 0.758 | 24.107 | 0.825 |
| Gam-Pois-Alt | 1.500 | 19.971 | 0.627 |  |  |  |  |
| GLMM | 1.500 | 19.781 | 0.642 | 22.526 | 0.747 | 23.940 | 0.827 |
| PI | 1.500 | 19.788 | 0.609 | 23.140 | 3.220 | 35.215 | 23.591 |
| Direct | 1.500 | 52.486 | 1.938 | 67.632 | 4.636 | 95.293 | 24.036 |

Notes: Gam-Pois = gamma-Poisson; GLMM = Generalized linear mixed model; IQR = Inter-quartile range; PI = Plug-in; $R B=$ Relative biases; $\mathrm{RRMSE}=$ Relative root mean square error.

### 4.2 Properties of bootstrap MSE estimator

We next consider the properties of the MSE estimators for the gamma-Poisson and GLMM models. The MSE estimator is defined in (2.5) for the gamma-Poisson model and in (3.3) for the GLMM. We calculate both MSE estimates under each data generating model. This allows us to evaluate the properties of the MSE estimates under model misspecification and when the model is correctly specified. The bootstrap sample size is $B=200$. The choice of $B=200$ follows from a recommendation in Hobza et al. (2020) that the bootstrap sample size be at least 200 . To reduce the computational requirements, we use $L=100$ for the simulations in this section. To evaluate the MSE estimators, we conduct a simulation with $M=250$ simulated samples using the same $x_{i j}$ used for the first simulation. We calculate the relative bias of the MSE estimator as well as empirical coverages of normal theory $95 \%$ prediction intervals.

We first define the relative bias of the MSE estimator. Let $\widehat{\mathrm{MSE}}_{i}^{(m)}$ denote an MSE estimate from simulated sample $m$ for $m=1, \ldots, M$, where $M=250$. We then define the $\%$ relative bias of the MSE estimator for area $i$ by

$$
\mathrm{RB}_{i}=100\left(\frac{\frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{MSE}}_{i}^{(m)}}{\mathrm{MSE}_{\mathrm{MC} i}}-1\right)
$$

where $\mathrm{MSE}_{\text {мСi }}$ is defined in (4.3) and is based a separate simulation with $M=5,000$. We use the output from a separate simulation with $M=5,000$ to define the denominator of the $\mathrm{RB}_{i}$ to reduce the variance of $\mathrm{RB}_{i}$.

Figure 4.1 contains box-plots of the relative biases for the four simulation configurations and the three parameters. The relative biases of the bootstrap MSE estimator (depicted in Figure 4.1) depend on the simulation model and the parameters. We first consider the Gam-Pois model with $\alpha=5$. For this configuration, the Gam-Pois MSE estimator has a slight negative bias, but the relative bias is usually between $-10 \%$ and $10 \%$. The GLMM MSE estimator is nearly unbiased for the MSE of the GLMM predictor in the sense that the relative biases for the GLMM MSE estimator are symmetric around zero and are usually between $-10 \%$ and $10 \%$. It is interesting that the GLMM MSE estimator performs well for this simulation configuration because the GLMM model is incorrectly specified. We next consider the Gam-Pois model with $\alpha=0.5$. The Gam-Pois MSE estimator remains nearly unbiased, in the sense that the median relative bias is close to zero and the relative biases are usually below $10 \%$ in absolute value. The GLMM MSE estimator for the GLMM predictor tends to have a positive bias under the Gam-Pois model with $\alpha=0.5$. We next consider the GLMM simulation model with $\sigma_{b}^{2}=0.5$. The GLMM MSE estimator for the GLMM predictor is nearly unbiased, with most relative biases between $-10 \%$ and $10 \%$. The Gam-Pois MSE estimator tends to underestimate the MSE of the Gam-Pois predictor for the GLMM simulation configuration with $\sigma_{b}^{2}=0.5$. The relative bias of the Gam-Pois MSE estimator is only about $-5 \%$ for this configuration. Increasing $\sigma_{b}^{2}$ to 1.5 has little impact on the properties of the GLMM MSE estimator for the GLMM predictors. For the GLMM simulation model with $\sigma_{b}^{2}=1.5$, the Gam-Pois MSE estimator can exhibit extreme values.

We define the relative root mean square error of the MSE estimator as

$$
\operatorname{RRMSEMSE}_{i}=\frac{\sqrt{M^{-1} \sum_{m=1}^{M}\left(\widehat{\mathrm{MSE}}_{i}^{(m)}-\mathrm{MSE}_{\mathrm{MC} i}\right)^{2}}}{\operatorname{MSE}_{\mathrm{MC} i}} .
$$

Figure 4.2 contains boxplots of the RRMSEMSE ${ }_{i}$. The two MSE estimators have consistently similar relative root mean square errors. The biases observed for certain simulation configurations seem to have a negligible effect on the MSE of the MSE estimator.

Figure 4.1 Relative biases ( $\mathrm{RB}_{i}$ ) of MSE estimators for four simulation configurations.


Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range; MSE = Mean square error.

Figure 4.2 Boxplots of relative root mean square errors of mean square error estimators (RRMSEMSE $i$ ) for four simulation configurations.


Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range.

We define the empirical coverage of prediction intervals for area $i$ by

$$
\mathrm{CR}_{i}=\frac{1}{M} \sum_{m=1}^{M} I\left\{\theta_{i}^{(m)} \in \mathrm{CI}_{i}^{(m)}\right\},
$$

where $\mathrm{CI}_{i}^{(m)}=\left[\hat{\theta}_{i}^{(m)}-1.96 \sqrt{\widehat{\mathrm{MSE}}_{i}^{(m)}}, \hat{\theta}_{i}^{(m)}+1.96 \sqrt{\widehat{\mathrm{MSE}}_{i}^{(m)}}\right]$. The empirical coverages of the nominal $95 \%$ prediction intervals are depicted in Figure 4.3. A surprising result is that the GLMM procedure tends to produce superior coverage rates than the Gam-Pois procedure under the Gam-Pois configuration with $\alpha=$ 0.5 . Generally, the departures of the coverage rates from the nominal level are not severe. The empirical coverages of prediction intervals tend to fall between $92 \%$ and $98 \%$.

In summary, the gamma-Poisson MSE estimator has reasonable properties when the gamma-Poisson model is the true model, and likewise, the GLMM MSE estimator has reasonable properties when the GLMM model is the true model. The properties of the MSE estimator under model misspecification depend on the parameter configuration. The GLMM MSE estimator is approximately unbiased under the Gam-Pois configuration when $\alpha=5$ but has positive bias when $\alpha=0.5$. The Gam-Pois MSE estimator tends to have a median relative bias of about $-5 \%$ under the GLMM configuration when $\sigma_{b}^{2}=0.5$ and can be erratic when $\sigma_{b}^{2}=1.5$.

Figure 4.3 Empirical coverages of nominal $\mathbf{9 5 \%}$ prediction intervals ( $\mathbf{C R}_{\boldsymbol{i}}$ ) for four simulation configurations.


Notes: GLMM = Generalized linear mixed model; GP = gamma-Poisson; IQR = Inter-quartile range.

## 5. Illustration with modeling observed vehicle occupants

We apply both the GLMM predictor and the Gam-Pois predictor to data from the 2018 Iowa Seat-Belt use survey. This is the same data set used in Berg (2022). While Berg (2022) only constructs predictors of means, we construct predictors of more general small area parameters.

The population consists of $N=65,313$ road segments in Iowa. The road segments are nested in $D=15$ counties that define the small areas. The area sample sizes are $n_{i}=5$ road segments for all but one county in which $n_{i}=14$ such that the total sample size is $n=84$. Due to the small county sample sizes, this is clearly a small area estimation problem. Each road segment in the sample is observed for 45 minutes, and the response variable is defined as
$y_{i j}=$ number of vehicle occupants observed during the 45-minute period on road segment $j$ of county $i$.

Two covariates are available for every road segment in the population from the sampling frame. The first is the road type of the road segment, where the three road types are primary, secondary, and local. The second is the vehicle miles traveled. As in Berg (2022), we define the model to contain indicators for road type as well as interactions between road type and VMT. We refer to Berg (2022) for estimates of the fixed model parameters.

For this analysis, our objective is to compare the two models. In practice, however, an analyst may need to select one of the two models. We recommend diagnosing the goodness of fit of the two models using standardized residuals. We define a residual for the Gam-Pois model by

$$
\frac{y_{i j}-\hat{\lambda}_{i j} \hat{u}_{i}}{\sqrt{\hat{\lambda}_{i j} \hat{u}_{i}}}
$$

where

$$
\hat{u}_{i}=\frac{y_{i .}+\hat{\alpha}}{\hat{\beta}+\hat{\lambda}} .
$$

Note that $\hat{u}_{i}$ is an estimate of $E\left[u_{i} \mid \mathbf{y}_{i s}\right]$ (Berg, 2022). The standardized residual for the GLMM is defined as

$$
\frac{y_{i j}-\hat{y}_{i j}^{\mathrm{PI}}}{\sqrt{\hat{y}_{i j}^{\mathrm{PI}}}} .
$$

Figure 5.1 contains plots of the standardized residuals against the predicted values for the Gam-Pois and GLMM models. The residuals for the two models are strikingly similar. The standardized residuals do not exhibit systematic trends, and the variance of the residuals remains constant as the mean increases. The residuals clearly do not have a standard normal distribution; however, normality is not one of the model
assumptions. It is possible that the residuals indicate that the data are over-dispersed relative to the specified Poisson models. Incorporating over-dispersion in the proposed framework is a possible direction for future research.

Figure 5.1 Standardized residuals against predicted values for gamma-Poisson (gam-Pois) and generalized linear mixed models (GLMM).


We consider three county-level parameters of interest: the mean number of occupants per road segment, the median number of occupants, and the interquartile range of the number of occupants per road segment. We apply the methods of Sections 2 and 3 to obtain predictors and associated estimates of the mean square error. Because we are interested in parameters other than the mean, we use the predictor (2.4) for the gammaPoisson model. We use the predictor (3.2) for the GLMM. We report the predictors and the coefficients of variation for each procedure. The coefficient of variation is the ratio of the square root of the estimated mean square error to the predictor. The MSE estimators are defined in (2.5) and (3.3), respectively, for the gammaPoisson and GLMM models. We use a bootstrap sample size of $B=200$. Table 5.1 contains the county level predictors and corresponding coefficients of variation based on the gamma-Poisson and GLMM models.

The two models produce consistently similar predictors. The predictors based on the gamma-Poisson model are nearly the same as the predictors based on the GLMM model. This empirical result is consistent with the finding in the simulation that both predictors tend to perform well, regardless of the true data generating model. For the data analysis, we do not know the "true" data generating model. Therefore, it is reassuring that the predictors based on the two models are similar.

Table 5.1
County predictors (pred) and coefficients of variation (cv) based on gamma-Poisson (Gam-Pois) and GLMM models.

| County | Mean |  |  |  | Median |  |  |  | IQR |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Gam-Pois |  | GLMM |  | Gam-Pois |  | GLMM |  | Gam-Pois |  | GLMM |  |
|  | pred | cV | pred | cv | pred | cV | pred | cv | pred | cv | pred | cv |
| 1 | 60.37 | 0.46 | 59.99 | 0.49 | 34.71 | 1.38 | 34.27 | 1.46 | 104.8 | 0.26 | 103.56 | 0.28 |
| 2 | 60.29 | 0.04 | 60.18 | 0.04 | 26.31 | 0.05 | 26.54 | 0.10 | 57.86 | 0.14 | 58.40 | 0.14 |
| 3 | 62.97 | 0.1 | 63.01 | 0.09 | 24.54 | 0.11 | 24.53 | 0.08 | 63.58 | 0.55 | 63.70 | 0.56 |
| 4 | 87.82 | 0.45 | 87.60 | 0.45 | 46.34 | 1.25 | 46.31 | 1.26 | 163.53 | 0.3 | 163.52 | 0.31 |
| 5 | 54.67 | 0.24 | 54.91 | 0.23 | 24.5 | 1.05 | 24.39 | 1.03 | 60.26 | 0.76 | 59.94 | 0.76 |
| 6 | 77.08 | 0.17 | 77.47 | 0.17 | 31.23 | 0.99 | 31.50 | 0.97 | 136.85 | 0.07 | 137.57 | 0.08 |
| 7 | 65.79 | 0.15 | 66.03 | 0.13 | 30.27 | 0.34 | 30.20 | 0.30 | 67.68 | 0.63 | 67.32 | 0.61 |
| 8 | 99.65 | 0.02 | 99.34 | 0.03 | 41.53 | 0.05 | 41.56 | 0.05 | 118.65 | 0.21 | 118.05 | 0.23 |
| 9 | 68.12 | 0.17 | 68.19 | 0.16 | 25.98 | 0.2 | 26.07 | 0.17 | 73.11 | 0.36 | 73.14 | 0.37 |
| 10 | 97.24 | 0.27 | 97.29 | 0.27 | 41.26 | 0.42 | 41.40 | 0.40 | 153.54 | 0.36 | 154.04 | 0.38 |
| 11 | 96.99 | 0.06 | 96.83 | 0.05 | 41.9 | 0.09 | 42.02 | 0.07 | 91 | 0.18 | 91.07 | 0.18 |
| 12 | 55.73 | 0.05 | 56.14 | 0.05 | 24.39 | 0.3 | 24.47 | 0.26 | 97.05 | 0.1 | 97.15 | 0.10 |
| 13 | 53.43 | 0.04 | 53.09 | 0.05 | 18.36 | 0.07 | 18.19 | 0.13 | 53.52 | 0.22 | 53.12 | 0.23 |
| 14 | 122.55 | 0.16 | 123.69 | 0.15 | 47.28 | 0.17 | 47.78 | 0.15 | 131.63 | 0.38 | 132.73 | 0.39 |
| 15 | 96.98 | 0.09 | 96.89 | 0.09 | 48.34 | 0.52 | 48.28 | 0.53 | 91.33 | 0.41 | 91.40 | 0.42 |

Notes: GLMM = Generalized linear mixed model; IQR = Inter-quartile range.

The coefficients of variation based on the two models are also strikingly similar. This also reflects the results of the simulation in that the two procedures tend to produce reasonable means square error estimates, even under model misspecification. The coefficients of variation are not uniformly below $20 \%$, a common threshold for determining an acceptable level of precision. Several of the coefficients of variation for the median and interquartile range exceed $30 \%$. The coefficients of variation can exceed $100 \%$ for the median. For both the gamma-Poisson model and the GLMM, the effect of the variance due to estimating the fixed parameters on the mean square error of the small area predictors is substantial. This data set only has 15 counties. This leads to substantial variation in $\hat{\beta}_{0}, \hat{\sigma}_{b}^{2}, \hat{\alpha}$, and $\hat{\beta}$.

## 6. Discussion

We develop predictors of nonlinear parameters based on two unit level models for count data. We first define procedures for a gamma-Poisson model with unit-level covariates. We compare the gamma-Poisson model to a standard generalized linear mixed model. We use standard parametric bootstrap procedures for both models.

A limitation of the bootstrap procedure that we employ is that the bootstrap MSE estimator is not secondorder unbiased. One can use the double bootstrap to construct a bias-corrected MSE estimator (Hall and Maiti, 2006a,b; Erciulescu and Fuller, 2014). We do not pursue the double bootstrap in this work because the proposed MSE estimator has adequate properties for the simulation configurations that we considered.

We study the empirical properties of the small area predictors when the model is correctly specified and under model misspecification. The main conclusion from the simulations is that the empirical best predictor
for the gamma-Poisson model is superior when the gamma-Poisson model is true, and the PI and GLMM predictors are superior when the GLMM is true, as expected. This illustrates the importance of validating the model assumptions in model-based small area estimation. However, an interesting result is that the loss of efficiency from using the GLMM or gamma-Poisson predictor when the other model is true is consistently small. The MSE estimators exhibit more bias when the model is incorrectly specified than when the model is correctly specified. Nonetheless, the relative biases of the MSE estimators are usually between $-20 \%$ and $20 \%$, regardless of which model is the true model. The coverage rates of confidence intervals do not exhibit severe and systematic over-coverage or under-coverage, even under model misspecification. The PI predictor has reasonable properties for the mean but is inefficient for nonlinear parameters.

The data analysis re-affirms the results of the simulation study. The gamma-Poisson and GLMM procedures lead to similar predictors in the data analysis. This result echoes an empirical finding of Clayton and Kaldor (1987) that estimates of lung cancer rates based on gamma-Poisson and Poisson-lognormal models are similar.

The gamma-Poisson model has two main strengths relative to the Poisson GLMM. An important benefit of the gamma-Poisson procedure is computational simplicity. Maximum likelihood estimation is easy for the gamma-Poisson model because calculation of the marginal likelihood does not require numerical integration. We do not use maximum likelihood estimation for the GLMM model. Instead, we use an IRLS algorithm (Schall, 1991) for computational simplicity. A formal comparison of maximum likelihood estimation to the IRLS algorithm for the purpose of small area estimation of count data is a possible future
 material (Berg, 2023b). Given estimates of the fixed parameters, the predictors are easier to calculate for the gamma-Poisson model than for the Poisson GLMM. The computing time to implement the gammaPoisson predictor is roughly half the time required to implement the GLMM predictor (see Section 4 of the supplementary material (Berg, 2023b) for further detail). A second strength of the gamma-Poisson procedure is that the estimators of fixed model parameters are consistent. In contrast, the IRLS estimators for the Poisson GLMM are known to be inconsistent. The bootstrap procedure for the GLMM relies on consistency of the model parameter estimators, an assumption that does not hold for our estimation procedure. The procedures for the GLMM model require approximations, and we evaluate the validity of these approximations through simulation.

The GLMM exhibits different strengths, relative to the gamma-Poisson procedure. The loss of efficiency from incorrect use of the GLMM is slightly below the loss from incorrect use of the Gam-Pois model. The coverage rates of confidence intervals for the GLMM predictor are somewhat closer to $95 \%$ than the coverage rates for the Gam-Pois model for certain parameters and configurations.

In practice, the analyst may need to choose one of the two models. We propose to use residuals to diagnose the goodness of fit of the models. Our experience is that it can be very difficult to distinguish between the two models. The two models differ only with respect to the distribution of the random effect and generate very similar types of data. Fortunately, the analysis in this paper suggests that the proposed
methods are fairly robust to specification of the incorrect distributional form. The methods for the gammaPoisson model work well when the Poisson GLMM is true and vice versa. Our analysis suggests that both the gamma-Poisson and Poisson GLMM will lead to similar results for many parameter configurations. We expect the conclusions for a specific application to be fairly insensitive to the choice of model. We encourage the analyst to consider the strengths and weaknesses of the two models for particular applications when selecting one of them in practice. We also note that an alternative to model selection is model averaging (Aitkin, Liu and Chadwick, 2009). An investigation of model averaging for the two models proposed here is a possible direction for future research.

Based on this analysis, we prefer the gamma-Poisson model for computational simplicity. The conjugate form of the model makes the predictor and MSE estimator straightforward to calculate. If one is only interested in the mean, then we recommend the closed-form predictor and MSE estimator of Berg (2022). For prediction of general parameters, we recommend the simulation-based predictor defined as $\hat{\theta}_{i, L}$ in (2.4). The gamma-Poisson predictor is straightforward to implement and has acceptable efficiency, even if the GLMM true.

## Acknowledgements

This research was supported by USDA NRCS Cooperative Agreement, 68-7482-17-009.

## References

Aitkin, M., Liu, C.C. and Chadwick, T. (2009). Bayesian model comparison and model averaging for smallarea estimation. Ann. Appl. Stat., 3(1), 199-221.

Berg, E. (2022). Empirical best prediction of small area means based on a unit-level gamma-poisson model. Journal of Survey Statistics and Methodology, 1-22.

Berg, E. (2023a). Small area prediction of seat-belt use rates using a bayesian hierarchical unit-level poisson model with multivariate random effects. Stat, 12(1), e544.

Berg, E. (2023b). Supplementary material for "Prediction of general small area parameters for unit-level count data". [Online]. Available from: https://github.com/emilyjb/Count-Data-Supplement.

Berg, E., and Eideh, A. (2023). Small area inference for exponential dispersion families under informative sampling. Manuscript under review by Journal of Survey Statistics and Methodology.

Boubeta, M., Lombardía, M.-J. and Morales, D. (2016). Empirical best prediction under area-level poisson mixed models. Test, 25(3), 548-569.

Clayton, D., and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics, 671-681.

Erciulescu, A.L., and Fuller, W.A. (2014). Parametric bootstrap procedures for small area prediction variance. Proceedings of the Survey Research Methods Section, American Statistical Association Washington, DC, 3307-3318.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. Statistics in Transition, New Series, 21(4), 1-22.

González-Manteiga, W., Lombardía, M.-J., Molina, I., Morales, D. and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Computational Statistics \& Data Analysis, 51(5), 2720-2733.

Guadarrama, M., Molina, I. and Rao, J.N.K. (2018). Small area estimation of general parameters under complex sampling designs. Computational Statistics \& Data Analysis, 121, 20-40.

Hall, P., and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. The Annals of Statistics, 34, 4, 1733-1750.

Hall, P., and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 221-238.

Hobza, T., Marhuenda, Y. and Morales, D. (2020). Small area estimation of additive parameters under unitlevel generalized linear mixed models. SORT-Statistics and Operations Research Transactions, 3-38.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. Test, 15(1), 1-96.

Marhuenda, Y., Molina, I., Morales, D. and Rao, J.N.K. (2017). Poverty mapping in small areas under a twofold nested error regression model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(4), 1111-1136.

Molina, I., Corral, P. and Nguyen, M. (2022). Estimation of poverty and inequality in small areas: Review and discussion. TEST, 1-24.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. The Annals of Applied Statistics, 8(2), 852-885.

Molina, I., and Rao, J.N.K. (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), 369-385.

Parker, P.A., Janicki, R. and Holan, S.H. (2019). Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions. arXiv preprint arXiv:1908.10488.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science, 28(1), 40-68.

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

Reluga, K., Lombardía, M.-J. and Sperlich, S. (2021). Simultaneous inference for empirical best predictors with a poverty study in small areas. Journal of the American Statistical Association, (accepted), 1-33.

Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2020). Data-driven transformations in small area estimation. Journal of the Royal Statistical Society: Series A (Statistics in Society), 183(1), 121-148.

Schall, R. (1991). Estimation in generalized linear models with random effects. Biometrika, 78(4), 719-727.

Smith, A.F., and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. The American Statistician, 46(2), 84-88.

Tzavidis, N., Ranalli, M.G., Salvati, N., Dreassi, E. and Chambers, R. (2015). Robust small area prediction for counts. Statistical Methods in Medical Research, 24(3), 373-395.

# A method for estimating the effect of classification errors on statistics for two domains 

Yanzhe Li, Sander Scholtus and Arnout van Delden ${ }^{1}$


#### Abstract

Being able to quantify the accuracy (bias, variance) of published output is crucial in official statistics. Output in official statistics is nearly always divided into subpopulations according to some classification variable, such as mean income by categories of educational level. Such output is also referred to as domain statistics. In the current paper, we limit ourselves to binary classification variables. In practice, misclassifications occur and these contribute to the bias and variance of domain statistics. Existing analytical and numerical methods to estimate this effect have two disadvantages. The first disadvantage is that they require that the misclassification probabilities are known beforehand and the second is that the bias and variance estimates are biased themselves. In the current paper we present a new method, a Gaussian mixture model estimated by an ExpectationMaximisation (EM) algorithm combined with a bootstrap, referred to as the EM bootstrap method. This new method does not require that the misclassification probabilities are known beforehand, although it is more efficient when a small audit sample is used that yields a starting value for the misclassification probabilities in the EM algorithm. We compared the performance of the new method with currently available numerical methods: the bootstrap method and the SIMEX method. Previous research has shown that for non-linear parameters the bootstrap outperforms the analytical expressions. For nearly all conditions tested, the bias and variance estimates that are obtained by the EM bootstrap method are closer to their true values than those obtained by the bootstrap and SIMEX methods. We end this paper by discussing the results and possible future extensions of the method.


Key Words: Bias; Variance; Misclassification; Binary classifier; Gaussian mixture model; EM algorithm.

## 1. Introduction

Accurate published output is crucial, especially in official statistics (Eurostat, 2009, page 32), where most outcomes are used for policy making. One of the important errors affecting the accuracy of output are measurement errors in the variable that is used to group output into subpopulations. The simplest type of estimator for which the effect of misclassifications has been studied are proportions in contingency tables. Bross (1954) gave an expression for the bias of an estimated proportion of a binary variable in the case of misclassifications, which is a special case of expression (C.1) in Appendix C. A more complicated situation concerns the accuracy of level estimators (totals, means) and ratios thereof as affected by misclassifications. For instance, one may be interested in the average risk of poverty for working persons by categories of educational level (Eurostat, 2022), while part of those education values are misclassified. Estimating population parameters of a continuous variable in each class is referred to as domain statistics. The current study focuses on estimation of the bias and variance of totals or means as affected by misclassifications. There is a large amount of literature on misclassifications, for instance Buonaccorsi (2010), Keogh, Shaw, Gustafson, Carroll, Deffner, Dodd, Küchenhoff, Tooze, Wallace, Kipnis and Freedman (2020) and Shaw, Gustafson, Carroll, Deffner, Dodd, Keogh, Kipnis, Tooze, Wallace, Küchenhoff and Freedman (2020). Van den Hout and Van der Heijden (2002) provide an extensive overview of literature on bias and variance

[^10]in the case of misclassifications. These references underline that quantifying the effect of misclassifications on population estimates is relevant for a wide range of disciplines, including medicine, epidemiology, statistical astronomy, sociology, land cover mapping, randomised response studies and data confidentiality. The latter two cases are special in the sense that their misclassifications are deliberately added to the data by a known mechanism.

Official statistics often make use of statistical registers to identify their target population and divide it into subpopulations. One example is a statistical business register containing a list of statistical units over time, with background variables such as economic activity and size class to stratify them into subpopulations (United Nations, 2015). Another example is a population register, with background variables such as date of birth, gender, place of residence and highest attained level of education (Bakker, Van Rooijen and Van Toor, 2014). These registers are often created with one or more administrative sources. Errors in the classification variables can occur due to errors during registration into those administrative sources, administrative delays, errors in the linkage of the administrative sources or changes in circumstances of the units which are not reported. Another source of error are differences between concepts used by register owners and those used in official statistics (Magnusson, Palm, Branden and Mörner, 2017). Sometimes classification variables are derived by applying machine learning algorithms [see, e.g., Meertens, Diks, Van den Herik and Takes (2020)] and errors made by those algorithms subsequently lead to misclassifications.

In practical situations where output is produced for official statistics, one tries to reduce the number of misclassifications for instance by automatic or manual editing. In the case of business statistics, economic activity codes are known to be prone to errors. In that case, manual editing is applied which is usually limited to the most influential units - the largest enterprises - while the remaining units are left uncorrected. One often hopes that errors in the smaller units do not have a large impact on the published figures. Unfortunately, this need not be true. Van Delden, Scholtus and Burger (2016) showed that in a particular case study, observed levels of misclassification in smaller and medium-sized enterprises resulted in considerable bias of turnover totals for some publication cells. The reason for this bias was that the inflow of turnover from units unjustly classified into the target class was not balanced by the outflow of turnover from units unjustly not classified into the target class. In the context of contingency tables, Schwartz (1985) underlined the importance of misclassifications on output accuracy by framing it as "a neglected problem".

Analytical expressions for the bias and variance of means or totals of subpopulations as affected by misclassifications have been published by Selén (1986) and Van Delden et al. (2016). Furthermore, the bias and variance of totals of subpopulations were discussed by Kooiman, Willenborg and Gouweleeuw (1997) in the context of data confidentiality. These analytical expressions have two disadvantages. The first disadvantage is that they rely on the assumption that the probabilities of all types of misclassification are known, or that they have been accurately estimated by means of a sample. In the context of data confidentiality these probabilities are known, because misclassifications are applied on purpose to avoid disclosure (Kooiman et al., 1997). In most applications, however, the probabilities of misclassification are
unknown. Estimated classification error probabilities are often obtained from previous knowledge or a comparable dataset (Edwards, Bakoyannis, Yiannoutsos, Mburu and Cole, 2019; Edwards, Cole and Fox, 2020), or from an audit sample (Gravel and Platt, 2018). Accurately estimating these probabilities using sufficiently large sample sizes requires a considerable amount of manual labour. The second disadvantage is that the bias and variance estimates based on those analytical expressions are biased themselves. This is mentioned by Kooiman et al. (1997) and worked out in more detail by Van Delden et al. (2016). The main reason is that the expressions for the bias and variance depend on true population totals which are unknown. When estimating the bias and variance, these true population totals are replaced by their biased estimators of the population totals, leading to biased estimates of the bias and (to a lesser extent) of the variance of the totals. In the special case that one is only interested to estimate a proportion without a numerical variable, different methods to correct for bias can be found in Kloos, Meertens, Scholtus and Karch (2021).

As an alternative to the use of analytical expressions, numerical approaches have been developed. The bootstrap approach may be used to estimate output accuracy in the case of misclassifications. Zhang (2011) used the bootstrap method to measure the variance of observed totals of subpopulations caused by misclassification of households. Van Delden et al. (2016) presented a bootstrap approach to estimate the bias and variance of means or totals of subpopulations. The bootstrap approach is very flexible and can be adapted to many estimators. Unfortunately, the bootstrap approach has the same two disadvantages as the analytical approach: it requires that the probabilities of misclassification are accurately known and, in general, it leads to biased estimates of bias and variance. In fact, for simple domain parameters such as totals and proportions, the bootstrap method and the analytical expressions lead to near-identical results; see for instance Van Delden et al. (2016). For non-linear domain parameters such as ratios, the bootstrap results are generally more accurate in the case of skewed distributions because some underlying assumptions of the approximation used in the analytical bias and variance expressions might be violated; see Van Delden, Scholtus, Burger and Meertens (2023).

Another numerical method that has been proposed for misclassifications is the SIMEX ("SIMulation and Extrapolation") method, which was first introduced by Cook and Stefanski (1994). It was developed to estimate the impact of measurement errors and is adapted in the field of misclassification by Küchenhoff, Mwalili and Lesaffre (2006) and Hopkins and King (2010). Similarly to the bootstrap approach, the SIMEX method starts with applying misclassifications to the observed data after which the estimators (totals, means) are recalculated. Additionally, the SIMEX method applies a simulation process that introduces multiple sets of extra errors to estimate the effect of misclassifications. Finally, one extrapolates the estimates to the errorfree condition. Here we have applied the ideas behind this SIMEX procedure to test whether it can be used to overcome the bias in the estimates of the bias and variance of means and totals by misclassifications

In the current study, we present a new method that uses a mixture model to estimate the bias and variance of means and totals by misclassifications. An Expectation-Maximisation (EM) algorithm is used to estimate the mixture model. Since the true classes of units are unknown in the case of misclassifications, these true classes can be regarded as an unobserved ("latent") variable in the mixture model. We model the numerical
target variable in each class as a mixture of different normal distributions (a Gaussian mixture), which can also accommodate target variables that do not have a normal distribution; see McLachlan and Peel (2000). The method is referred to as "the EM bootstrap method". In the current study we limit ourselves to a binary classification variable. Our approach can be extended to a situation with multiple classes, which is further treated in the discussion.

The new method performs better with respect to the two mentioned disadvantages than the bootstrap and the SIMEX method. First of all, it does not require an estimate of the misclassification probabilities beforehand, although in more complicated cases it is computationally efficient to have a rough estimate of the misclassification rates since it can be used as an informed start for the EM algorithm. Second, we will show that - at least in the binary case - the new method leads to more accurate estimates of the bias and variance of means and totals affected by misclassifications than the bootstrap and the SIMEX method.

In the current paper we evaluate the accuracy of bias and variance estimates under misclassification, by comparing the proposed EM bootstrap method with the bootstrap method and with the adapted SIMEX method (referred to as "the SIMEX bootstrap method") in a simulation study and in a case study. In both studies, the proportions of two classes are varied as well as the misclassification rates. In the simulation study the data distribution in each class is a Gaussian mixture. The case study uses empirical distributions from a dataset with log turnover per enterprise in the Netherlands. For the empirical data we use a Gaussian mixture model in which the optimal number of components has to be estimated. In both studies we compare true bias and variance values with their estimates.

The remainder of this paper is organised as follows. Section 2 describes the details of the three methods. The evaluation of the methods for the simulation study is given in Section 3 and for the case study in Section 4. Section 5 discusses the results of our study and proposes future directions. Furthermore, Appendices A, B and C provide additional material. Appendix A discusses three different extrapolation functions in the SIMEX method. Appendix B compares the difference of using BIC and sBIC as a criterion for selecting the optimal number of components in the case study. Appendix C includes some theoretical properties of the bootstrap and the EM bootstrap method. R code that implements the methods used in this study can be found at https://github.com/Yanzhee/EM-bootstrapping.

## 2. Methodology

### 2.1 General settings

We consider a situation where a population of $N$ units is classified into two classes. The indicator variable for the true membership of one of these classes is denoted as $z \in\{0,1\}$. For convenience, from now on we will refer to the classes of interest as class 1 and class 0 and use the indicator $z$ interchangeably with the classification itself. The values of $z$ are assumed to contain no classification errors and are considered to be fixed for the population of interest. In practice, we do not know the true classes of all units. We can
only infer them through, for example, manual checking by experts, automatic classification from multiple machine-learning algorithms, etc. The true proportion of class 1 in the population is denoted by $\alpha_{1}$; the true proportion of class 0 is $1-\alpha_{1}$.

The observed classification variable is denoted as $\hat{z} \in\{0,1\}$. In practice, the observed classes of units contain classification errors. The classification errors can be from misunderstandings of class definition, miscommunication or simply typos.

In practice, one is often interested in estimating population parameters of a continuous variable $y$ in each class, referred to as domain statistics. For example, when $y$ represents turnover of enterprises in various industries, the total turnover of each industry will be an interesting indicator. We use $\zeta$ to denote a true population parameter, based on $y$ and the true classification variable $z$. In what follows, the continuous variable $y$ is assumed to be error-free.

Examples of common domain parameters $\zeta$ that are included in our study are: the total sum of $y$ for class $1\left(T_{1}\right)$, the proportion of class $1\left(\alpha_{1}\right)$, and the standard deviation of $y$ for class $1\left(\sigma_{1}\right)$. Table 2.1 lists the formulas for $\zeta$ given variables $y$ and $z$. Since our study is under the setting of a binary classifier, the accuracy of domain statistics in class 0 is directly related to the accuracy of the corresponding estimates in class 1.

Table 2.1
List of formulas for examples of domain parameters $\zeta$.

| Statistics of Interest | Notation | Formula given $\boldsymbol{y}$ and $z$ |
| :--- | :---: | :---: |
| Total sum of $y$ for class 1 | $T_{1}$ | $\sum_{i} z_{i} y_{i}$ |
| Proportion of class 1 | $\alpha_{1}$ | $\sum_{i} z_{i} / N$ |
| Standard deviation of $y$ for class 1 | $\sigma_{1}$ | $\sqrt{\frac{1}{\sum_{i} z_{i}} \sum_{i} z_{i}\left(y_{i}-\mu_{1}\right)^{2}}$ |

Notes: 1. $i$ stands for a unit in the population.
2. $\mu_{1}=\sum_{i} z_{i} y_{i} / \sum_{i} z_{i}=T_{1} /\left(N \alpha_{1}\right)$ is the mean of $y$ for class 1 .
3. By replacing $z_{i}$ with $\hat{z}_{i}$ in the formulas, $\hat{T}_{1}, \hat{\alpha}_{1}$, and $\hat{\sigma}_{1}$ can be calculated.

The parameter $\zeta$ requires the true values of $z$ and therefore cannot be computed in practice. An obvious estimator is obtained by replacing the unknown true $z$ with the observed $\hat{z}$ in the expressions for $\zeta$; this yields the domain statistic $\hat{\zeta}$. For example, by replacing $z_{i}$ with $\hat{z}_{i}$ in Table 2.1, formulas are obtained for the domain statistics $\hat{T}_{1}, \hat{\alpha}_{1}$, and $\hat{\sigma}_{1}$. Note that the hat in $\hat{\zeta}$ signifies an estimator, whereas in $\hat{z}_{i}$ it has no meaning other than to distinguish the observed indicator from the true indicator $z_{i}$.

Our study uses bias and variance to measure the effects that classification errors bring to $\hat{\zeta}$. Bias is defined as the difference between the expected values of the estimated output and the true value of domain parameters. Variance is a measure of the expected amount that the estimated domain statistics will change if different classification variables with the same error distribution are used. Mathematically, we define:

$$
\begin{array}{ll}
\text { Bias } \quad=\mathrm{E}(\hat{\zeta}-\zeta), \\
\text { Variance } & =\operatorname{Var}(\hat{\zeta})=\mathrm{E}\left((\hat{\zeta}-\mathrm{E}(\hat{\zeta}))^{2}\right) . \tag{2.1}
\end{array}
$$

Besides classification errors, for simplicity we assume that no other errors occur. In line with a common type of application in official statistics, we are interested here in bias and variance due to classification errors for a fixed finite population. That is to say, in (2.1) we implicitly condition on the realised values of $z_{1}, \ldots, z_{N}$ and $y_{1}, \ldots, y_{N}$.

### 2.2 Mixture model

### 2.2.1 Model setup

Our proposed new method to estimate the bias and variance of $\hat{\zeta}$ requires a model for the observed values $y_{1}, \ldots, y_{N}$ and $\hat{z}_{1}, \ldots, \hat{z}_{N}$. Here, we propose to use a mixture model, where the distributions in class $z=1$ and class $z=0$ may be different. For simplicity, we assume that the values of units $i=1, \ldots, N$ are drawn independently of each other. In addition, we assume that classification errors in $\hat{z}_{i}$ occur with the same probabilities for all units, which also means that they are independent of the continuous variable $y_{i}$. The latter assumption allows us to model the observed values $y_{i}$ and $\hat{z}_{i}$ separately, since it implies that, within each true class, the joint density of $y_{i}$ and $\hat{z}_{i}$ is factorised as follows:

$$
\begin{equation*}
f\left(y_{i}, \hat{z}_{i}=b \mid z_{i}=a\right)=f\left(y_{i} \mid z_{i}=a\right) \cdot P\left(\hat{z}_{i}=b \mid z_{i}=a\right) \tag{2.2}
\end{equation*}
$$

for all $i=1, \ldots, N$, with $a, b \in\{0,1\}$. Here, $P(\hat{z}=b \mid z=a)$ denotes a classification error probability and $f(y \mid z=a)$ denotes the density of the continuous variable in class $a$. We will now describe these two parts of the model in more detail.

Probability matrix. The probabilities of classification errors are modelled by a $2 \times 2$ transition matrix $\mathbf{P}$ (Formula 2.3). It describes the relationship between the true classes $z$ (rows) and the observed classes $\hat{z}$ (columns).

$$
\mathbf{P}=\left(\begin{array}{cc}
p_{11} & 1-p_{11}  \tag{2.3}\\
1-p_{00} & p_{00}
\end{array}\right) .
$$

The value $p_{a b}$ indicates the probability of observing a unit in class $b$ when its true class is $a$. Thus, for unit $i$, if its true class is 1 , its probability to be observed in class 1 is $P\left(\hat{z}_{i}=1 \mid z_{i}=1\right)=p_{11}$, and its probability to be observed in class 0 is $P\left(\hat{z}_{i}=0 \mid z_{i}=1\right)=1-p_{11}$; similarly, for a unit in true class $0, P\left(\hat{z}_{i}=0 \mid z_{i}=0\right)=$ $p_{00}$ and $P\left(\hat{z}_{i}=1 \mid z_{i}=0\right)=1-p_{00}$. For reasonable classifiers, the values of $p_{11}$ and $p_{00}$ should be above 0.5 .

Gaussian mixture model. We assume that the distribution of $y$ for each class $z$ follows a Gaussian mixture model. The number of Gaussian components in class 1 is $q_{1}$, and the number in class 0 is $q_{0}$. Note that this means that the overall model for $y$ can be seen as a "mixture of mixtures", with the first mixing
level given by the true class $z$ and the second mixing level by the Gaussian mixture components within a true class.

For better explanation, a component variable $m$ is defined to identify which component of the Gaussian mixture model each unit belongs to. The variable $m$ is unobserved. The distribution of $y_{i}$ depends on the class it belongs to (the value of $z_{i}$ ) and also which component in this class it belongs to (the value of $m_{i}$ ); we make the conventional assumption that each unit $i$ belongs to a unique class-component pair $\left(z_{i}, m_{i}\right)$. By the law of total probability, the density of $y_{i}$ conditional on $z_{i}$ is:

$$
f\left(y_{i} \mid z_{i}=a\right)= \begin{cases}\sum_{j=1}^{q_{1}} P\left(m_{i}=j \mid z_{i}=1\right) \cdot f\left(y_{i} \mid z_{i}=1, m_{i}=j\right), & \text { if } a=1, \\ \sum_{k=1}^{q_{0}} P\left(m_{i}=k \mid z_{i}=0\right) \cdot f\left(y_{i} \mid z_{i}=0, m_{i}=k\right), & \text { if } a=0\end{cases}
$$

where $P(m=j \mid z=1)$ is the mixture weight of a component in class 1 , denoted as $\xi_{1 j}\left(j \in\left\{1, \ldots, q_{1}\right\}\right)$; $P(m=k \mid z=0)$ is the mixture weight of a component in class 0 , denoted as $\xi_{0 k}\left(k \in\left\{1, \ldots, q_{0}\right\}\right)$. These mixture weights satisfy $\sum_{j=1}^{q_{1}} \xi_{1 j}=1$ and $\sum_{k=1}^{q_{0}} \xi_{0 k}=1$.

In a Gaussian mixture, it is assumed that each component follows a normal distribution. Hence, in this case we obtain:

$$
f\left(y_{i} \mid z_{i}=a\right)= \begin{cases}\sum_{j=1}^{q_{1}} \xi_{1 j} \cdot \varphi\left(y_{i} ; \mu_{1 j}, \sigma_{1 j}^{2}\right), & \text { if } a=1  \tag{2.4}\\ \sum_{k=1}^{q_{0}} \xi_{0 k} \cdot \varphi\left(y_{i} ; \mu_{0 k}, \sigma_{0 k}^{2}\right), & \text { if } a=0\end{cases}
$$

where $\mu_{1 j}$ is the mean of component $j$ in class 1 and $\sigma_{1 j}$ is its standard deviation; $\mu_{0 k}$ is the mean of component $k$ in class 0 and $\sigma_{0 k}$ is its standard deviation; $\varphi\left(y ; \mu, \sigma^{2}\right)=\sigma^{-1}(2 \pi)^{-1 / 2} \exp \left\{-(y-\mu)^{2} /\left(2 \sigma^{2}\right)\right\}$ denotes the density of a normal distribution with parameters $\mu$ and $\sigma^{2}$.

Let $\boldsymbol{\theta}=\left(\alpha_{1}, p_{11}, p_{00}, \xi_{11}, \mu_{11}, \sigma_{11}, \ldots, \xi_{1 q_{1}}, \mu_{1 q_{1}}, \sigma_{1 q_{1}}, \xi_{01}, \mu_{01}, \sigma_{01}, \ldots, \xi_{0 q_{0}}, \mu_{0 q_{0}}, \sigma_{0 q_{0}}\right)^{\prime}$ denote the vector of unknown parameters of the Gaussian mixture model. The identification of all parameters in $\boldsymbol{\theta}$ requires that the order of the components in each class is fixed. For simplicity, we assume here that $\mu_{11}<\ldots<\mu_{1 j}<\ldots<$ $\mu_{1 q_{1}}$ and $\mu_{01}<\ldots<\mu_{0 k}<\ldots<\mu_{0 q_{0}}$.

In practice, the appropriate numbers of components in the Gaussian mixture models for class 1 and class $0, q_{1}$ and $q_{0}$, are not known and need to be determined from the observed data. Here, the main purpose of the mixture model is to provide a flexible way to model the distribution of $y$ in each class. For this type of application, a commonly-used approach is to fit several mixture models to the data with different numbers of components and use the Bayesian information criterion (BIC) to select the optimal number of components (McLachlan and Peel, 2000, page 175 and pages 209-210). However, more recent research suggests that when the components in a mixture model have a very similar mean and variance, the Fisher information matrix may become singular and the BIC is no longer justified as a criterion (Drton and Plummer, 2017).

For such situations, Drton and Plummer (2017) proposed a modified BIC, referred to as the sBIC. In the case study to be discussed in Section 4, we compare the optimal number of components according to the BIC and sBIC criteria.

### 2.2.2 EM algorithm

Maximum likelihood estimation of the above Gaussian mixture model can be achieved using an EM algorithm (Dempster, Laird and Rubin, 1977; Little and Rubin, 2002). This type of algorithm is often applied when there are unobserved variables in statistical models. As illustrated by its name, it contains two steps: an E step and an M step. The E step builds the expected value of the complete-data log-likelihood function, conditional on the observed data. The $M$ step then estimates the model parameters, which in turn provide input for the next E step. The algorithm proceeds in an iterative way until convergence.

It should be noted that the EM algorithm can estimate all parameters of the mixture model (including the matrix $\mathbf{P}$ ) from a data set of observations $\left(\hat{z}_{i}, y_{i}\right)$; thus, it is not necessary to have observed the true class $z_{i}$ for any unit. Loosely stated, this is possible because, on the one hand, the probability that an observation $\left(\hat{z}_{i}, y_{i}\right)$ belongs to class 1 or class 0 can be predicted based on differences in the distribution of $y$ for class 1 and class 0 (which is done during the E step) and, on the other hand, the distributions of $y$ and $\hat{z}$ within each true class can be estimated based on these predicted probabilities (which is done during the M step). Technically, estimation is possible because (under normal circumstances) there exists a unique set of parameter values for $\boldsymbol{\theta}$ for which the complete-data log-likelihood function achieves its global maximum. In other words: the model is identified; see McLachlan and Peel (2000) for more details.

To derive the complete-data log-likelihood function, we note from expressions (2.2), (2.3), and (2.4) that

$$
f\left(z_{i}, m_{i}, \hat{z}_{i}, y_{i} ; \boldsymbol{\theta}\right)=\prod_{j=1}^{q_{1}} \omega_{1 j i}^{z_{i} \mathbf{1}_{\left(m_{i}-j\right)}} \prod_{k=1}^{q_{0}} \omega_{0 k i}^{\left(1-z_{i}\right) \mathbf{1}_{\left(m_{i}-k\right)}},
$$

where

$$
\begin{gathered}
\omega_{1 j i} \triangleq \alpha_{1} p_{11}^{\hat{i}_{i}}\left(1-p_{11}\right)^{1-\hat{z}_{i}} \frac{\xi_{1 j}}{\sigma_{1 j} \sqrt{2 \pi}} \exp \left\{-\frac{\left(y_{i}-\mu_{1 j}\right)^{2}}{2 \sigma_{1 j}^{2}}\right\}, \\
\omega_{0 k i} \triangleq\left(1-\alpha_{1}\right)\left(1-p_{00}\right)^{\hat{z}_{i}} p_{00}^{1-\hat{z}_{i}} \frac{\xi_{0 k}}{\sigma_{0 k} \sqrt{2 \pi}} \exp \left\{-\frac{\left(y_{i}-\mu_{0 k}\right)^{2}}{2 \sigma_{0 k}^{2}}\right\},
\end{gathered}
$$

and the indicator functions for $m_{i}$ are defined as:

$$
\mathbf{1}_{\left(m_{i}=j\right)}=\left\{\begin{array}{ll}
1 & m_{i}=j \\
0 & m_{i} \neq j
\end{array} ; \quad \mathbf{1}_{\left(m_{i}=k\right)}=\left\{\begin{array}{ll}
1 & m_{i}=k \\
0 & m_{i} \neq k
\end{array} .\right.\right.
$$

Note that $z_{i} \sum_{j=1}^{q_{1}} \mathbf{1}_{\left(m_{i}=j\right)}=z_{i}$ and $\left(1-z_{i}\right) \sum_{k=1}^{q_{0}} \mathbf{1}_{\left(m_{i}=k\right)}=1-z_{i}$ for all units $i$.

It follows that the complete-data log-likelihood of the Gaussian mixture model can be written as:

$$
\begin{equation*}
\operatorname{LL}(\boldsymbol{\theta})=\sum_{i=1}^{N} \log f\left(z_{i}, m_{i}, \hat{z}_{i}, y_{i} ; \boldsymbol{\theta}\right)=\sum_{i=1}^{N}\left\{z_{i} \sum_{j=1}^{q_{i}} \mathbf{1}_{\left(m_{i}=j\right)} \log \omega_{1 j i}+\left(1-z_{i}\right) \sum_{k=1}^{q_{0}} \mathbf{1}_{\left(m_{i}=k\right)} \log \omega_{0 k i}\right\}, \tag{2.5}
\end{equation*}
$$

where we used the assumption that the units are drawn independently of each other.
E step. In the E step of the algorithm, the unobserved quantities $z_{i} \mathbf{1}_{\left(m_{i}-j\right)}$ and $\left(1-z_{i}\right) \mathbf{1}_{\left(m_{i}=k\right)}$ in (2.5) are replaced by their conditional expectations, given the observed values $\hat{z}_{i}$ and $y_{i}$ :

$$
\begin{gathered}
\mathrm{E}\left(z_{i} \mathbf{1}_{\left(m_{i}=j\right)} \mid \hat{z}=\hat{z}_{i}, y=y_{i}\right)=P\left(z_{i}=1, m_{i}=j \mid \hat{z}=\hat{z}_{i}, y=y_{i}\right)=\frac{\omega_{1 j i}}{\sum_{j=1}^{q_{1}} \omega_{1 j i}+\sum_{k=1}^{q_{0}} \omega_{0 k i}} \triangleq A_{1 j i} ; \\
\mathrm{E}\left(\left(1-z_{i}\right) \mathbf{1}_{\left(m_{i}=k\right)} \mid \hat{z}=\hat{z}_{i}, y=y_{i}\right)=P\left(z_{i}=0, m_{i}=k \mid \hat{z}=\hat{z}_{i}, y=y_{i}\right)=\frac{\omega_{0 k i}}{\sum_{j=1}^{q_{1}} \omega_{1 j i}+\sum_{k=1}^{q_{0}} \omega_{0 k i}} \triangleq A_{0 k i} .
\end{gathered}
$$

During iteration $t$ of the algorithm, these expressions are evaluated using the current parameter estimates $\hat{\boldsymbol{\theta}}^{(t-1)}$, yielding the values $A_{1 j i}^{(t)}$ and $A_{0 k i}^{(t)}$.

M step. In the M step of the algorithm, the log-likelihood function (2.5) is maximised with respect to the model parameters, with the unobserved quantities replaced by $A_{1 j i}^{(t)}$ and $A_{0 k i}^{(t)}$ from the most recent E step. By setting the first-order partial derivatives of this expected log-likelihood equal to zero, the following formulas are obtained to update the model parameters:

$$
\begin{array}{ll}
\hat{\alpha}_{1}^{(t+1)}=\frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_{1}} A_{1 j i}^{(t)}\right)}{N} ; & \\
\hat{p}_{11}^{(t+1)}=\frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_{1}} A_{1 j i}^{(t)}\right) \hat{z}_{i}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_{1}} A_{1 j i}^{(t)} ;\right.} ; & \hat{p}_{00}^{(t+1)}=\frac{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_{0}} A_{0 k i}^{(t)}\right)\left(1-\hat{z}_{i}\right)}{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_{0}} A_{0 k i}^{(t)}\right)} ; \\
\hat{\xi}_{1 j}^{(t+1)}=\frac{\sum_{i=1}^{N} A_{1 j i}^{(t)}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_{1}} A_{1 j i}^{(t)}\right)} ; & \hat{\xi}_{0 k}^{(t+1)}=\frac{\sum_{i=1}^{N} A_{0 k i}^{(t)}}{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_{0}} A_{0 k i}^{(t)}\right)} ; \\
\hat{\mu}_{1 j}^{(t+1)}=\frac{\hat{\mu}_{0 k}^{(t+1)}=\frac{\sum_{i=1}^{N} A_{1 j}^{(t)} A_{i}^{(t)} y_{i}}{\sum_{i=1}^{N} A_{1 j i}^{(t)} ;} ;}{\sum_{i=1}^{N} A_{0 k i}^{(t)} ;} \\
\hat{\sigma}_{1 j}^{(t+1)}=\sqrt{\frac{\sum_{i=1}^{N} A_{1 j i}^{(t)}\left(y_{i}-\hat{\mu}_{1 j}^{(t+1)}\right)^{2}}{\sum_{i=1}^{N} A_{1 j i}^{(t)}} ;} & \hat{\sigma}_{0 k}^{(t+1)}=\sqrt{\frac{\sum_{i=1}^{N} A_{0 k i}^{(t)}\left(y_{i}-\hat{\mu}_{0 k}^{(t+1)}\right)^{2}}{\sum_{i=1}^{N} A_{0 k i}^{(t)}} .}
\end{array}
$$

As noted above, the EM algorithm can be used to estimate the mixture model even when the true class $z_{i}$ is never observed in the available data. In general, however, the EM algorithm will merely converge to
a local maximum of the log-likelihood function. To ensure that the global maximum is found, it may be necessary to run the algorithm multiple times using different (randomly chosen) starting values and retain the best solution. For general suggestions on how to choose suitable random starting values for mixture models, see McLachlan and Peel (2000, Section 2.12).

Alternatively, observations $\left(z_{i}, \hat{z}_{i}, y_{i}\right)$ including the true class may have been obtained for a random subsample of the data (an audit sample). If available, an audit sample can be used to improve the convergence of the EM algorithm to the global maximum of the likelihood function, reducing the need for re-runs with different starting values. First, the parameters $\alpha_{1}, p_{11}$ and $p_{00}$ can be estimated directly from the audit sample to provide reasonable starting values for the EM algorithm. In addition, improved starting values for the other parameters (related to the components of the Gaussian mixture inside each class) could be obtained by applying a $k$-means clustering algorithm to each true class in the audit sample (Li, 2020b). Finally, for observations from the audit sample, $A_{1 j i}$ and $A_{0 k i}$ can be replaced during the E step by more narrowly defined expected values:

$$
\begin{gathered}
\mathrm{E}\left(z_{i} \mathbf{1}_{\left(m_{i}=j\right)} \mid z_{i}=1, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=P\left(m_{i}=j \mid z_{i}=1, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=\frac{\omega_{1 j i}}{\sum_{j=1}^{q_{j} \omega_{1 j i}}} \triangleq Q_{1 j i} ; \\
\mathrm{E}\left(\left(1-z_{i}\right) \mathbf{1}_{\left(m_{i}=k\right)} \mid z_{i}=0, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=P\left(m_{i}=k \mid z_{i}=0, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=\frac{\omega_{0 k i}}{\sum_{k=1}^{q_{0}} \omega_{0 k i}} \triangleq Q_{0 k i} ;
\end{gathered}
$$

whereas $\mathrm{E}\left(z_{i} \mathbf{1}_{\left(m_{i}=j\right)} \mid z_{i}=0, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=\mathrm{E}\left(\left(1-z_{i}\right) \mathbf{1}_{\left(m_{i}=k\right)} \mid z_{i}=1, \hat{z}=\hat{z}_{i}, y=y_{i}\right)=0$.

### 2.3 Methods

In this study we will compare three methods that try to estimate the bias and variance of a domain statistic $\hat{\zeta}$ as defined in (2.1): the bootstrap method, the EM bootstrap method and the SIMEX bootstrap method. We did not compare the outcomes with analytical expressions, since we already know that for domain statistics those expressions yield results that are either similar to those of the bootstrap method or less accurate (see Introduction).

As noted at the end of Section 2.1, we are interested here in bias and variance due to classification errors for a finite population, conditional on the values $z_{1}, \ldots, z_{N}$ and $y_{1}, \ldots, y_{N}$. Due to assumption (2.2), this means that the bias and variance of interest are completely determined by the random process described by the matrix $\mathbf{P}$, applied to the fixed values $z_{1}, \ldots, z_{N}$.

An important practical consideration is that the bootstrap and SIMEX bootstrap methods assume that (an estimate of) the matrix $\mathbf{P}$ is available beforehand, whereas an estimate of $\mathbf{P}$ is obtained as part of the EM bootstrap method. For the other methods, $\mathbf{P}$ could be estimated in practice from an audit sample or by running the EM algorithm separately. In our study to be discussed in Sections 3 and 4, we used the estimated $\mathbf{P}$ from the EM algorithm as input for the bootstrap and SIMEX bootstrap methods.

### 2.3.1 Bootstrap method

The bootstrap method as applied here originates from Van Delden et al. (2016). It is summarised in Algorithm 1. The classification matrix $\mathbf{P}$ is applied to the observed class $\hat{z}$ and bootstrapped classes $z^{*}$ are obtained. The probability that the bootstrapped class is 1 given the observed class 1 is $P\left(z_{i}^{*}=1 \mid \hat{z}_{i}=1\right)=p_{11}$ and given the observed class 0 is $P\left(z_{i}^{*}=1 \mid \hat{z}_{i}=0\right)=1-p_{00}$. Through bootstrapping, random classification errors are introduced to observed classes. As a result, estimated bias is computed by comparing bootstrapped statistics $\zeta^{*}$ to the observed statistic $\hat{\zeta}$, and the variance of the bootstrapped statistics $\zeta^{*}$ is an estimate of the variance of the observed domain statistic $\hat{\zeta}$. In practice, the theoretical bootstrap bias and variance are usually approximated using a finite number $(S)$ of bootstrap samples. For certain simple statistics such as $\alpha_{1}$ and $T_{1}$, it is also possible to derive an exact formula for the theoretical bootstrap bias and variance (Van Delden et al., 2016).

```
Algorithm 1 The bootstrap method
    Input: Observations \(\left(y_{i}, \hat{z}_{i}\right)\) for \(i=1, \ldots, N\), matrix \(\mathbf{P}\) and \(S\).
    for \(s=1 \ldots S\) do
        Generate \(z_{i}^{*}\) by \(\mathbf{P}\), conditional on \(\hat{z}_{i}\), for every unit \(i\) in the data set
        Calculate the corresponding \(\zeta^{*}\) based on \(\left(y_{i}, z_{i}^{*}\right)\) instead of \(\left(y_{i}, \hat{z}_{i}\right)\)
        end for
        Calculate \(\operatorname{Bias}_{\text {boot }}=\mathrm{E}\left(\zeta^{*} \mid \hat{z}\right)-\hat{\zeta}, \operatorname{Var}_{\text {boot }}=\operatorname{Var}\left(\zeta^{*} \mid \hat{z}\right)\) based on \(S\) simulations
    Output: Bias \(_{\text {boot }}\) and Var \(_{\text {boot }}\)
```

It is known that, in general, the bias and variance estimates from Algorithm 1 are biased, due to the fact that the observed classes $\hat{z}_{i}$ are used as a starting point for the bootstrap. This problem is illustrated in Appendix C. 1 using the parameter $\zeta=T_{1}$, for which an exact analysis is possible. The next two methods attempt to correct for this bias.

### 2.3.2 EM bootstrap method

In the EM bootstrap method, the observed data $(y, \hat{z})$ are assumed to follow the Gaussian mixture model from Section 2.2. The EM algorithm estimates the parameters $\boldsymbol{\theta}$ of the model, which include the classification probabilities $p_{11}$ and $p_{00}$. Then we apply a nested simulation process; see Algorithm 2. The purpose of the first level of simulation is to restore (in expectation) an error-free status. It generates classes $\tilde{z}$ from $P(z \mid y, \hat{z} ; \hat{\boldsymbol{\theta}})$, leading to unbiased statistics $\tilde{\zeta}$. [Note that the required probabilities are available directly from the EM algorithm, since $P\left(z_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=\sum_{j=1}^{q_{1}} A_{1 j i}$ and $P\left(z_{i}=0 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=\sum_{k=1}^{q_{0}} A_{0 k i}$.] After that, classes $\tilde{z}^{*}$ are generated by a bootstrapping process with matrix $\mathbf{P}$, which leads to bias and variance estimation similar to Algorithm 1. Finally, the results of this inner bootstrap simulation are averaged over the first level of simulation, to reduce the effect of noise due to drawing $\tilde{z}$ from a probability distribution.

```
Algorithm 2 The EM bootstrap method
    Input: Observations \(\left(y_{i}, \hat{z}_{i}\right)\) for \(i=1, \ldots, N\) and \(S_{1}, S_{2}\).
    Estimate model parameters \(\boldsymbol{\theta}\), including \(p_{11}\) and \(p_{00}\), from the EM algorithm, conditional on \(\hat{z}\) and \(y\)
    Calculate \(P\left(z_{i} \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)\) for every unit \(i\) in the data set
    for \(s_{1}=1, \ldots, S_{1}\) do
        Generate \(\tilde{z}_{i}\) by \(P\left(z_{i} \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)\) for every unit \(i\) in the data set
        Calculate the corresponding \(\tilde{\zeta}\) based on \(\left(y_{i}, \tilde{z}_{i}\right)\) instead of \(\left(y_{i}, \hat{z}_{i}\right)\)
        for \(s_{2}=1, \ldots, S_{2}\) do
            Generate \(\tilde{z}_{i}^{*}\) by \(\mathbf{P}\), conditional on \(\tilde{z}_{i}\), for every unit \(i\) in the data set
            Calculate the corresponding \(\tilde{\zeta}^{*}\) based on \(\left(y_{i}, \tilde{z}_{i}^{*}\right)\) instead of \(\left(y_{i}, \hat{z}_{i}\right)\)
        end for
        end for
        Calculate \(\operatorname{Bias}_{\text {comb }}=\mathrm{E}_{\tilde{z}}\left(\mathrm{E}\left(\tilde{\zeta}^{*} \mid \hat{z}, \tilde{z}\right)-\tilde{\zeta} \mid \hat{z}\right)\) and \(\operatorname{Var}_{\text {comb }}=\mathrm{E}_{\tilde{z}}\left(\operatorname{Var}\left(\tilde{\zeta}^{*} \mid \hat{z}, \tilde{z}\right) \mid \hat{z}\right)\) based on \(\quad S_{1} \quad\) and \(\quad S_{2}\)
        simulations
    Output: Bias \(_{\text {comb }}, \mathbf{V a r}_{\text {comb }}\) and estimated matrix \(\mathbf{P}\)
```

It is shown in Appendix C. 2 that, for $S_{1}, S_{2} \rightarrow \infty$, Algorithm 2 indeed yields approximately unbiased bias and variance estimators in the special cases $\zeta=T_{1}$ and $\zeta=\alpha_{1}$, provided that the observed data follow the assumed mixture model. For other, non-linear parameters such as $\zeta=\sigma_{1}$, no exact proof is available, but we will investigate the behaviour of Algorithm 2 in a simulation study.

### 2.3.3 SIMEX bootstrap method

SIMEX was introduced by Cook and Stefanski (1994) for numerical variables. It uses a bootstrapping process to add various extra errors, through which a sequence of estimates under different error-included conditions is obtained. Then a function is applied to extrapolate the sequence back to the estimate without error. Here, we use an extension of the SIMEX method to categorical variables that was introduced by Küchenhoff et al. (2006).

Traditionally, the SIMEX method would be applied to obtain a bias-corrected estimate of the target parameter itself ( $\zeta$ in our notation). Here, we use it to obtain bias-corrected bootstrap estimates of the bias and variance of $\hat{\zeta}$. We will refer to this approach as the SIMEX bootstrap method.

The SIMEX bootstrap method simulates multiple conditions where classification errors are added to the data according to the matrix $\mathbf{P}^{\lambda}$ [with $\mathbf{P}$ given by (2.3)], for different values of $\lambda \geq 0$. For each value of $\lambda$, the bootstrap method of Algorithm 1 is applied to the adjusted data to obtain bias and variance estimates. Finally, the SIMEX bias and variance estimates are obtained by extrapolating the sequence of bias and variance estimates as functions of $\lambda$ to the value $\lambda=-1$. This can be understood as follows. The available observed data can be viewed as one realisation of applying the matrix $\mathbf{P}$ to the true data. So, starting from the observed data at $\lambda=0$ we want to extrapolate the sequence of bias and variance estimates back to what
would have been found at the unobserved point of zero misclassifications, which corresponds to $\lambda=-1$. The SIMEX bootstrap method is summarised in Algorithm 3.

```
Algorithm 3 The SIMEX bootstrap method
    Input: Observations \(\left(y_{i}, \hat{z}_{i}\right)\) for \(i=1, \ldots, N\), matrix \(\mathbf{P}\) and \(S_{1}, S_{2}\).
            for \(\lambda\) ranging from 0 to 5 with increments of 0.5 do
        Calculate \(\mathbf{P}^{\lambda}\)
        for \(s_{1}=1, \ldots, S_{1}\) do
            Generate \(z_{i}^{(\lambda)}\) by \(\mathbf{P}^{\lambda}\), conditional on \(\hat{z}_{i}\), for every unit \(i\) in the data set
            Calculate the corresponding \(\zeta^{(\lambda)}\) based on \(\left(y_{i}, z_{i}^{(\lambda)}\right)\) instead of \(\left(y_{i}, \hat{z}_{i}\right)\)
            for \(s_{2}=1, \ldots, S_{2}\) do
                        Generate \(\hat{z}_{i}^{(\lambda)}\) by \(\mathbf{P}\), conditional on \(z_{i}^{(\lambda)}\), for every unit \(i\) in the data set
                        Calculate the corresponding \(\hat{\zeta}^{(\lambda)}\) based on ( \(y_{i}, \hat{z}_{i}^{(\lambda)}\) ) instead of \(\left(y_{i}, \hat{z}_{i}\right)\)
            end for
        end for
        Calculate \(\operatorname{Bias}_{\text {simex }}^{(\lambda)}=\mathrm{E}_{z^{(\lambda)}}\left(\mathrm{E}\left(\hat{\zeta}^{(\lambda)} \mid \hat{z}, z^{(\lambda)}\right)-\zeta^{(\lambda)} \mid \hat{z}\right)\) and \(\operatorname{Var}_{\text {simex }}^{(\lambda)}=\mathrm{E}_{z^{(\lambda)}}\left(\operatorname{Var}\left(\hat{\zeta}^{(\lambda)} \mid \hat{z}, z^{(\lambda)}\right) \mid \hat{z}\right)\) based on \(S_{1}\) and \(S_{2}\)
            simulations
        12: end for
        13: Extrapolate \(\operatorname{Bias}_{\text {simex }}^{(\lambda)}\) and \(\operatorname{Var}_{\text {simex }}^{(\lambda)}\) to \(\lambda=-1\) to get \(\boldsymbol{B i a s}_{\text {simex }}\) and \(\mathbf{V a r}_{\text {simex }}\)
    Output: Bias \(_{\text {simex }}\), Var \(_{\text {simex }}\)
```

Calculate $\mathbf{P}^{\lambda}$. When $\lambda=0, \mathbf{P}^{0}$ is an identity matrix. In this special case, no additional classification errors are introduced in line 4 of the SIMEX bootstrap algorithm. For any real-valued $\lambda>0, \mathbf{P}^{\lambda}$ can be computed using the eigenvalue decomposition of $\mathbf{P}$. Through it, we get $\mathbf{P}=\mathbf{Q A} \mathbf{Q}^{-1}$, where $\mathbf{A}$ is a diagonal matrix of eigenvalues and $\mathbf{Q}$ is a matrix of eigenvectors of $\mathbf{P}$. Then $\mathbf{P}^{\lambda}$ is calculated by $\mathbf{P}^{\lambda}=\mathbf{Q A}^{\lambda} \mathbf{Q}^{-1}$. The corresponding probabilities $p_{11}^{(\lambda)}$ and $p_{00}^{(\lambda)}$ are obtained:

$$
\mathbf{P}^{\lambda}=\left(\begin{array}{cc}
p_{11}^{(\lambda)} & 1-p_{11}^{(\lambda)} \\
1-p_{00}^{(\lambda)} & p_{00}^{(\lambda)}
\end{array}\right),
$$

where $p_{11}^{(\lambda)}$ is the probability of $z^{(\lambda)}=1$ when $\hat{z}=1$, and $p_{00}^{(\lambda)}$ is the probability of $z^{(\lambda)}=0$ when $\hat{z}=0$. These probabilities are used to draw $z_{i}^{(\lambda)}$, given $\hat{z}_{i}$, in line 4 of the algorithm. Note that this procedure works only if $\mathbf{P}^{\lambda}$ is a true probability matrix in the sense that $0 \leq p_{11}^{(\lambda)}, p_{00}^{(\lambda)} \leq 1$. For $\lambda \geq 0$, this is guaranteed provided that $p_{11}+p_{00}>1$ (Küchenhoff et al., 2006). For $\lambda<0$ this property does not hold; hence, extrapolation is a necessary step of the SIMEX method.

Extrapolation functions. A SIMEX method yields a consistent estimator of a parameter of interest (in our case: bias and variance of $\hat{\zeta}$ ) when the extrapolation function, which describes how the uncorrected estimator varies as a function of $\lambda$, is correctly specified; see Küchenhoff et al. (2006) for more details. In
the literature on SIMEX, extrapolation functions that have been suggested include: local linear regression (LOESS) on $\lambda$ (Hopkins and King, 2010) and standard regression on a quadratic or cubic polynomial of $\lambda$ (Küchenhoff et al., 2006). We have compared all three approaches in the simulation study of Section 3.

## 3. Simulation study

### 3.1 Settings

We simulated a population of size $N=2,000$ with two classes. For target variable $y$ we used the following Gaussian mixture distribution: Class 0 has one component with $\mu_{0}=15, \sigma_{0}=3$; class 1 has two components with $\left(\xi_{11}, \xi_{12}\right)=(0.5,0.5),\left(\mu_{11}, \mu_{12}\right)=(2,4)$, and $\left(\sigma_{11}, \sigma_{12}\right)=(1,2)$.

With respect to the true classification variable $z$, we tested two different proportions of class 1: $\alpha_{1}$ is 0.3 or 0.5 . The observed classification variable $\hat{z}$ was generated from $z$ by using the transition matrix $\mathbf{P}$ given in equation (2.3). In the simulation study, values of $p_{11}$ and $p_{00}$ were set at $0.6,0.75$, or 0.9 . For each setting of $\alpha_{1}, p_{11}$ and $p_{00}$, we used $S_{0}=100$ implying that 100 sets of $\hat{z}$ were generated from $z$. In each set $s_{0}, 5 \%$ of the units from the population (so 100 in total) were randomly selected as an audit sample which was used to obtain the starting values for the EM algorithm of $\hat{p}_{11}, \hat{p}_{00}$ and the other class-level parameters. The starting values of the component-level parameters were obtained by k-means; see Li (2020b) for more details. In a preliminary study, we have also tested the EM algorithm without an audit sample (see Section 3.3).

For each set $s_{0}$, the bootstrap method, the SIMEX bootstrap method and the EM bootstrap method were applied to estimate the corresponding bias and variance of the estimated domain parameters $\hat{\zeta}$ that are given in Table 2.1: the total sum for class $1\left(T_{1}\right)$, the proportion of class $1\left(\alpha_{1}\right)$ and the standard deviation for class $1\left(\sigma_{1}\right)$. The bias and variance estimates are given here as the average over the $S_{0}$ sets.

With respect to the SIMEX bootstrap method, we tested the above settings of the simulation study for three different extrapolation functions, as noted in Section 2.3.3; see also Appendix A. The estimates of the bias and of the standard error of the LOESS function and of the third order polynomial were closer to the true values than those of the second order polynomial. Since the LOESS function was used in a previous study on misclassifications (see Hopkins and King (2010)) we used this extrapolation function in the remainder of this paper.

The EM algorithm was stopped either when none of the parameter estimates changed by more than 0.001 between two iterations, or after 5,000 iterations. In practice, both in this simulation study and in the case study of Section 4, this maximum number of iterations was never reached. For about $1 \%$ of the simulated data sets, the EM algorithm did not converge properly due to numerical issues. These were caused by an unfortunate choice of starting values estimated from the audit sample, for instance a starting value of $p_{11}$ exactly equal to 1 . In principle, this problem could be avoided easily by a slight change of starting values. However, as this issue only affected a small number of cases, for convenience we ignored these cases in the results below.

The number of iterations in the methods, $S, S_{1}$ and $S_{2}$ were all set at 100 . To obtain a benchmark for the true bias and variance of the estimated domain parameters $\hat{\zeta}, 1,000$ sets of $\hat{z}$ were simulated.

### 3.2 Results

Bias estimation. Figure 3.1 shows the bias of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ for $\alpha_{1}=0.3$ (row 1,3 and 5 respectively) and for $\alpha_{1}=0.5$ (row 2, 4 and 6 respectively). Each column shows results under the three $p_{11}$ settings. In each subplot, the horizontal axis indicates values of $p_{00}$, the vertical axis defines the estimates under different conditions and methods. Furthermore, the different methods are given by different symbols: the true values $(\cdot)$, the bootstrap $(\Delta)$, the SIMEX bootstrap $(\nabla)$ and the EM bootstrap method $(\times)$. Figure 3.2 shows the standard error (square root of the variance) of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ for $\alpha_{1}=0.3$ and $\alpha_{1}=0.5$ for the same settings.

Overall, for a given value of $p_{11}$, the bias of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ decreased with a larger value of $p_{00}$. For a given value of $p_{00}$, the bias increased with a larger value of $p_{11}$. This result can be understood as follows. In the case of the total, an analytical expression for the bias of $\hat{T}_{1}$ is given by (C.1). Normally, this expression cannot directly be computed since $T_{0}$ and $T_{1}$ are unknown in real situations, but in the simulations we know their values. In our example, at $\alpha_{1}=0.3$ one finds that $T_{0}=15 \times(1-0.3) \times 2,000=21,000$ and $T_{1}=(0.5 \times 2+$ $0.5 \times 4) \times(0.3) \times 2,000=1,800$. An increase of $p_{00}$ from 0.6 to 0.9 for a given value of $p_{11}$ reduces the bias since the contribution $\left(1-p_{00}\right) T_{0}$ drops from 8,400 to 2,100 . This refers to units with true class 0 that are erroneously observed as class 1 (overestimation). Conversely, an increase of $p_{11}$ from 0.6 to 0.9 for a given value of $p_{00}$ leads to a small increase of the bias due to an increase in the contribution $\left(p_{11}-1\right) T_{1}$ from -720 to -180 . This contribution refers to units with true class 1 that erroneously have an observed class 0 (underestimation). Generally, for a given value of $p_{11}$ the bias of the statistics of interest ( $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ ) decreases with a larger value of $p_{00}$ because their overestimation decreases. Conversely, for a given value of $p_{00}$ the bias of the statistics increases with a larger value of $p_{11}$ because its underestimation decreases.

With respect to the three estimation methods (see Figure 3.1), we found that the bias estimates for the statistics of interest from the EM bootstrap method were closest to the true values. The bias estimates from the SIMEX method were closer to the true values than the estimates from the bootstrap method, but they still had a considerable distance to the true bias. Furthermore, we found that the bias estimates from all three methods were closer to their true bias for $T_{1}$ and $\sigma_{1}$ when the misclassification probabilities were reduced ( $p_{11}$ or $p_{00}$ values closer to 1 ). When $p_{11}$ and $p_{00}$ were equal to 0.9 , bias estimates from the bootstrap method and the SIMEX bootstrap method even overlapped with the corresponding true bias. Finally note that when $\alpha_{1}=0.5$ and $p_{00}=p_{11}$, then the true bias of $\hat{\alpha}_{1}$ equals 0 and all three methods estimated this bias correctly.

Variance estimation. Figure 3.2 shows the true and estimated standard error of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ for $\alpha_{1}=0.3$ (row 1,3 and 5 respectively) and for $\alpha_{1}=0.5$ (row 2,4 and 6 respectively). The true standard errors of $\hat{T}_{1}$ and of $\hat{\alpha}_{1}$ were reduced with less misclassifications with $p_{00}$ and $p_{11}$ going from 0.6 to 0.9 . This result follows from the analytical expression (C.2) for the variance of $\hat{T}_{1}$.

Figure 3.1 Bias estimation in the simulation study.


Notes: Rows represent different domain parameters $\left(T_{1}, \alpha_{1}, \sigma_{1}\right)$ for two settings of $\alpha_{1}: \alpha_{1}=0.3$ and $\alpha_{1}=0.5$. Columns represent different $p_{11}$ values, within each column different $p_{00}$ values are given.

Figure 3.2 Standard error estimation in the simulation study.


Notes: For row and column settings see Figure 3.1.

Surprisingly, the standard error of $\hat{\sigma}_{1}$ clearly increased for a given value of $p_{11}$ when $p_{00}$ increased from 0.6 to 0.9 . Furthermore, there was a very small reduction of this standard error for a given value of $p_{00}$ when $p_{11}$ increased from 0.6 to 0.9 . This result can be explained as follows. The mean turnover level is much higher in class 0 than in class 1 . When most units observed in class 1 truly come from class 1 and relatively few units from class $0\left(p_{00}=0.9\right)$, the standard error of $\hat{\sigma}_{1}$ is large since there is a considerable variation over the replicates in which turnover values of class 0 will be observed as class 1 . In some of the replicates this concerns outlying values compared to the true distribution in class 1 . When the number of units from class 0 increases ( $p_{00}=0.75$ ), this variation in turnover values of the replicates decreases and thus the standard error of $\hat{\sigma}_{1}$ decreases. Since turnover values of class 0 are larger than those of class 1 , the impact of varying values of $p_{00}$ is larger than for $p_{11}$.

With respect to the three estimation methods (see Figure 3.2) we found, similar as with the bias, that the standard error estimates for the statistics of interest from the EM bootstrap method were closest to the true values. For $\hat{T}_{1}$ and $\hat{\alpha}_{1}$, the standard error estimates from all three methods were almost equally good in most conditions. In the other conditions, the bootstrap method estimates were least accurate, followed by those of the SIMEX bootstrap method which were close to the true values. The estimated standard errors of $\hat{\sigma}_{1}$ were close to their true values in the case of the EM bootstrap method, much closer than for the other two methods. Larger values of the standard error of $\hat{\sigma}_{1}$ lead to larger estimation differences among the three methods.

### 3.3 Additional simulation studies

Below we summarise the results of three additional simulation studies.
Audit sample. In a preliminary study we have compared the estimation of bias and standard error with and without an audit sample (Li, 2020a). We tested $p_{00}$ and $p_{11}$ values of $0.6,0.75$ and $0.9, N=2,000, \alpha_{1}$ equal to $0.1,0.3,0.5,0.7$ and 0.9 ; a single Gaussian component in each class, with $\mu_{1}$ values of $2,10,12$ and $15, \mu_{0}$ fixed at $15, \sigma_{1}=1$ and $\sigma_{0}=2$. In situations without an audit sample, different starting values were tested for the EM algorithm. For $p_{00}$ and $p_{11}$, starting values of $0.6,0.75$ and 0.9 were used, leading to nine combinations for the pair $\left(p_{00}, p_{11}\right)$. By choosing different values for $p_{00}$ and $p_{11}$, the starting points can be seen as representative in the parameter space. The starting value of $\alpha_{1}$ was set according to $\alpha_{1}=\sum \hat{z}_{i} / N \times\left(3-p_{11}-p_{00}\right)-\left(1-p_{00}\right)$ which is an unbiased estimate of $\alpha_{1}$ (Kloos et al., 2021). Means and variances were started by robust statistics obtained from observed classes, where $\mu_{g}$ for the two classes was initialised at the median of the corresponding target variable, and $\sigma_{g}$ started with $k \times$ MAD where $k=1 / \Phi^{-1}(3 / 4) \approx 1.48$ and MAD is the median absolute deviation; see Rousseeuw and Croux (1993).

We found that the estimated bias of the statistics of interest using the EM bootstrap with and without an audit sample yielded nearly the same results, except for difficult estimation conditions. These difficult conditions were when $\mu_{0}=\mu_{1}$ combined with lower values for $\alpha_{1}$ and smaller $p_{00}$ and $p_{11}$ values (not shown). Under those conditions the bias estimates were less accurate, but the true (relative) bias was small. The bias estimates were close to zero while the true relative bias was up to 0.03 . For the standard error of the statistics of interest, the EM bootstrap with and without an audit sample yielded nearly the same results under all tested conditions.

Results for $\mu_{1}$. Besides the estimated bias and standard error of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ for $\alpha_{1}=0.3$, we have also computed them for $\hat{\mu}_{1}$. Note that $\hat{\mu}_{1}=\hat{T}_{1} /\left(\hat{\alpha}_{1} \times N\right)$, so results for $\hat{\mu}_{1}$ follow from those of $\hat{T}_{1}$ and $\hat{\alpha}_{1}$. With respect to the comparison of the three methods, we concluded that the true bias and standard error were estimated most accurately by the EM bootstrap method. The specific results can be found in Li (2020a).

Confidence intervals. So far, we have presented the results as averaged over $S_{0}=100$ sets of $\hat{z}$ that were generated from $z$. In a practical situation one would have only a single set of $\hat{z}$ values. This raises the question whether the EM bootstrap method also leads to more accurate bias estimates than the other two
methods in the case of a single sample or only on average. To that end, we estimated a $95 \%$ confidence interval for the bias estimates of the statistics of interest for the three methods using the $S_{0}=100$ replicates. This interval was estimated as 1.96 times the standard error of the estimated bias which in turn was derived from the $S_{0}=100$ bias estimates.

From Section 3.2 we have already concluded that in some settings all three methods yielded accurate bias estimates for $\hat{T}_{1}$ and $\hat{\alpha}_{1}$, but otherwise the bias estimate by the EM bootstrap method was clearly closer to the true value than that of the SIMEX bootstrap and bootstrap method, as averaged over $S_{0}$ replicates. Since the $95 \%$ confidence intervals were very small for all three methods (see Figure 3.3), in most settings also for a single set of $\hat{z}$ values, the EM bootstrap estimate of the bias was closer to the true bias than the estimates from the SIMEX bootstrap and bootstrap method.

Figure 3.3 Bias estimation in the simulation study with $\mathbf{9 5 \%}$ confidence interval.


Notes: For row and column settings see in Figure 3.1.

## 4. Case study

### 4.1 Data

In order to assess the performance of our methods in real applications, a case study was conducted. In the case study, the bias and variance of the estimated domain statistics $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ were estimated, using the same methods as in the simulation study.

For the case study, we started with a data set that contains the logarithm of the yearly turnover for a population of enterprises for which we know their website address; see Oosterveen (2020). The enterprises are classified by economic activity codes, according to the European NACE rev. 2 classification (Eurostat, 2008). The enterprises, their NACE codes and the website addresses were obtained from a statistical business register (SBR). For some enterprises we obtained the website address from a data set with URLs that we retrieved from an external company DataProvider.

These NACE code values are prone to classification errors. NACE codes of larger and more complex enterprises are checked manually and corrected if needed. For the case study we therefore limited ourselves to the simpler enterprises, that are composed of three legal units or fewer, since those are the enterprises whose NACE codes are the most prone to misclassification in practice. Furthermore, we started with a shortlist of 25 economic activities; the NACE codes are given in Figure 4.1. This shortlist contains a few groups of NACE codes with similar classes within a group, such as wholesale of clothes and retail trade of clothes, and dissimilar to others, such as taxi operation.

Similar to the simulation study, in the case study we wanted to start with a data set that was free of misclassifications, and then introduce the misclassifications on purpose in order to test the performance of the three methods. However, the data that we extracted from the SBR concerned observed NACE codes that could already contain misclassifications. We therefore did not use all smaller enterprises with a website ( 76,270 enterprises), but we used a selection of enterprises that had a relatively high probability of having the correct NACE code. This selection was made by predicting the NACE code of the enterprises based on the text of the main page of the website of the enterprise. These texts were scraped and preprocessed; see Oosterveen (2020) on how this was done. Three different machine learning algorithms (Naïve Bayes, Support Vector Machine and Random Forest) were trained to predict the NACE code, using a ten fold crossvalidation procedure. In each fold, the model was trained on $90 \%$ of the data and the fitted model was used to predict the remaining $10 \%$ of the data. An observed NACE code was considered to be correct when the fitted models of all three algorithms predicted the same code or when it was predicted by two of the algorithms and the prediction confidence of the models was relatively high; see Oosterveen (2020) for a more detailed description of the selection. This selection led to 45,965 enterprises.

In the present paper we limit ourselves to binary misclassifications. We therefore made a further selection of pairs of two NACE codes out of the shortlist of 25 NACE codes. We selected different pairs, where each selected pair is referred to as a "case". The pairs differed in the extent of overlap between the log-turnover distributions and in the shapes of the distributions. Furthermore, we ensured that the two group sizes were
not too small and not too unbalanced, which meant that $\alpha_{1}$ was not close to 0 or 1 . The full set of tests that we did can be found in Li (2020b). Here we present only the results for the two most interesting pairs which are given in Table 4.1.

Figure 4.1 Density distribution of $\log$ turnover for all groups.


Notes: The labels refer to NACE codes. NACE $=$ Nomenclature générale des activités économiques dans les Communautés européennes.

Table 4.2 gives some basic statistics of each of the four subpopulations defined by the NACE codes selected in Table 4.1: the size, total, mean and standard deviation of the log yearly turnover per enterprise. Case 1 has two well-separated distributions and a large number of enterprises per class. By contrast, in case 2 the two distributions are less well-separated and they have a smaller number of enterprises per class. Table 4.2 describes the true domain statistics $\zeta$ of each class.

Before we applied our methods, we removed outliers. We removed enterprises with a turnover value that was below $Q_{1}-1.5 \times \mathrm{IQR}$ or above $Q_{3}+1.5 \times \mathrm{IQR}$, where $Q_{1}$ is the first quartile of the enterprises in the true class, $Q_{3}$ is the third quartile and $\mathrm{IQR}=Q_{3}-Q_{1}$. Removal of outliers was only necessary in case 2 . We removed 17 outliers from NACE 4932 and 6 from NACE 8121.

Table 4.1
The selected groups and their allocations in the case study.

| NACE Code | Description of Economic Activity | Case | Class |
| :---: | :--- | :---: | :---: |
| 56101 | Restaurants | Case 1 | Class 1 |
| 96022 | Beauty treatment, pedicures and manicures, make-up and image consulting | Case 1 | Class 0 |
| 4932 | Taxi operation | Case 2 | Class 1 |
| 8121 | General cleaning of buildings | Case 2 | Class 0 |

Note: NACE = Nomenclature générale des activités économiques dans les Communautés européennes.

### 4.2 Settings

In contrast to the simulation study, here the number of components in the Gaussian mixture model is not known, and needs to be determined. To that end, we fitted a standard Gausian mixture model (i.e. without a misclassification component) for each class per case separately. We then determined the BIC and sBIC of these fitted models, since those measures could be used to select the number of components, see Section 2.2. For convenience, we fitted this Gausian mixture model to the true data rather than to the generated data with misclassifications, thereby avoiding that we had to run it for $100\left(S_{0}\right)$ sets times nine misclassification conditions (see below). In practice, one has to fit the standard Gaussian mixture model to observed data containing misclassifications, which is expected to result in slightly more components than when the model is run on true data. In Appendix B, the estimated optimal number of components for case 1 and 2 are shown according to the BIC and sBIC criteria. The optimal number of components in class 0 of case 2 was two using sBIC and one using BIC. For the other three classes no differences in the optimal number of components were found. For the remainder, we used the number of components according to the sBIC criterion, shown in Table 4.2.

Similar to the set up of the simulation study (Section 3.1), the observed classification variable $\hat{z}$ was generated from $z$ by using $\mathbf{P}$ (equation 2.3) with values of $p_{11}$ and $p_{00}$ of $0.6,0.75$, or 0.9 . For each setting of $p_{11}$ and $p_{00}$, we generated $S_{0}=100$ sets of $\hat{z}$ from $z$. In each set $s_{0}, 5 \%$ units from the population were randomly selected as the audit sample, which was used to estimate $\hat{p}_{11}$ and $\hat{p}_{00}$. These estimates were used as starting values for the EM algorithm. The final estimates of $\hat{p}_{11}$ and $\hat{p}_{00}$ by the EM algorithm were input for the bootstrap and SIMEX bootstrap methods. The number of iterations $S, S_{1}$ and $S_{2}$ was 100. The overall bias and variance estimates of the three methods were computed as the average over $S_{0}$ bias and variance estimates. As before, the estimated true bias and variance of the estimated domain parameters were based on 1,000 sets of $\hat{z}$.

Table 4.2
Domain statistics for each class in the case study.

| Case | Class | Number of <br> Components | Size | Total | Mean | Standard <br> Deviation |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
| Case 1 | Class 1 | 1 | 3,076 | 17,415 | 5.66 | 0.358 |
|  | Class 0 | 2 | 6,993 | 30,377 | 4.34 | 0.501 |
| Case 2 | Class 1 | 2 | 642 | 3,294 | 5.13 | 0.597 |

### 4.3 Results

The pattern of the bias (see Figures 4.2 and 4.3) and of the standard error (Figures 4.4 and 4.5) as a function of $p_{00}$ and $p_{11}$ is the same as has been found previously in the simulation study. For most settings tested, the estimation of the bias of $\hat{T}_{1}, \hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ was most accurate by the EM bootstrap method, followed by the SIMEX bootstrap method, while the bootstrap led to the least accurate results. An exception occurred in case $2, p_{00}=0.9$ for $\hat{\sigma}_{1}$ where the SIMEX bootstrap method was most accurate followed by the EM bootstrap method. In some of the settings with $p_{00}=0.75$ the EM bootstrap method and the SIMEX bootstrap method yielded near-identical results for the bias.

The bias estimates by the EM bootstrap method almost overlapped with the corresponding true values. Only in case 2 , the bias estimates of $\hat{\sigma}_{1}$ showed some distance from the true values, while the bias for $\hat{T}_{1}$ and $\hat{\alpha}_{1}$ remained accurate. It is seen in Figure 4.1 that the true distribution of $y$ in class 1 in this case (NACE code 4932) is more right-skewed than the other distributions considered in this study. This may particularly affect the statistic $\hat{\sigma}_{1}$, which is relatively sensitive to values in the right tail of the distribution. A follow-up analysis showed that the EM bootstrap method performed somewhat better in this example when the number of components per class was increased from two to three (see Figures B. 1 and B. 2 in Appendix B). This suggests that it may be beneficial to choose a relatively large number of mixture components if the distribution is known to be asymmetrical and if non-linear, non-robust parameters such as $\sigma_{1}$ are of interest. Note, however, that the statistic $\hat{\sigma}_{1}$ is not directly published as output in official statistics. Overall, the results suggest that the EM bootstrap method usually has accurate performance even in difficult situations, when the two classes contain fewer units and their distributions overlap.

Figure 4.2 Bias estimation in case 1.


Notes: Rows represent different domain parameters $\left(T_{1}, \alpha_{1}, \sigma_{1}\right)$. Columns represent different $p_{11}$ values, within each column different $p_{00}$ values are given.

Figure 4.3 Bias estimation in case 2.


Notes: For row and column settings see Figure 4.2.

Figure 4.4 Standard error estimation in case 1.


Notes: For row and column settings see Figure 4.2.

Figure 4.5 Standard error estimation in case 2.


Notes: For row and column settings see Figure 4.2.
Differences among the three methods with respect to the accuracy of the estimated standard error of $\hat{T}_{1}$, $\hat{\alpha}_{1}$ and $\hat{\sigma}_{1}$ were smaller than for the estimated bias. In settings where the estimates of the three methods were clearly different, most of the times the standard error estimate by the EM bootstrap method was closest to its true value, followed by the SIMEX bootstrap while the bootstrap method was the least accurate. Sometimes the estimated standard error of the SIMEX bootstrap was closest to the true standard error, for instance for case $1, p_{00}=0.6, p_{11}=0.9$ and target statistic $\hat{T}_{1}, \hat{\alpha}_{1}$ but differences with the EM bootstrap method were small.

## 5. Discussion

In this paper, we have proposed an EM bootstrap method for estimating the accuracy of domain statistics, in terms of bias and variance, in the presence of misclassifications. The use of an EM algorithm is not new in studies where classification errors occur. For instance, Sinclair and Hooker (2017) and Kosinski and Flanders (1999) both aim to estimate parameters of a model in the presence of classification errors in one or more of its variables. Our EM bootstrap method assumes a Gaussian mixture model. In our study, we do not use the mixture model to directly improve the accuracy of our (total or mean) estimates, but instead we use it to estimate the accuracy of a given estimator. The main reason is that National Statistical Institutes tend to avoid using model-based estimators directly for output, particularly if assumptions of the model are not verifiable (Van den Brakel and Bethlehem, 2008). However, using models to estimate the quality of output is accepted. An additional reason why we used the method to estimate the accuracy rather than to (directly) improve the accuracy is that the performance needed for the former is less than for the latter. Once
the bias and variance of the (given) estimates are obtained, one either concludes that those estimates are sufficiently accurate and publishes them, or one aims to first improve the accuracy of the estimates. In the latter case one might apply data editing first to reduce the rate of misclassifications. Alternatively, the mixture model could then be used to construct an improved estimate, if the model is considered sufficiently trustworthy.

We compared how well one can estimate the bias and variance of statistics in the presence of misclassification with the EM bootstrap method, the bootstrap method and the SIMEX bootstrap method using simulated data sets and real applications. The simulated data concerned (mixtures of) normal distributions whereas the real data concerned empirical distributions. For most of the conditions tested, we found that the EM bootstrap method outperformed the bootstrap and the SIMEX bootstrap method. The estimated bias of statistics from the EM bootstrap method was closer to the true bias than for the bootstrap and the SIMEX bootstrap method. The estimated variance of statistics based on the EM bootstrap method was also more accurate than for the other methods, but here the relative differences between the estimated values of the three methods were smaller than for the bias. Only in situations where the means of the two distributions were very close together, and populations were small the $\hat{\sigma}_{1}$ statistic was better estimated with SIMEX, but still the bias estimate of the total remained accurate.

Based on the results obtained we expect that the EM bootstrap outperforms the bootstrap for a binary variable as long as the model parameters are well estimated and the distribution of $y$ is captured well. We found that the EM algorithm only had difficulties in estimating the model parameters well when the two distributions had means that were close together while the population size was small. Still, under those difficult conditions the bias of the total was estimated well. We expect that the combination of similar class means and large standard deviations is also more problematic. The SIMEX method only works well when the parameter of interest is a smooth function of $\lambda$ that can be extrapolated well. Since our estimation method does not depend on these conditions we expect it to perform better than SIMEX for binary classification variables. When the distribution of $y$ is very skewed and contains a number of outliers, then the SIMEX algorithm might outperform the EM algorithm.

The above results were obtained by averaging over 100 simulation runs. By comparing the $95 \%$ confidence intervals of the bias estimated in the simulation study, we showed that even in a practical situation where there is only one single set of $\hat{z}$, the EM bootstrap method leads to better bias estimates than the bootstrap method and the SIMEX bootstrap method. We obtained these results, by using the classification error probabilities as estimated by the EM algorithm as input for the bootstrap and SIMEX bootstrap. That way, we gave the bootstrap and SIMEX bootstrap the best possible starting position to compete with the EM bootstrap. As an alternative, we also used estimated probabilities from an audit sample as input to the bootstrap and SIMEX bootstrap methods, which resulted in poorer results than before (not shown here).

Besides giving more accurate bias and variance estimates, the EM bootstrap method has two further advantages over the bootstrap and the SIMEX bootstrap method. The first advantage is that the EM bootstrap does not require that classification error probabilities $p_{11}$ and $p_{00}$ are accurately estimated beforehand; in fact, the EM algorithm provides estimates of these probabilities as part of its output. In previous studies these misclassification probabilities could only be obtained from audit samples (Van Delden et al., 2016).

The results of Li (2020a) and Li (2020b) suggest that the EM bootstrap also works well without an audit sample, except in some extreme cases (where two classes had the same mean). When there is no audit sample available, one should use multiple starting values to avoid finding a local maximum. If an audit sample is available, it is useful to incorporate it into the EM algorithm (Li, 2020a). In our study, the starting values obtained from the audit sample were sufficient to obtain accurate bias and variance estimates. The second advantage of the EM bootstrap method is that its output includes unit-specific probabilities $P\left(z_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=\sum_{j=1}^{q_{1}} A_{1 j i}$ and $P\left(z_{i}=0 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=\sum_{k=1}^{q_{0}} A_{0 k i}$, which could be used to predict the probability of a classification error in $\hat{z}_{i}$ for each unit in the data set. That outcome could subsequently be used to manually check and correct units with a potentially incorrect code, in an efficient way. This might save considerable time and effort compared to simply checking all units in a (sub)population based on $p_{11}$ and $p_{00}$.

There are two points of attention with respect to the practical use of the EM bootstrap method. First, the EM bootstrap method assumes that the actual distribution can be described with a Gaussian mixture model. According to McLachlan and Peel (2000) a Gaussian mixture model can accommodate various distributions for the continuous variable. In the case studies we found that the empirical distributions could already be approximated with 2-3 components. We do not expect that the Gaussian mixture assumption will pose great limits to its application in practice, although a caveat should be made that we did not test our method in situations that require more than three Gaussian components per class. Furthermore, in our case studies the number of components per class was determined using sBIC on the true data, whereas in practice it would have to be determined on data with misclassifications. The results in Section 4.3 suggest that, for skewed data, it could be beneficial in practice to err on the side of including too many components rather than too few. These matters could be investigated further.

Second, the number of iterations of the two loops will need to be determined with care since it will influence the accuracy of the bias and variance estimates of the EM bootstrap method. (This similarly holds for the other two methods.) In our study, we used a fixed number of iterations for the experiments. Based on the results, we judged that we had performed enough iterations to draw valid conclusions. In practice though, the number of iterations should be adjusted according to properties of the data sets, such as size, distribution of each class, distance between the two classes, etc. Efron and Tibshirani (1993) provide some theoretical considerations to take into account when choosing the number of iterations in a bootstrap algorithm; for instance, more iterations are usually needed for a smaller population size (op. cit., Section 6.4). In addition, it has been shown for similar nested algorithms that the number of iterations in the outer for-loop has the strongest effect on convergence (Chang and Hall, 2015), which suggests that increasing $S_{1}$ in Algorithm 2 is more beneficial than increasing $S_{2}$. However, as each application is different, it is good practice to examine the convergence of bootstrap estimates, e.g., by plotting intermediate results against the number of iterations to check when they become sufficiently stable. Finally, when one is only interested to estimate the bias, and not the variance, then the inner bootstrap loop within the EM bootstrap method is not needed, which saves computation time; see the "EM method" in Li (2020b).

In a future study, a number of extensions of our approach would be useful. A first, important, extension would be to generalise the bias and variance estimation to a situation with a classification variable with $D \geq 2$ classes. With $D$ classes, the probabilities of misclassification will be given by a $D \times D$ matrix $\mathbf{P}$
with $D(D-1)$ probabilities that are to be estimated. For each additional domain $d \in\{1, \ldots, D\}$, the number of other parameters in the mixture model increases linearly according to $\left(q_{d}-1\right)+2 q_{d}+1=3 q_{d}$ where $q_{d}$ is the number of components per additional domain $d$. Because the number of parameters increases with the number of domains the audit sample becomes more important to give the model reasonable starting values. It is also important to test whether such a $D$ class EM bootstrap converges well. Furthermore, it needs to be tested whether the mixture model performs better than the existing SIMEX and bootstrap methods for $D>2$.

A second extension would be to account for sample data rather than census data. In that case output quality is affected by both classification error and sampling error. In the case of simple random sampling the estimated bias will not be affected by the sampling error. For the variance one could use two approaches. One approach is that the sampling procedure is also bootstrapped by including it in the EM bootstrap procedure. Alternatively, a hybrid approach could be used in which the EM bootstrap estimate is used for the classification errors while an analytical expression is used for the sampling error.

A third extension would be to relax the assumptions for the probabilities of classification errors. In our study, the probabilities of making classification errors, $p_{11}$ and $p_{00}$, were assumed independent of the continuous variable. However, in real situations, this assumption does not always hold, at least not without conditioning on other covariates. It would therefore be interesting to make an extension in which the misclassification probabilities depend on covariates.

A fourth extension would be to use multiple numerical target variables, such as height and weight of patients in medical records. Then there will be more than one target variable in the general model. A multivariate Gaussian mixture model can be a suitable model for this case (McLachlan and Peel, 2000). A fifth possible extension is to take missing values in the target variable(s) into account.

Finally, we note that Algorithm 1 in our study is a standard, single bootstrap method. In the literature, double and higher-order bootstraps have also been proposed as a way of correcting for bias in the bias and variance estimators from a single bootstrap (Chang and Hall, 2015; Hall and Martin, 1988). These methods do not require an explicit model for the data but, like the single bootstrap, they do require (an estimate of) the matrix $\mathbf{P}$ as input. It would be interesting to compare the performance of the EM bootstrap method and a double bootstrap in a future study, in particular for the extended problem with $D \gg 2$ classes and/or several numerical target variables, for which estimating a Gaussian mixture model may become challenging.

## Acknowledgements

The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors would like to thank the Associate Editor and two anonymous referees for their constructive and detailed comments on a draft version of this article.

## Appendix

## A. Extrapolation function in the SIMEX bootstrap method

In this appendix, three extrapolation functions used in the SIMEX bootstrap method are compared: a local polynomial regression (LOESS), a second order polynomial regression (poly2) and a third order
polynomial regression (poly3). For the LOESS function we used default settings of the loess function in R $($ span $=0.75$ and nls.control $($ maxiter $=1000)$ ).

The bias and standard error estimates from the LOESS function and of the third order polynomial regression were similar, see Figures A. 1 and A.2. The estimates of both models were closer to the true values than those of the second order polynomial regression. Considering that the LOESS function had been used in a previous study on misclassifications (Hopkins and King, 2010), we decided to apply the LOESS function in the present paper.

Figure A. 1 Comparison of the bias estimation performance of the three extrapolation functions in the SIMEX bootstrap method.


Notes: Rows represent different domain parameters $\left(T_{1}, \alpha_{1}, \mu_{1}, \sigma_{1}\right)$ for two settings of $\alpha_{1}: \alpha_{1}=0.3$ and $\alpha_{1}=0.5$. Columns represent different $p_{11}$ values, within each column different $p_{00}$ values are given.

Figure A. 2 Comparison of the standard error estimation performance of the three extrapolation functions in the SIMEX bootstrap method.


Notes: For row and column settings see Figure A.1.

## B. BIC vs. sBIC

Before fitting the Gaussian mixture model, one must select its number of components. In this appendix, we compare the bias and the standard error estimation outcomes when using the BIC versus the sBIC as criterion to select the number of components. The sBIC is more justified when the components of the mixture model have very similar means and variance; see Section 2.2.1.

Table B. 1 shows the optimal number of components selected by the BIC and sBIC criteria. The two criteria led to the same number of components for both classes of case 1 and for class 1 of case 2 . For class 0 of case 2, the optimal number of components selected by sBIC was two and by BIC it was one. Hence,
only for case 2, we compared the bias and the standard error estimates, using two (sBIC) or one (BIC) component for class 0 .

Table B. 1
The optimal number of components selected by BIC and sBIC.

| Case No. | Class | Optimal Number |  |
| :--- | :--- | :---: | :---: |
|  |  | BIC | sBIC |
| Case 1 | Class 1 | 1 | 1 |
| Case 2 | Class 0 | 2 | 2 |
|  | Class 1 | 2 | 2 |

Using two components for class 0 led to more accurate bias estimates than using one component; see Figure B.1. For some of the settings the standard error estimates based on two components were also more accurate than those based on one component, see Figure B.2, although the accuracy differences were smaller for the standard error than for the bias. We therefore decided to use the sBIC in our case study to estimate the optimal number of components.

For case 2, we also tried a model with three components in both classes instead of two. The resulting bias estimates and standard errors are also shown in Figures B. 1 and B. 2 ("comp3"). As mentioned in Section 4.3, it is seen that increasing the number of components to three led to more accurate results.

Figure B. 1 Bias estimation using BIC vs. sBIC in case 2.


Notes: For row and column settings see Figure A.1.

Figure B. 2 Standard error estimation using BIC vs. sBIC in case 2.


Notes: For row and column settings see Figure A.1.

## C. Theoretical properties of Algorithms 1 and 2

## C. 1 Algorithm 1: The bootstrap method

In general, Algorithm 1 yields biased estimates of the bias and variance of $\hat{\zeta}$. We will illustrate this using the estimated domain total $\hat{T}_{1}$ as an example. Note that the results below also apply to $\hat{\alpha}_{1}$, since it is obtained as a special case of $\hat{T}_{1}$ with $y_{i} \equiv 1 / N$.

For $\hat{T}_{1}$, it can be derived that its true bias equals

$$
\begin{equation*}
\operatorname{Bias}\left(\hat{T}_{1}\right)=\left(1-p_{00}\right) T_{0}+\left(p_{11}-1\right) T_{1}, \tag{C.1}
\end{equation*}
$$

whereas, for $S \rightarrow \infty$, the bootstrap bias estimator from Algorithm 1 converges to $\left(1-p_{00}\right) \hat{T}_{0}+\left(p_{11}-1\right) \hat{T}_{1}$; see, e.g., Burger, van Delden and Scholtus (2015). Thus, in general, the bootstrap bias estimator is biased unless $\hat{T}_{1}$ itself happens to be an unbiased estimator of $T_{1}$. The latter situation occurs only for particular combinations of ( $p_{11}, p_{00}$ ) (Kloos et al., 2021).

Similarly, it can be derived that the true variance of $\hat{T}_{1}$ is

$$
\begin{equation*}
\operatorname{Var}\left(\hat{T}_{1}\right)=p_{00}\left(1-p_{00}\right) K_{0}+p_{11}\left(1-p_{11}\right) K_{1}, \tag{C.2}
\end{equation*}
$$

with $K_{1}=\sum_{i=1}^{N} z_{i} y_{i}^{2}$ and $K_{0}=\sum_{i=1}^{N}\left(1-z_{i}\right) y_{i}^{2}$, whereas the bootstrap variance estimator converges to the same expression with $K_{1}$ and $K_{0}$ replaced by $\hat{K}_{1}=\sum_{i=1}^{N} \hat{z}_{i} y_{i}^{2}$ and $\hat{K}_{0}=\sum_{i=1}^{N}\left(1-\hat{z}_{i}\right) y_{i}^{2}$, respectively
(Burger et al., 2015). Note that, in general when $T_{1} \not \neq \alpha_{1}$, the conditions under which the bootstrap bias estimator and bootstrap variance estimator are unbiased are not equivalent.

## C. 2 Algorithm 2: The EM bootstrap method

As suggested in Section 2.3.2, the purpose of generating $0-1$-values $\tilde{z}_{i}$ in the outer for-loop of Algorithm 2, with $P\left(\tilde{z}_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=P\left(z_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right)=\sum_{j=1}^{q_{1}} A_{1 j i}$, is to correct the bootstrap bias and variance estimators for the bias that occurs in Algorithm 1. To illustrate the underlying idea, we will show here that in the case of the estimated domain total $\hat{T}_{1}$, for which exact analytical expressions are available, bias correction is indeed achieved by both $\mathbf{B i a s}_{\text {comb }}$ and $\mathbf{V a r}_{\text {comb }}$ provided that the assumed mixture model holds. A simplified version of the derivation below can be given for the estimated proportion $\hat{\alpha}_{1}$.

Denote $\tilde{T}_{1}=\sum_{i=1}^{N} \tilde{z}_{i} y_{i}$ and $\tilde{T}_{0}=\sum_{i=1}^{N}\left(1-\tilde{z}_{i}\right) y_{i}$; also denote $\tilde{T}_{1}^{*}=\sum_{i=1}^{N} \tilde{z}_{i}^{*} y_{i}$ and $\tilde{T}_{0}^{*}=\sum_{i=1}^{N}\left(1-\tilde{z}_{i}^{*}\right) y_{i}$. It can be derived analogously to (C.1) that $\mathrm{E}\left(\tilde{T}_{1}^{*}-\tilde{T}_{1} \mid \hat{z}, \tilde{z}\right)=\left(1-p_{00}\right) \tilde{T}_{0}+\left(p_{11}-1\right) \tilde{T}_{1}$. Hence, for $S_{1}, S_{2} \rightarrow \infty$, the expected value of the bias estimator in the EM bootstrap method is

$$
\begin{equation*}
\mathrm{E}\left\{\operatorname{Bias}_{\text {comb }}\left(\hat{T}_{1}\right)\right\}=\left(1-p_{00}\right) \mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{T}_{0} \mid \hat{z}\right)\right\}+\left(p_{11}-1\right) \mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{T}_{1} \mid \hat{z}\right)\right\} . \tag{C.3}
\end{equation*}
$$

Furthermore, it is seen that

$$
\mathrm{E}_{\hat{z}}\left(\tilde{T}_{1} \mid \hat{z}\right)=\sum_{i=1}^{N} \mathrm{E}\left(\tilde{z}_{i} \mid \hat{z}_{i}\right) y_{i}=\sum_{i=1}^{N} P\left(z_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right) y_{i}=\sum_{i=1}^{N} \sum_{j=1}^{q_{1}} A_{1 j i} y_{i}=N \hat{\alpha}_{1} \sum_{j=1}^{q_{1}} \hat{\xi}_{1 j} \hat{\mu}_{1 j},
$$

where the last expression follows from the formulas applied during the M step of the EM algorithm (see Section 2.2.2). Under the assumption that the mixture model holds, it follows that

$$
\mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{T}_{1} \mid \hat{z}\right)\right\}=\mathrm{E}\left(N \hat{\alpha}_{1} \sum_{j=1}^{q_{1}} \hat{\xi}_{1 j} \hat{\mu}_{1 j}\right) \approx N \alpha_{1} \sum_{j=1}^{q_{1}} \xi_{1 j} \mu_{1 j}=N \alpha_{1} \mu_{1}=T_{1} ;
$$

cf. note 2 at Table 2.1 for the final two equalities. Similarly, it can be shown that $\mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{T}_{0} \mid \hat{z}\right)\right\} \approx T_{0}$ if the model holds. Substituting both results into (C.3) and recalling (C.1), we conclude that $\mathrm{E}\left\{\boldsymbol{\operatorname { B i a s }}_{\text {comb }}\left(\hat{T}_{1}\right)\right\} \approx$ $\operatorname{Bias}\left(\hat{T}_{1}\right)$.

For the variance estimator, we can proceed in a similar fashion. Denote $\tilde{K}_{1}=\sum_{i=1}^{N} \tilde{z}_{i} y_{i}^{2}$ and $\tilde{K}_{0}=$ $\sum_{i=1}^{N}\left(1-\tilde{z}_{i}\right) y_{i}^{2}$. Analogously to (C.2) it can be shown that $\operatorname{Var}\left(\tilde{T}_{1}^{*} \mid \hat{z}, \tilde{z}\right)=p_{00}\left(1-p_{00}\right) \tilde{K}_{0}+p_{11}\left(1-p_{11}\right) \tilde{K}_{1}$. Hence, for $S_{1}, S_{2} \rightarrow \infty$,

$$
\begin{equation*}
\mathrm{E}\left\{\operatorname{Var}_{\text {comb }}\left(\hat{T}_{1}\right)\right\}=p_{00}\left(1-p_{00}\right) \mathrm{E}\left\{\mathrm{E}_{\tilde{z}}\left(\tilde{K}_{0} \mid \hat{z}\right)\right\}+p_{11}\left(1-p_{11}\right) \mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{K}_{1} \mid \hat{z}\right)\right\} . \tag{C.4}
\end{equation*}
$$

From the formulas applied during the EM algorithm, it follows that

$$
\mathrm{E}_{\hat{z}}\left(\tilde{K}_{1} \mid \hat{z}\right)=\sum_{i=1}^{N} \mathrm{E}\left(\tilde{z}_{i} \mid \hat{z}_{i}\right) y_{i}^{2}=\sum_{i=1}^{N} P\left(z_{i}=1 \mid y_{i}, \hat{z}_{i} ; \hat{\boldsymbol{\theta}}\right) y_{i}^{2}=\sum_{i=1}^{N} \sum_{j=1}^{q_{1}} A_{1 j i} y_{i}^{2}=N \hat{\alpha}_{1} \sum_{j=1}^{q_{1}} \hat{\xi}_{1 j}\left(\hat{\sigma}_{1 j}^{2}+\hat{\mu}_{1 j}^{2}\right) .
$$

Hence, assuming that the mixture model holds, we obtain:

$$
\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}\left(\tilde{K}_{1} \mid \hat{z}\right)\right\} \approx N \alpha_{1} \sum_{j=1}^{q_{1}} \xi_{1 j}\left(\sigma_{1 j}^{2}+\mu_{1 j}^{2}\right)=N \alpha_{1}\left(\sigma_{1}^{2}+\mu_{1}^{2}\right)=K_{1} .
$$

In the same way, it can be derived that $\mathrm{E}\left\{\mathrm{E}_{\hat{z}}\left(\tilde{K}_{0} \mid \hat{z}\right)\right\} \approx K_{0}$. Thus, it is seen using $(\mathrm{C} .2)$ that $\mathrm{E}\left\{\operatorname{Var}_{\text {comb }}\left(\hat{T}_{1}\right)\right\} \approx$ $\operatorname{Var}\left(\hat{T}_{1}\right)$ if the mixture model holds.

## References

Bakker, B.F.M., Van Rooijen, J. and Van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. Statistical Journal of the IAOS, 30, 411-424.

Bross, I. (1954). Misclassification in $2 \times 2$ tables. Biometrics, 10, 478-486.
Buonaccorsi, J.P. (2010). Measurement Error: Models, Methods and Applications. Chapman and Hall/CRC Press.

Burger, J., van Delden, A. and Scholtus, S. (2015). Sensitivity of mixed-source statistics to classification errors. Journal of Official Statistics, 31(3), 489-506.

Chang, J., and Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. Biometrika, 102, 203-214.

Cook, J.R., and Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association, 89(428), 1314-1328.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

Drton, M., and Plummer, M. (2017). A Bayesian information criterion for singular models. Journal of the Royal Statistical Society, 79(2), 323-380.

Edwards, J.K., Bakoyannis, G., Yiannoutsos, C.T., Mburu, M.W. and Cole, S.R. (2019). Non-parametric estimation of the cumulative incidence function under outcome misclassification using external validation data. Statistics in Medicine, 38(29), 5512-5527.

Edwards, J.K., Cole, S.R. and Fox, M.P. (2020). Flexibly accounting for exposure misclassification with external validation data. American Journal of Epidemiology, 189(8), 850-860. Retrieved from https://doi.org/10.1093/aje/kwaa011 doi: 10.1093/aje/kwaa011.

Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. London: Chapman and Hall/CRC.

Eurostat (2008). NACE Rev. 2 Statistical Classification of Economic Activities in the European Community. Retrieved from https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_ DTL\&StrNom=NACE_REV2.

Eurostat (2009). ESS Handbook for Quality Reports (Tech. Rep.). Office for Official Publications of the European Communities, Methodologies and Working papers. Retrieved from https://ec.europa.eu/ eurostat/web/products-manuals-and-guidelines/-/ks-ra-08-016.

Eurostat (2022). Table: In-Work at-Risk-of-Poverty Rate by Educational Attainment Level - EU-SILC survey. Retrieved from https://ec.europa.eu/eurostat/databrowser/view/ilciw04/default/table?lang=en.

Gravel, C.A., and Platt, R.W. (2018). Weighted estimation for confounded binary outcomes subject to misclassification. Statistics in Medicine, 37(3), 425-436.

Hall, P., and Martin, M.A. (1988). On bootstrap resampling and iteration. Biometrika, 75, 661-671.

Hopkins, D.J., and King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229-247.

Keogh, R.G., Shaw, P.A., Gustafson, P., Carroll, R.J., Deffner, V., Dodd, K.W., Küchenhoff, H., Tooze, J.A., Wallace, M.P., Kipnis, V. and Freedman, L.S. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-Basic theory and simple methods of adjustments. Statistics in Medicine, 39(16), 2197-2231. doi: 10.1002/sim. 8532.

Kloos, K., Meertens, Q., Scholtus, S. and Karch, J. (2021). Comparing correction methods to reduce misclassification bias. In Proceedings of BNAIC/BeneLearn, (Eds., L. Cao, W.A. Kosters and J. Lijffijt), 103-129. Springer, Leiden. Retrieved from https://bnaic.liacs.leidenuniv.nl/wordpress/wp-content/ uploads/papers/BNAICBENELEARN_2020_Final_paper_64.pdf.

Kooiman, P., Willenborg, L. and Gouweleeuw, J. (1997). PRAM: A Method for Disclosure Limitation of Microdata. Research paper no. 9705. Voorburg/Heerlen: Statistics Netherlands.

Kosinski, A.S., and Flanders, W.D. (1999). Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: A regression approach. Statistics in Medicine, 18(20), 2795-2808.

Küchenhoff, H., Mwalili, S.M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. Biometrics, 62(1), 85-96.

Li, Y. (2020a). Bias Correction for Classification Errors. Retrieved from https://www.researchgate.net/ publication/362862838_Bias_Correction_for_Classification_Errors (Internship Report, Leiden University).

Li, Y. (2020b). Estimating the Effect of Classification Errors on Domain Statistics (Master's thesis, Leiden University). Retrieved from https://hdl.handle.net/1887/3280845.

Little, R.J., and Rubin, D.B. (2002). Statistical Analysis with Missing Data (2 ${ }^{\text {nd }}$ ed., Vol. 793). New York: John Wiley \& Sons, Inc.

Magnusson, P., Palm, A., Branden, E. and Mörner, S. (2017). Misclassification of hypertrophic cardiomyopathy: Validation of diagnostic codes. Clinical Epidemiology, 9, 403.

McLachlan, G.J., and Peel, D. (2000). Finite Mixture Models. New York: John Wiley \& Sons, Inc.

Meertens, Q., Diks, C., Van den Herik, H. and Takes, F. (2020). A data-driven supply-side approach for estimating cross-border internet purchases within the European Union. Journal of the Royal Statistical Society: Series A (Statistics in Society), 183(1), 61-90.

Oosterveen, V. (2020). Notice the Noise: Detecting Misclassifications in Register Data (Master's thesis, Utrecht University, the Netherlands). Retrieved from https://www.researchgate.net/publication/ 354533688_Notice_the_noise_detecting_misclassifications_in_register_data.

Rousseeuw, P.J., and Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88(424), 1273-1283.

Schwartz, J.E. (1985). The neglected problem of measurement error in categorical data. Sociological Methods and Research, 13(4), 435-466.

Selén, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. Journal of the American Statistical Association, 81, 75-81.

Shaw, P.A., Gustafson, P., Carroll, R.J., Deffner, V., Dodd, K.W., Keogh, R.H., Kipnis, V., Tooze, J.A., Wallace, M.P., Küchenhoff, H. and Freedman, L.S. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2-More complex methods of adjustment and advanced topics. Statistics in Medicine, 39(16), 2232-2263. doi: 10.1002/sim.8531.

Sinclair, D.G., and Hooker, G. (2017). An Expectation Maximization algorithm for high-dimensional model selection for the Ising model with misclassified states. arXiv preprint arXiv:1704.05995. Retrieved from https://arxiv.org/abs/1704.05995.

United Nations (2015). Guidelines on Statistical Business Registers. New York and Geneva: United Nations.

Van Delden, A., Scholtus, S. and Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. Journal of Official Statistics, 32(3), 619-642. Retrieved from https://content. sciendo.com/view/journals/jos/32/3/article-p619.xml doi: https://doi.org/10.1515/jos-2016-0032.

Van Delden, A., Scholtus, S., Burger, J. and Meertens, Q.A. (2023). Accuracy of estimated ratios as affected by dynamic classification errors. Journal of Survey Statistics and Methodology, 11(4), 942-966. doi: https://doi.org/10.1093/jssam/smac015.

Van den Brakel, J., and Bethlehem, J. (2008). Model-Based Estimation for Official Statistics (Tech. Rep.). Retrieved from https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-officialstatistics.

Van den Hout, A., and Van der Heijden, P.G.M. (2002). Randomised response, statistical disclosure control and misclassification: A review. International Statistical Review, 70(2), 269-288.

Zhang, L.C. (2011). A unit-error theory for register-based household statistics. Journal of Official Statistics, 27(3), 415-432.

# Model-based stratification of payment populations in Medicare integrity investigations 

Don Edwards, Piaomu Liu and Alexandria Delage ${ }^{1}$


#### Abstract

When a Medicare healthcare provider is suspected of billing abuse, a population of payments $X$ made to that provider over a fixed timeframe is isolated. A certified medical reviewer, in a time-consuming process, can determine the overpayment $Y=X-$ (amount justified by the evidence) associated with each payment. Typically, there are too many payments in the population to examine each with care, so a probability sample is selected. The sample overpayments are then used to calculate a $90 \%$ lower confidence bound for the total population overpayment. This bound is the amount demanded for recovery from the provider. Unfortunately, classical methods for calculating this bound sometimes fail to provide the $90 \%$ confidence level, especially when using a stratified sample. In this paper, 166 redacted samples from Medicare integrity investigations are displayed and described, along with 156 associated payment populations. The 7,588 examined $(Y, X)$ sample pairs show (1) Medicare audits have high error rates: more than $76 \%$ of these payments were considered to have been paid in error; and (2) the patterns in these samples support an "All-or-Nothing" mixture model for ( $Y, X$ ) previously defined in the literature. Model-based Monte Carlo testing procedures for Medicare sampling plans are discussed, as well as stratification methods based on anticipated model moments. In terms of viability (achieving the $90 \%$ confidence level) a new stratification method defined here is competitive with the best of the many existing methods tested and seems less sensitive to choice of operating parameters. In terms of overpayment recovery (equivalent to precision) the new method is also comparable to the best of the many existing methods tested. Unfortunately, no stratification algorithm tested was ever viable for more than about half of the 104 test populations.


Key Words: Medicare fraud; All-or-nothing mixture model; Dalenius-Hodges stratification; Anticipated moments; Neyman allocation.

## 1. Introduction

According to the U.S. Centers for Medicare and Medicaid Services (CMS) 2022 Trustees Report (https://www.cms.gov/files/document/2022-medicare-trustees-report.pdf), the Medicare Trust Fund Hospital Insurance trust is estimated to be depleted by 2028. In an effort to extend the depletion date, over the last decade the CMS and the legislatures have focused efforts on eliminating fraud, improving quality and reducing overall cost of care (Huffman, 2021; Salmond and Echevarria, 2017). The CMS uses the Comprehensive Error Rate Testing (CERT, https://www.cms.gov/Research-Statistics-Data-and-Systems/ Monitoring-Programs/Improper-Payment-Measurement-Programs/CERT) program to estimate the Medicare Fee-for-Service (FFS) program's improper payment rate each year by sampling claims to determine whether they were paid properly under Medicare coverage, coding and payment rules. In 2022, the improper payment rate was estimated at 7.46 percent, representing $\$ 31.46$ billion in improper payments. It is clear that even with the focused effort on eliminating fraud, waste, and abuse from the Medicare FFS program that there are still program integrity concerns related to improperly paid claims, and that these will continue (Clemente, McGrady, Repass, Paul III and Coustasse, 2018).

In this paper, a "provider" is any entity that bills Medicare: physicians, home health providers, hospitals, hospice providers, durable medical goods providers, ambulance services, etc. The CMS uses contractors at multiple levels to investigate providers suspected of abusing the system. The current guidelines governing the investigation process are given in the Medicare Program Integrity Manual (MPIM), Chapter 8: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/pim83c08.pdf

When there is reason to suspect that a provider is billing improperly to Medicare FFS, a Unified Program Integrity Contractor (UPIC), often a subsidiary of an insurance company, conducts a sampling investigation. First, the UPIC obtains detailed information on all Medicare claims paid to that provider for a specified period, usually 1-2 years: a population of payments. A sampling unit is chosen and a probability sample is designed and implemented. A Certified Medical Reviewer then examines the evidence in support of each sampled payment, obtaining the overpayment amount Y as

$$
Y=\text { overpayment }=(\text { amount paid } \mathrm{X})-(\text { amount justified by the evidence }) .
$$

Note that since the amount justified by the evidence is non-negative, the overpayment amount is bounded above by the payment amount. The medical review usually requires several months for a sample of moderate size such as those displayed in Section 2. The UPIC then uses the overpayment amounts to "extrapolate": the UPIC calculates a $90 \%$ lower confidence bound for the total overpayment made to the provider over the specified time period. This lower bound is the amount demanded for recovery from the provider. (If the point estimate achieves "high precision", it can be used as the demand amount. This is very rare).

The sampling unit in these investigations is usually either the paid claim or all claims paid for services to a particular beneficiary (more precisely, to a Medicare identification number, formerly called a Health Insurance Claim Number HICN). The sampling plans are nearly always simple random samples or stratified random samples with strata determined by payment amounts. The extrapolation method is usually the "mean per unit" approximate method based on the finite population Central Limit Theorem (Hájek, 1964; Li and Ding, 2017).

Specifically, consider a stratified random sample with L strata, denote stratum sizes by $N_{l}$ and sample sizes by $n_{l}$, for $l=1,2, \ldots, L$. Let $N=\sum_{l=1}^{L} N_{l}$ and $n=\sum_{l=1}^{L} n_{l}$. For each $l$, let $W_{l}=N_{l} / N$ and let $\bar{Y}_{l}$ and $S_{l}$ be the sample mean and standard deviation of stratum $l$ overpayments, respectively. The CLT-based extrapolation has the form

$$
\begin{equation*}
N \sum_{l=1}^{L} W_{l} \bar{Y}_{l}-t_{v} N \sqrt{\sum_{l=1}^{L} W_{l}^{2}\left(\frac{S_{l}^{2}}{n_{l}}\right)\left(1-\frac{n_{l}}{N_{l}}\right)} \tag{1.1}
\end{equation*}
$$

where $t_{v}$ is the $90^{\text {th }}$ percentile of Student's t -distribution with $v$ degrees of freedom. Some contractors use the approximate degrees of freedom due to Satterthwaite (1946), though some use the less conservative choices $(n-L)$ or infinite degrees of freedom (i.e., the standard Normal percentile point). The Satterthwaite degrees of freedom are at most $(n-L)$, so it is the most conservative of the three choices. For simple random sampling, equation (1.1), suppressing subscript $l$, reduces to

$$
\begin{equation*}
N \bar{Y}-t_{n-1} N\left(\frac{S}{\sqrt{n}}\right) \sqrt{1-\frac{n}{N}} . \tag{1.2}
\end{equation*}
$$

The provider has the right to appeal the reviewer's overpayment determinations and the sampling methods employed by the UPIC. In this event, an independent Medicare Administrative Contractor (MAC) conducts a review of the sampling plan and overpayment determinations as the first level of appeal. If the provider further appeals the MAC's decision, a Qualified Independent Contractor conducts another review as the second level of appeal. If the provider is still unsatisfied, the matter can be appealed to the third level and heard in a formal Administrative Law Judge hearing. Further appeals are also possible. At any level of appeal, the provider may submit additional documentation to support payment on their claims.

Section 2 of this paper examines the raw data for 166 redacted samples examined by a single reviewer at a MAC during the period 2013-2020. This data motivates a statistical model for Medicare sample overpayments, discussed in Section 3. Section 4 discusses appropriate Monte Carlo testing methods for sampling plans given the model. Section 5 reviews existing methods for stratifying populations and proposes a new class of stratification methods. Section 6 gives the results of a Monte Carlo efficiency study comparing these stratification methods, conducted using 104 of the payment populations.

## 2. Honor the data: An exploration of $\mathbf{1 6 6}$ Medicare samples

The redacted samples, in rough chronological order 2013-2020, are numerically summarized and graphically displayed using scatterplots of overpayment versus payment amount at: https://drive.google. com/drive/folders/1CIzQKzN4-RIIY38WonSAU7ydICwb76wh. The reader is encouraged to take a few minutes to page through the plots to form his/her own opinions on patterns in the data. The investigations summarized here were conducted by four different program integrity contractors. Sixty-eight percent (113) of the samples were simple random samples and the remaining $32 \%$ (53) were stratified by payment amount. Total sample sizes varied between 25 and 159 with $90 \%$ of the samples having between 30 and 70 total payments. Seventy-five percent (124) of these samples were from home health providers, $16 \%$ (27) were from hospice providers, and the remaining $9 \%$ (15) included physicians, skilled nursing facilities, ambulance services, etc. Nine of the samples used the beneficiary's HICN as the sampling unit, 149 used the paid claim, and 8 were drawn and analyzed as pennysamples (Edwards, Gilliland, Ward-Besser and Lasecki, 2015).

Define the "taint" for a sample payment as the ratio of overpayment to payment. Figure 2.1 provides a schematic classifying the 7,588 sample points from the 166 samples into four zones based on taint value: Zone 1 (taint $\geq 0.95$ ), Zone $0(0 \leq$ taint $\leq 0.05)$, the Negative Zone (taint $<0$ ), and the Partial Zone. Some observations:

1. More than $76 \%$ of these payments were adjudged to have been paid either partially or totally in error. The conventional "wisdom" that auditing error rates are typically low does not apply to Medicare investigations.
2. Only one negative overpayment occurred in 7,588 reviewed payments.
3. Except for "complete error" samples (overpayment equals payment, 24 samples), overpayment amount is not linearly related to payment amount in any of these samples, though the two variables are sometimes highly correlated (the important distinction between "correlated" and "linearly related", so eloquently sung by Anscombe's (1973) quartet, is unfortunately blurred at times in the literature, and by some data analysts). A handful of samples have numerous partial overpayments, in which case the overpayment-payment relationship is essentially formless within the triangle formed by Zones 0 and 1 . For the majority of these samples, though, the relationship is best described as approximately bilinear, with a horizontal line in Zone 0 , a $45^{\circ}$ line in Zone 1, and possibly a handful of partial overpayments in between.
4. It is less obvious to the naked eye, but there is no evidence in this data that larger payments are more likely to have larger taints. Figure 2.2, admittedly crowded, shows a plot of taint versus the logarithm base 10 of payment amount, showing no obvious monotone trend. Figure 2.3 displays the results of a generalized additive model fit of taint on $\log _{10}$ (payment) using a separate intercept for each sample and a common smooth function for the taint vs. $\log _{10}$ (payment) relationship. The figure includes Scheffé-style simultaneous $95 \%$ confidence bands for the true regression function; these bands include a horizontal line. Analogous results were obtained using payment amounts instead of $\log _{10}$ (payment amounts) as regressor.

Figure 2.1 A classification of the $\mathbf{7 , 5 8 8}$ overpayment-payment pairs in the $\mathbf{1 6 6}$ samples.


Figure 2.2 Taint versus $\log _{10}$ (Payment amount) for 7,588 sample points.


Figure 2.3 Graphical summary of a generalized additive model analysis of taint vs. $\log _{10}$ (payment). Gray shading is a $95 \%$ Scheffé-style confidence region for the true nonparametric regression function.


## 3. A model for Medicare sample data

The patterns in the 166 samples suggest a simple mixture model for Medicare sample data. This model has previously been discussed by several authors, including King (1996), Edwards, Ward-Besser, Lasecki, Parker, Wieduwilt, Wu and Moorhead (2003), and King and Madansky (2013). The population payments $X_{i}, i=1,2, \ldots, N$ are known constants. Let $Z_{i}=1$ with probability $P_{E}, 0 \leq P_{E} \leq 1$, and $Z_{i}=0$ otherwise, regardless of the value of $X_{i}$. The "All or Nothing" mixture model with error rate $P_{E}$, here abbreviated AN $\left(P_{E}\right)$, has overpayment $Y_{i}=Z_{i} X_{i}$. That is, with probability $P_{E}$ the overpayment equals the payment, and
with probability $1-P_{E}$ the overpayment is zero. The majority of the 166 samples discussed in Section 2 conform to this model with negligible numbers of partial or negative overpayments.

Let $\mu_{X}$ and $\sigma_{X}^{2}$ denote the known population mean and variance of payment amounts. Under the AN $\left(P_{E}\right)$ model it is straightforward to derive the mean and variance of overpayment Y for a randomly selected X (King, 1996):

$$
\begin{align*}
E(Y) & =\mu_{Y}=P_{E} \mu_{X} \\
\operatorname{Var}(Y) & =\sigma_{Y}^{2}=\mu_{X}^{2} P_{E}\left(1-P_{E}\right)+P_{E} \sigma_{X}^{2} . \tag{3.1}
\end{align*}
$$

Note that the variance of the overpayment amounts can be dramatically different from the variance of the payment amounts. To illustrate, Figure 3.1 shows histograms of three payment populations. We include 0 on the horizontal axis of these histograms in order to help envision the shape of the corresponding overpayment population under a given error rate $P_{E}$, as explained below. For population B , for values of $P_{E}$ near $1 / 2$, the variance of overpayment amounts $\sigma_{Y}^{2}$ is more than 800 times the variance of payment amounts $\sigma_{X}^{2}$. For populations C and A the ratio $\sigma_{Y}^{2} / \sigma_{X}^{2}$ reaches a maximum of 4.7 and 1.5 respectively. Hence, the common practice of using the variance of payment amounts as an estimate of the variance of overpayment amounts in sample-size determination formulas has no relevance except when $P_{E}$ is very near to 1.

Figure 3.1 Example payment populations.


Notes: (A) $\mathrm{N}=59,804$ hospice HICN payments; (B) $\mathrm{N}=249$ power wheelchair claim payments; (C) $\mathrm{N}=570$ hospice claim payments.

## 4. Monte Carlo testing of sample designs under the AN model

Consider a sampling-and-extrapolation plan proposed for a particular payment population. The $\operatorname{AN}\left(P_{E}\right)$ model motivates a simple Monte Carlo testing procedure for the plan, implemented by the freely-distributed $\mathbf{R}$ ( $\mathbf{R}$ Core Team, 2021) function samptest available from the authors. The testing proceeds as follows: for each choice of $P_{E}$ on a grid spanning the interval ( 0,1 ],

1. A plausible overpayment population is created by randomly choosing a proportion $P_{E}$ to have overpayment $=$ payment, leaving the remaining proportion $1-P_{E}$ to have zero overpayment.
2. The sampling plan is applied to the created overpayment population. The extrapolation is computed and compared to the total overpayment for the population.

Steps 1 and 2 are repeated independently a large number of times, enough to estimate achieved confidence level to high accuracy. The plan investigated can be a simple random sample or a stratified random sample. The function summarizes the testing with plots of the estimated achieved confidence level, and the average percent of overpayment recovery for the plan, versus error rate $P_{E}$ (the latter quantity is equivalent to average precision, as explained in Section 6). For most payment populations, the samptest results are returned in a few seconds, even using a computer of modest computing power.

For example, Figure 4.1 shows samptest results on achieved confidence levels for a simple random sample of 30 payments using the standard extrapolation (1.2) applied to the Figure 3.1 payment populations. The plot tells us that this modest sampling plan provides confidence level above or near the MPIM-required $90 \%$ level for population A for all error rates, but for populations B and C it fails to provide the $90 \%$ confidence level if the error rate $P_{E}$ exceeds 0.6 .

Figure 4.1 Achieved confidence levels (\%) versus error rate for the Figure 3.1 populations, using a simple random sample of 30 payments with the standard extrapolation (1.2).


Note: Estimates are accurate to $\pm 1 \%$ with $95 \%$ confidence.

These results illustrate the strong effect of skewness in the overpayment population on the achieved confidence level of the bound (1.2). Cochran (1977, pages 39-44) noted that a right-skewed population yielded a conservative lower bound and a liberal upper bound for a two-sided confidence interval for the population mean. Using this "principle of skew", which in our experience is pervasive, the conservatism of the lower bound (1.2) for Population A at all error rates can be anticipated: at any error rate $P_{E}$ the overpayment population has a "spike" of zeros of (approximate) size $\mathrm{N}\left(1-P_{E}\right)$, with the remaining $\mathrm{N} P_{E}$ overpayments having the same shape as the payment population. Hence at any error rate the overpayment population for population A is also right skewed; hence the bound (1.2) is conservative. At any error rate $P_{E}$, overpayment population B has two spikes, with $\mathrm{N}\left(1-P_{E}\right)$ zeros and $\mathrm{N} P_{E}$ payments near $\$ 4,000$. For values of $P_{E}$ above $1 / 2$ the spike at 0 is the shorter of the two, hence the overpayment population is leftskewed, hence the bound (1.2) is liberal. Finally, payment population C is left-skewed. For small values of $P_{E}$ the large spike at 0 in the overpayment population will counterbalance this left skew and yield a conservative lower bound, but the spike at 0 gradually disappears as $P_{E}$ grows, yielding a liberal lower bound at high error rates.

If problems with the achieved confidence level occur using extrapolation (1.2) they invariably occur at high error rates due to left skew in the overpayment population. For this reason, the presence of partial overpayments improves the achieved confidence level: at any particular $P_{E}>1 / 2$, partial overpayments reduce the severity of any left skew. This has been repeatedly confirmed in our experience using samptest, which has the capability to flexibly model the occurrence of partial overpayments. We conjecture that any sampling plan which achieves the $90 \%$ confidence level under the $\operatorname{AN}\left(P_{E}\right)$ model for a given error rate $P_{E}$ will also achieve it in the presence of partial overpayments at that value. That is, the presence of partial overpayments tends to improve the confidence level of the bound (1.2). This effect is usually modest since the frequency of partial overpayments is usually modest.

Right-skewed payment populations such as Population A occur frequently in Medicare investigations. For such a population, a simple random sample will usually achieve confidence level above $90 \%$ for all error rates. This conservatism signals the potential for improving the sampling plan via stratification. Care must be taken, however, for in a sample stratified by payment amount the effects of skew are mixed. For example, particularly for $L \geq 3$, stratification by payment amount creates payment subpopulations similar in shape to that shown in Figure 3.1B. These strata give rise to left-skewed overpayment strata at high error rates, which will not preserve the confidence level. King and Madansky (2013) say it well: "it is quite possible to destroy the achieved confidence level by stratifying carelessly. We therefore underscore the cautionary note that a misapplication of stratification (e.g., not using the appropriate stratification boundaries) will do more than invalidate the precision of the bounds; it can invalidate the normal confidence coefficient as well". Any stratification scheme must be tested using Monte Carlo.

Note that in some situations a transformation (e.g., the $\log$ transformation) may normalize the distribution well enough to provide a viable $90 \%$ confidence bound for the population mean of $\log (\mathrm{Y})$, but this cannot be back-transformed to obtain a confidence bound for the population mean of Y: the mean of
the logs is not the log of the mean. If the transformation symmetrizes the distribution, the back-transformed lower bound is a confidence bound for the population median of Y , not the population mean, and when multiplied by N does not provide a bound for the population total, but something much smaller. The same argument applies to other non-linear transformations, e.g., the square root transformation, inverse transformation, etc.

## 5. Methods for constructing strata using an auxiliary variable $X$

Assume that $X$ is known for all sampling units. In Medicare investigations, $X$ is usually the total paid amount for the sampling unit. This section reviews some existing methods for stratifying the population using $X$ and defines a new method.

In some designs, a few of the largest payment amounts are examined in their entirety in a "take all" or "certainty" stratum. This reduces the severity of right skew in the remaining population. If such a certainty stratum is used, the question remains as to how one should stratify the remaining population, and that is our focus here.

It should be stated at the outset that there is no need to use an algorithm from the peer-reviewed literature to determine values of $X$ ("cutpoints"), $X C_{1}<X C_{2}<\ldots<X C_{L-1}$ defining the $L \geq 2$ strata. In practice, any a priori choice achieving confidence level very near to or above $90 \%$ for all error rates is viable, regardless of how it was obtained. The time required to find such cutpoints by trial and error, testing each choice using Monte Carlo, may be prohibitive, however. Also, it is sometimes useful to cite the use of an existing algorithm from the peer-reviewed literature when creating a stratified design.

### 5.1 Existing methods

Several popular methods to determine cut points or stratum boundaries using X exist. With $Y$ the measured variable of interest, Tschuprow (1923) and Neyman (1934) proved that the variance of the linear population mean estimator $\bar{y}_{s}=\sum_{l=1}^{L} W_{l} \bar{y}_{l}$ is minimized for a fixed total sample size $n$ when the sample size $n_{l}$ in stratum $l$ is proportional to the product of stratum size $N_{l}$ and stratum standard deviation $\sigma_{Y l}, l=1,2, \ldots, L$. This "Neyman allocation" is the goal for several stratification algorithms under various assumptions. Methodological research in this area dates back to Dalenius and Gurney (1951) and Cochran (1977, Chapter 5A). Relevant literature is extensive, including Ekman (1959), Serfling (1968), Singh (1971), Wang and Aggarwal (1984), Hidiroglou and Srinath (1993), Hidiroglou (1994), Hedlin (2000), Kozak and Verma (2006), Jurina and Gligorova (2017), Hidiroglou and Kozak (2018) and Reddy and Khan (2019).

When the auxiliary variable $X$ is linearly related to $Y$, and highly correlated with $Y$, the well-known Dalenius and Hodges (1959) method can achieve approximate Neyman allocation. The method begins with a frequency distribution $f$ for $X$. Under the additional assumption that the distribution of $X$ within the
frequency distribution cells is approximately uniform, choosing stratum boundaries that equate the cumulative $\sqrt{f}$ between strata achieves approximate Neyman allocation for equal stratum sample sizes.

Lavallée and Hidiroglou (1988) take an iterative approach to choosing stratum cutpoints to minimize the total sample size $n$ given a specified relative coefficient of variation for the point estimator of the population total; this problem is equivalent to minimizing the relative coefficient of variation for fixed $n$, which is our goal. Their approach assumes that the auxiliary variable $X$ is "closely related to" $Y$, and they warn that their method will not achieve desired efficiency if $X$ and $Y$ are not "highly correlated". The iterative algorithm they use, due to Sethi (1963), was improved by Kozak (2004).

For sampling plans where the stratification variable $X$ is not linearly related to the survey variable $Y$, the "anticipated moments" under a model (for example, Dayal (1985), Sigman and Monsour (1995) and Sweet and Sigman (1995)) is a critical concept. Generalizing the algorithm in Lavallée and Hidiroglou (1988) and using "anticipated moments" of $Y$ given $X$, Rivest (1999, 2002) proposed stratification algorithms and models that account for discrepancies between $Y$ and $X$, in particular (1) a Log-Linear (LL) model, i.e., $\log (Y)$ is linearly related to $\log (X)$ and (2) a "random replacement" model. The latter models $Y=X$ with high probability, but otherwise $Y$ is equal to a randomly selected value of $X$. The algorithms in Rivest $(1999,2002)$ choose sample size to achieve a pre-specified level of precision, or to maximize precision for a fixed sample size, allowing for different sample allocation rules. The statistical models and anticipated moment approach in Rivest $(1999,2002)$ provide a very general and flexible method covering a wide range of survey scenarios.

Baillargeon and Rivest (2009) extend the Log-Linear (LL) model in Rivest (2002) to include a survival probability that models the probability of survey variable $Y$ taking on value 0 (equation (5.1)). The survival probability can be specified to vary by stratum. This model is particularly helpful for business surveys where a business is no longer in operation when the survey takes place but the $X$ variable has been collected resulting in a zero value for $Y$. Let $p_{l}$ denote survival probability for the $l^{\text {th }}$ stratum; assuming $\epsilon \sim N\left(0, \sigma^{2}\right)$, the LL model is

$$
Y= \begin{cases}\exp (\alpha+\beta \log X+\epsilon), & \text { with probability } p_{l}  \tag{5.1}\\ 0, & \text { with probability } 1-p_{l}\end{cases}
$$

For business surveys, the $p_{l}$ values typically increases with $X$, which means a greater chance of business survival or Y being non-zero when the value of $X$ is large. When a constant survival probability is specified across all strata, the probability of $Y$ being zero is the same for all sampling units. In this case, if $\beta=0, \sigma^{2}=0$, the model in equation (5.1) is equivalent to the All-or-Nothing model in Section 2 of this paper. This is a testimony to the generality of the models proposed by Rivest $(1999,2002)$ and Baillargeon and Rivest (2009).

A simpler method for determining cut points of stratification for right-skewed populations is due to Gunning and Horgan (2004) and Gunning, Horgan and Yancey (2004). The method is based on an observation of Cochran (1961) that the coefficients of variation across different strata are comparable in
near-optimum stratification. Assuming the coefficient of variation is constant across all strata, boundary points can be written as terms in a geometric series. Once the minimum and maximum values of the set of cut points are specified, the geometric relationship produces cutpoints for all strata. Gunning and Horgan state that their method assumes that $Y$ is "highly correlated with" $X$. It also assumes that distributions within strata are uniform for efficiency

### 5.2 Sample allocations

For a given sample size $n$ or a precision level, values of the boundary points are affected by the sample allocation scheme (Lavallée and Hidiroglou, 1988; Hidiroglou and Srinath, 1993; Horgan, 2006). Common allocation schemes include but are not limited to the power allocations, such as the Y-proportional and N proportional power allocations (Lavallée and Hidiroglou, 1988; Hidiroglou and Kozak, 2018). A general expression of allocation schemes is described in equation (5.2) below, which is included in the $\mathbf{R}$ package stratification. For stratum $l$, the sample size is $n_{l}=n a_{l}$. Combinations of parameters $q_{1}, q_{2}$ and $q_{3}$ produce different sample allocations, where $0 \leq 2 q_{1} \leq 1,0 \leq 2 q_{2} \leq 1$ and $0 \leq 2 q_{3} \leq 1 . N_{l}, \bar{Y}_{l}$ and $S_{l}$ are the size, mean and standard deviation of stratum $l$, respectively.

$$
\begin{equation*}
a_{l}=\frac{N_{l}^{2 q_{1}} \bar{Y}_{l}^{q_{2}} \bar{S}_{l}^{2 q_{3}}}{\sum_{l=1}^{L} N_{l}^{q_{l}} \bar{Y}_{l}^{q_{2}} S_{l}^{2 q_{3}}} . \tag{5.2}
\end{equation*}
$$

For instance, using the Dalenius-Hodges method where equal sample size achieves the goal of approximating the Neyman allocation, setting $q_{1}=q_{2}=q_{3}=0$ imples allocating an equal number of observations across the strata. For a power allocation where the power is 0.7 , one specifies $q_{1}=q_{2}=0.35$ and $q_{3}=0$. For a Y-proportional power allocation, the combination of parameters $q_{1}=q_{3}=0$, and $q_{2}=$ 0.35 yields a power of 0.7 .

To implement the aforementioned stratification methods under various sample allocation schemes, the $\mathbf{R}$ package stratification (see Baillargeon and Rivest, 2011) provides functions strata.cumrootf(), strata.LH() and strata.geo() to implement the Dalenius-Hodges, Lavallée-Hidiroglou, and Gunning-Horgan methods to produce cut points for a specified number of strata, respectively. The package also implements the alternatives studied by Rivest (2002) and Baillargeon and Rivest (2009).

### 5.3 A simple new method using anticipated AN-model moments and equal sample sizes

The well-known stratification method due to Dalenius and Hodges, under the model $\mathrm{Y} \approx \mathrm{X}$, chooses stratum cutpoints to achieve near-equality of the quantities $N_{l} \sigma_{Y}$. In that case, using equal sample sizes corresponds to Neyman allocation. The relatively simple new method described in this section mimics this approach using the AN model at a particular value $P_{E}$. For a given value $P_{E}$, the $\operatorname{ESS}\left(P_{E}\right)$ algorithm determines cutpoints by nearly equalizing the quantities $N_{l} \sigma_{Y l}$ determined by the AN -model moments (3.1).

Thus, for equal sample sizes and this value of $P_{E}$, the $\operatorname{ESS}\left(P_{E}\right)$ method provides Neyman allocation, maximizing precision for a given $n$.

For example, when $L=2$, a specified cutpoint $X C_{1}$ determines $N_{1}$ and $N_{2}$ as well as $\mu_{1 X}, \sigma_{1 X}^{2}, \mu_{2 X}$ and $\sigma_{2 X}^{2}$. Specifying $P_{E}$, we then use equation (3.1) to compute $\sigma_{1 Y}^{2}$ and $\sigma_{2 Y}^{2}$. The final cutpoint is found by iterating on $X C_{1}$ until $\left(N_{1} \sigma_{1 Y}-N_{2} \sigma_{2 Y}\right)^{2}$ is minimized. For general L , our algorithm (available from the second author) uses a nonlinear search through the R function optimize for univariate optimization when $L=2$, and constrOptim for higher-dimensional optimization when $L \geq 3$, to find cut points to minimize the corrected sum of squares for the quantities $N_{l} \sigma_{I Y}$. Exhaustive searches are available but computationally feasible only for $L \leq 3$.

The question remains: how does one choose $P_{E}$ ? Fortunately, as will be seen in the next section, the operating characteristics of $\operatorname{ESS}\left(P_{E}\right)$ seem rather robust to the choice. An idea worth exploring, suggested by the Associate Editor, is to initially choose $P_{E}$ to maximize (3.1). We recommend exploring the properties of a grid of choices using Monte Carlo prior to making the final choice.

## 6. Efficiency comparisons for the stratification methods

Equations (1.1) and (1.2) can be written informally as

$$
\begin{equation*}
(\text { overpayment recovery })=(\text { point estimate })-(\text { margin of error }) . \tag{6.1}
\end{equation*}
$$

The precision of a sample is defined to be its margin of error expressed as a percentage of its point estimate, with precision being considered "high" if this percentage is small. Some would not define precision to include the $t$ critical point; the gist of what is said below (that the average percent of overpayment recovered is inversely related to average precision) still holds under different definitions of precision. Traditionally, analysts seek to create a sample design having high average precision. Averaging (6.1) over all possible samples, and using the fact that the point estimator is unbiased, we obtain
(Average overpayment recovery) $=$ (true totol overpayment) - (average margin of error).
Dividing both sides of the above by true total overpayment and multiplying by $100 \%$, we obtain
(Average overpayment recovery percent) $=100 \%-$ (average precision),
Therefore, seeking a design with high average precision is equivalent to seeking a design with high average overpayment recovery. For example, a sample design with $10 \%$ average precision has $90 \%$ average overpayment recovery. For any sample design, both quantities will vary dramatically depending on the error rate $P_{E}$.

We prefer quantifying the efficiency of a design in terms of the average overpayment recovery as opposed to average precision. This facilitates determination of the cost-effectiveness of certain design decisions, such as increasing sample size or paying an analyst to spend an extra hour searching for efficient stratification schemes. In this section, we compare the efficiency of the major stratification algorithms discussed in Section 5.

Of the 166 samples discussed in Section 2, a total of 156 payment populations, matched by number, were available for study. These populations are numerically summarized and graphically depicted at https://drive. google.com/drive/folders/1-7M-4R3KPcCgfPmPWOo24Qc5AOAdOfV7?usp=sharing. Of these, 104 payment populations were selected as candidates for stratification. For these 104 test populations, simple random samples (SRS) of size 30 achieved estimated confidence level above $89 \%$ across all 7 tested error rates $P_{E}=0.10$, $0.30,0.50,0.70,0.9,0.95$, and 1.0 using 10,000 generated overpayment populations, each with its sample and extrapolation. It is assumed without loss of generality that, if a certainty (or "take-all") stratum is to be used that these few largest payments have already been removed, to be examined in their entirety. Stratification strategies apply only to the remaining payments

In our testing, designs with total sample sizes 30,60 , or 90 were considered, with $L=2,3$, or 4 strata, for a total of $104 * 9=936$ test cases. The stratification methods due to Delanius-Hodges (DH), LavalléeHidiroglou (LH), Gunning-Horgan (Geo) and the Log-Linear (LL) models were tested using the $\mathbf{R}$ package stratification to generate cutpoints. Three sample allocations were tested for the LH, Geo and LL methods: Y-proportional and N-proportional power allocations with power $p=0.7$ as well as a power allocation where $q_{1}=q_{2}=0.35, q_{3}=0$ are specified in equation (5.2). The stratum sample sizes were determined by the $\mathbf{R}$ functions that generated the cutpoints. For the other methods, stratum sample sizes were taken to be equal, except when $L=4$ and $n=30$ or 90 , where slight inequalities occurred.

To fully define an ESS stratification method, we must choose a value for $P_{E}$. Hopefully the choice will be viable and have good recovery properties regardless of the true value of $P_{E}$, and that is what the simulation testing is meant to answer. In this study, the ESS cutpoint algorithm was tested at $P_{E}=0.2$ (ESS20), $P_{E}=0.5$ (ESS50), $P_{E}=0.8$ (ESS80). The LL models are also tested with survival probabilities $0.2,0.5$ and 0.8 . The Monte Carlo testing for all cases and methods with 10,000 iterations at each error rate required approximately 27 hours on Bentley University's High Performance Computing cluster. The conservative Satterthwaite (1946) degrees of freedom were used for all methods.

For each of the 936 test cases, a stratification method was considered viable if it achieved estimated confidence above $89 \%$ for all 7 tested error rates. Table 6.1 shows, for each of the nine sample size - number of strata combinations, the number of test populations for which each method produced a viable stratification scheme. Only simulation results from the Y-proportional power allocation are reported as these allocation methods yielded better viability than the other two sample allocations. The table shows that stratification is more likely to be viable for small $L$ and large $n$. The Geo method is viable more often than any other stratification method; unfortunately, it will also be seen that it is usually less efficient than a simple random sample of the same size. Of the other stratification methods, the LL20 method, the ESS20 and the ESS50 methods were comparable in terms of viability and noticeably better in this respect than the other methods tested (see the last column of Table 6.1). Notably, these three methods were viable almost twice as often as the Dalenius-Hodges method, and almost three times as often as the Lavallée-Hidiroglou method. Unfortunately, in no case did any algorithm (except Geo) achieve viability for more than about half of the test populations.

In order to measure the efficiency of (for example) the Dalenius-Hodges method relative to simple random sampling, for each sample size, L-value, and population for which the DH method was viable, the
difference in average overpayment recovery in the order DH-SRS was computed for each error rate. Analogous differences were computed for the other stratification methods. Figure 6.1 shows boxplots of these differences. It is evident from this graphic that the Geo method improved on SRS for less than half of the cases for which it was viable. Additionally, when the DH, LH, LL and ESS methods were viable, they improved the overpayment recovery compared to SRS by a median amount of $4-6 \%$. In more than a few cases they improved the recovery dramatically, by $10-22 \%$. However, the recovery using DH can sometimes be worse than SRS by more than $9 \%$. In contrast, SRS was never better than any LL method by more than $1.5 \%$, and never better than any ESS method by more than $3 \%$. Both the LL and ESS methods can improve on SRS by up to $22 \%$.

Table 6.1
Method viability: number of test populations for which a stratification method (see text) achieved estimated confidence at least $\mathbf{8 9 \%}$ over all tested error rates, versus number of strata $L$ and total sample size $n$ under Yproportional power allocation with power value 0.7 .

| Methods | $\mathrm{L}=2$ |  |  | $\mathrm{L}=3$ |  |  | $\mathrm{L}=4$ |  |  | Total viability frequency |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathrm{n}=30$ | $\mathrm{n}=60$ | $\mathrm{n}=90$ | $\mathrm{n}=30$ | $\mathrm{n}=60$ | $\mathrm{n}=90$ | $\mathbf{n}=30$ | $\mathrm{n}=60$ | $\mathrm{n}=90$ |  |
| DH | 22 | 30 | 37 | 5 | 8 | 10 | 1 | 3 | 4 | 120 |
| LH | 19 | 22 | 25 | 4 | 5 | 5 | 0 | 0 | 0 | 80 |
| Geo | 90 | 91 | 83 | 52 | 62 | 61 | 5 | 15 | 23 | 482 |
| LL20 | 31 | 43 | 57 | 19 | 19 | 26 | 9 | 12 | 15 | 231 |
| LL50 | 28 | 42 | 47 | 14 | 18 | 20 | 5 | 11 | 10 | 195 |
| LL80 | 21 | 28 | 36 | 7 | 9 | 14 | 2 | 4 | 5 | 126 |
| ESS20 | 33 | 41 | 51 | 17 | 22 | 27 | 10 | 13 | 13 | 227 |
| ESS50 | 33 | 44 | 45 | 18 | 23 | 24 | 8 | 8 | 12 | 215 |
| ESS80 | 30 | 33 | 44 | 11 | 13 | 17 | 5 | 5 | 6 | 164 |
| SRS | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 936 |

Notes: DH = Dalenius-Hodges; ESS = Equal sample sizes; Geo = Gunning-Horgan; LH = Lavallée-Hidiroglou; LL = Log-linear; SRS $=$ Simple random samples.

Figure 6.1 Improvement in overpayment recovery versus a simple random sample of the same size for each viable method (see text for details).


Notes: $\mathrm{DH}=$ Dalenius-Hodges; ESS = Equal sample sizes; Geo = Gunning-Horgan; $\mathrm{LH}=$ Lavallée-Hidiroglou; LL $=$ Log-linear.

The top three algorithms were further compared via calculating differences in the order ESS50-ESS20, ESS50-LL20, and ESS20-LL20 whenever the differenced algorithms were both viable, for all error rates and test cases. These differences are displayed in Figure 6.2. The figure shows that the ESS50 and ESS20 methods had similar efficiency, suggesting that choice of $P_{E}$ in this range is not of critical importance for efficiency. Overall, the ESS methods and LL20 yielded similar expected overpayment recovery. On average, the ESS methods produced slightly higher medians than those of the LL20. However, the LL20 method had slightly higher mean expected overpayment recovery. In some cases, the ESS methods outperformed the LL20 method by over $4 \%$ while in other cases they were worse-off than the LL20 method by a little bit over $6 \%$.

Figure 6.2 Relative efficiency of the ESS50, ESS20 and LL20 methods (see text for details).


Notes: ESS = Equal sample sizes; LL = Log-linear.

## 7. Discussion and conclusion

This paper's major accomplishment is its unprecedented sharing of raw data from Medicare investigations. Despite this, these 166 shared samples cannot be considered representative of all such investigations. In particular these samples are taken from the first level of appeal; error rates tend to decrease at higher levels of appeal as providers challenge the overpayment determinations. The samples displayed here are also primarily from home health or hospice providers, a major shortcoming.

We see no reason why similar sharing of thoroughly redacted sample and payment population data could not be done by all MACs and at higher levels of appeal. If this was done, say, on a triannual basis, the anonymity of the individual providers and UPICs would be protected. More data sharing would allow for further refinement of models for the overpayment - payment relationship. It would also be very useful to know if negative overpayments are as rare and negligible at other levels of appeal, and for other provider lines of business, as they were for our samples.

Our data showed 24 "complete error" samples, i.e., samples where every overpayment equaled its payment. Use of the bounds (1.1) or (1.2) based on the finite population Central Limit Theorem can be problematic in these cases: at very high error rates, it is not unusual for these lower confidence bounds for total overpayment to exceed the total payment amount, which renders the bound indefensible. The bounds (1.1) and (1.2) can also fail to provide confidence level at least $90 \%$ at high error rates for payment populations similar to populations B and C in Figure 3.1. Alternative extrapolation methods based on the hypergeometric probability distribution (Edwards et al., 2003; Gilliland and Feng, 2010; Edwards et al., 2015) have been developed for situations where the bounds (1.1) and (1.2) tend to fail. These alternative methods always provide lower $90 \%$ confidence bounds less than the total payment amount. They are mathematically conservative - guaranteed to provide confidence level at least $90 \%$ - as long as negative overpayments do not occur with any great frequency or severity. More sharing of data is needed to shed light on the safety of this assumption.

With Medicare data, the frequency for which extrapolations using (1.1) and (1.2) are not viable is disappointing, to say the least. Alternative approaches using empirical likelihood confidence bounds (Chen, Chen and Rao, 2003; Rao and Wu , 2009) hold great promise. Unfortunately, simulation studies on empirical likelihood methods to date have not included error rates greater than $40 \%$, which are the norm for Medicare samples. A comprehensive study of empirical likelihood approaches using Medicare data is beyond the scope of this paper but is the focus for an ongoing project by the authors.

This paper proposes a new stratification method as a special case of the anticipated moment method based on the All-or-Nothing model. The efficiency study provided here found this new ESS method to be competitive with the best of the many existing methods tested from the $\mathbf{R}$ package stratification. Specifically, the Log-Linear model with survival probability of 0.2 (LL20) under the Y-proportional power allocation ( $p=0.7$ ) showed a slight advantage in viability over the best ESS methods. Finding this particular choice of operating parameters in the wide range available in stratification required the advice of an expert and hours of testing, however. In contrast, the near-equivalent ESS choices ( $P_{E}=0.2$ or 0.5 ) were among the first methods we tested. However, no one method dominated all other methods in all situations. And, for every choice of $n$ and $L$, there were many test populations for which no stratification algorithm was viable.

As is often the case, these results lead to new questions. Are there aspects of the payment population that can provide clues as to when stratification will lead to improved efficiency? Are there aspects of the payment population that can provide clues as to which method is the best choice? The authors will be pursuing these and other questions and invite collaboration from others.

## Acknowledgements

The authors thank the CMS for permission to use and display the data from these samples and populations, and thank the associate editor and reviewers for their hard work and helpful suggestions.

Piaomu Liu thanks Bentley University for providing High Performance Computing capabilities to perform efficiency studies in the research.

## References

Anscombe, F.J. (1973). Graphs in statistical analysis. The American Statistician, 27(1), 17-21.

Baillargeon, S., and Rivest, L.-P. (2009). A general algorithm for univariate stratification. International Statistical Review, 77(3), 331-344.

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

Chen, J., Chen, S.-Y. and Rao, J. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. Canadian Journal of Statistics, 31(1), 53-68.

Clemente, S., McGrady, R., Repass, R., Paul III, D.P. and Coustasse, A. (2018). Medicare and the affordable care act: Fraud control efforts and results. International Journal of Healthcare Management, 11(4), 356362.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. Bulletin of the International Statistical Institute, 35, 345-358.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley \& Sons, Inc.

Dalenius, T., and Gurney, M. (1951). The problem of optimum stratification. ii. Scandinavian Actuarial Journal, 1951(1-2), 133-148.

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

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristic. Journal of Statistical Planning and Inference, 11(3), 321-328.

Edwards, D., Gilliland, D., Ward-Besser, G. and Lasecki, J. (2015). Conservative penny sampling. Journal of Survey Statistics and Methodology, 3(4), 504-523.

Edwards, D., Ward-Besser, G., Lasecki, J., Parker, B., Wieduwilt, K., Wu, F. and Moorhead, P. (2003). The minimum sum method: A distribution-free sampling procedure for medicare fraud investigations. Health Services and Outcomes Research Methodology, 4(4), 241-263.

Ekman, G. (1959). An approximation useful in univariate stratification. The Annals of Mathematical Statistics, 30(1), 219-229.

Gilliland, D., and Feng, W. (2010). An adaptation of the minimum sum method. Health Services and Outcomes Research Methodology, 10(3), 154-164.

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

Gunning, P., Horgan, J.M. and Yancey, W. (2004). Geometric stratification of accounting data. Contaduría y Administración, (214), 0.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 1491-1523.

Hedlin, D. (2000). A procedure for stratification by an extended ekman rule. Journal of Official Statistics, 16(1), 15.

Hidiroglou, M.A. (1994). Sampling and estimation for establishment surveys: Stumbling blocks and progress. Proceedings of the Survey Research Methods Section, American Statistical Association, 693698.

Hidiroglou, M.A., and Kozak, M. (2018). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. International Statistical Review, 86(1), 87-105.

Hidiroglou, M.A., and Srinath, K. (1993). Problems associated with designing subannual business surveys. Journal of Business \& Economic Statistics, 11(4), 397-405.

Horgan, J.M. (2006). Stratification of skewed populations: A review. International Statistical Review, 74(1), 67-76.

Huffman, M. (2021). Value-based care: An executive briefing. Nurse Leader, 19(1), 82-86.

Jurina, I., and Gligorova, L. (2017). Determination of the optimal stratum boundaries in the monthly retail trade survey in the croatian bureau of statistics. Romanian Statistical Review, (4).

King, B. (1996). Sampling design issues when dealing with zeros. Proceedings of the Survey Research Methods Section, 400-405.

King, B., and Madansky, A. (2013). On sampling design issues when dealing with zeros. Journal of Survey Statistics and Methodology, 1(2), 144-170.

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

Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. Survey Methodology, 32, 2, 157-163. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9550-eng.pdf.

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

Li, X., and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. Journal of the American Statistical Association, 112(520), 1759-1769.

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

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http: //www.R-project.org/.

Rao, J., and Wu, C. (2009). Empirical likelihood methods. Elsevier, Handbook of Statistics, 29, 189-207.

Reddy, K.G., and Khan, M.G. (2019). Optimal stratification in stratified designs using weibull-distributed auxiliary information. Communications in Statistics-Theory and Methods, 48(12), 3136-3152.

Rivest, L. (1999). Stratum jumpers: Can we avoid them. Proceedings of the Survey Research Methods Section, American Statistical Association, 64-72.

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

Salmond, S.W., and Echevarria, M. (2017). Healthcare transformation and changing roles for nursing. Orthopedic Nursing, 36(1), 12.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2(6), 110-114.

Serfling, R. (1968). Approximately optimal stratification. Journal of the American Statistical Association, 63(324), 1298-1309.

Sethi, V. (1963). A note on optimum stratification of populations for estimating the population means. Australian Journal of Statistics, 5(1), 20-33.

Sigman, R.S., and Monsour, N.J. (1995). Selecting samples from list frames of businesses. Business Survey Methods, 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. Journal of the American Statistical Association, 66(336), 829-833.

Sweet, E.M., and Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. Proceedings of the Section on Survey Research Methods, 1, 491-496.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. Metron, 2(461-493), 646-683.

Wang, M., and Aggarwal, V. (1984). Stratification under a particular pareto distribution. Communications in Statistics-Theory and Methods, 13(6), 711-735.

## ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2023.

- B. Baffour, Australian National University
- C. Bayart, University Claude Bernard Lyon 1
- Béliveau, University of Waterloo
- Bell, U.S. Census Bureau
- E. Berg, Iowa State University
- K. Blocksom, US EPA
- Bocci, Statistics Canada
- H.-J. Boonstra, Statistics Netherlands Heerlen
- Boulet, Statistics Canada
- J. Breidt, NORC at the University of Chicago
- J.M. Brick, Westat Inc.
- J. Brown, University of Technology Sydney
- P.J. Cantwell, U.S. Census Bureau
- G. Chauvet, École nationale de la statistique et de l'analyse de l'information
- S. Chen, University of Oklahoma Health Sciences Center
- Y. Cheng, USDA National Agricultural Statistics Service
- J. Chipperfield, Australian Bureau of Statistics
- M. Cohen, American Institutes for Research
- M. Dagdoug, Laboratoire de Mathématiques de Besançon
- G. Datta, University of Georgia
- N. Dean, Emory University
- M. DeBell, Stanford University
- M. del Mar Rueda, University of Granada
- P. Duchesne, Université de Montréal
- J.L. Eltinge, US Census Bureau
- N. English, National Opinion Research Center
- Erciulescu, Westat Inc.
- Franco, NORC at the University of Chicago
- W.A. Fuller, Iowa State University
- M. Ghosh, University of Florida
- Girard, Statistics Canada
- K.C.M. Gonçalves, Universidade Federal do Rio de Janeiro
- Han, Westat Inc.
- Haziza, University of Ottawa
- M.A. Hidiroglou, Statistics Canada
- P. Hoff, Duke University
- Judkins, ABT Associates Inc Bethesda
- J.K. Kim, Iowa State University
- H. Kitada Smalley, Willamette University
- S. Kolenikov, NORC at the University of Chicago
- P.S. Kott, RTI International
- P. Lahiri, University of Maryland
- D. Lee, The University of Alabama
- S. Lee, University of Michigan
- É. Lesage, L'Institut national de la statistique et des études économiques
- T. Lewis, George Mason University
- P. Lugtig, Utrecht University
- J.M. Marin, University Carlos III of Madrid
- Matei, Université de Neuchâtel
- K. McConville, Reed College
- T. McElroy, US Census Bureau
- T. Merly-Alpa, INED
- Molina, Universidad Complutense de Madrid
- J. Moore, University of Essex
- R. Münnich, University of Trier
- Nandram, Worcester Polytechnic Institute
- Nettel Aguirre, University of Wollongong
- J. Opsomer, Westat Inc.
- Pfeffermann, University of Southampton
- J.N.K. Rao, Carleton University
- R. Rapley, Defence Science and Technology Laboratory
- L.-P. Rivest, Université Laval
- Ruiz-Gazen, Toulouse 1 University Capitole
- T. Saegusa, University of Maryland at College Park
- F.J. Scheuren, National Opinion Research Center
- Schouten, Statistics Netherlands
- N. Shlomo, The University of Manchester
- P.L.d.N. Silva, Escola Nacional de Ciências Estatísticas
- P. Smith, University of Southampton
- Steel, University of Wollongong
- Struminskaya, Utrecht University
- Y. Tillé, University of Neuchatel Faculty of Sciences
- R. Tiller, US Bureau of Labor Statistics
- M. Torabi, University of Manitoba
- Toth, U.S. Bureau of Labor Statistics
- J. van den Brakel, Statistics Netherlands
- J. Wagner, University of Michigan
- J. Wakefield, University of Washington
- L. Wang, National Cancer Institute
- Z. Wang, Xiamen University
- B. West, University of Michigan
- Wu, University of Waterloo
- W. Yung, Statistics Canada
- L.-C. Zhang, University of Southampton
- Y. Zhao, Georgia State University

Acknowledgements are also due to those who assisted during the production of the 2023 issues: Céline Ethier of Economic Statistics Methods Division; Patrick O'Leary of Social Statistics Methods Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Kathy Charbonneau, Sophie Pilon, Travis Robinson, Niruja Sivalingam and Karo-Lynn Audy as well as our partners in the Communications Division.

## ANNOUNCEMENTS

## Nominations Sought for the 2025 Waksberg Award

The journal Survey Methodology has established an annual invited paper series in honor of the late Joseph Waksberg to recognize his outstanding contributions to survey statistics and methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey statistics and methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2025 Waksberg Invited Address at the Statistics Canada Symposium, expected to be held in the autumn of 2025. The paper will be published in an upcoming issue of Survey Methodology (targeted for December 2025).

The author of the 2025 Waksberg paper will be selected by a four-person committee appointed by Survey Methodology and the American Statistical Association. Nomination of individuals to be considered should be sent by email before February 15, 2024 to the chair of the committee, Denise Silva (denisebritz@gmail.com). Nominations should include a CV and a nomination letter. Nominations will remain active for 5 years.

Members of the Waksberg Paper Selection Committee (2023-2024)

Denise Silva, Brazilian Institute of Geography and Statistics (Chair)<br>Jae-Kwang Kim, Iowa State University<br>Paul Smith, University of Southampton<br>Kristen Olson, University of Nebraska-Lincoln

## Past Chairs:

Graham Kalton (1999-2001)
Chris Skinner (2001-2002)
David A. Binder (2002-2003)
J. Michael Brick (2003-2004)

David R. Bellhouse (2004-2005)
Gordon Brackstone (2005-2006)
Sharon Lohr (2006-2007)
Robert Groves (2007-2008)
Leyla Mojadjer (2008-2009)
Daniel Kasprzyk (2009-2010)
Elizabeth A. Martin (2010-2011)
Mary E. Thompson (2011-2012)
Steve Heeringa (2012-2013)
Cynthia Clark (2013-2014)
Louis-Paul Rivest (2014-2015)
Tommy Wright (2015-2016)
Kirk Wolter (2016-2017)
Danny Pfeffermann (2017-2018)
Michael A. Hidiroglou (2018-2019)
Robert E. Fay (2019-2020)
Jean Opsomer (2020-2021)
Jack Gambino (2021-2022)
Maria Giovanna Ranalli (2022-2023)

## 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. 2, June 2023
Effects of Changing Modes on Item Nonresponse in Panel Surveys Oliver Lipps, Marieke Voorpostel and Gian-Andrea Monsch ..... 139
Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach
Daniela Marella ..... 151
Design and Sample Size Determination for Experiments on Nonresponse Followup using a Sequential Regression Model
Andrew M. Raim, Thomas Mathew, Kimberly F. Sellers, Renee Ellis and Mikelyn Meyers ..... 173
Estimating Intra-Regional Inequality with an Application to German Spatial Planning Regions Marina Runge ..... 203
Constructing Building Price Index Using Administrative Data
Masahiro Higo, Yumi Saita, Chihiro Shimizu and Yuta Tachi ..... 229
From Quarterly to Monthly Turnover Figures Using Nowcasting Methods
Daan Zult, Sabine Krieg, Bernd Schouten, Pim Ouwehand and Jan van den Brakel ..... 253

## 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. 3, September 2023
Letter to Editor Quality of 2017 Population Census of Pakistan by Age and Sex Asif Wazir and Anne Goujon ..... 275
Looking for a New Approach to Measuring the Spatial Concentration of the Human Population Federico Benassi, Massimo Mucciardi and Giovanni Pirrotta ..... 285
Predicting Days to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach Stephanie Coffey and Michael R. Elliott. ..... 325
Towards Demand-Driven On-The-Fly Statistics Tjalling Gelsema and Guido van den Heuvel. ..... 351
Database Reconstruction Is Not So Easy and Is Different from Reidentification
Krishnamurty Muralidhar and Josep Domingo-Ferrer. ..... 381
Comment to Muralidhar and Domingo-Ferrer (2023) - Legacy Statistical Disclosure Limitation Techniques Were Not An Option for the 2020 US Census of Population And Housing Simson Garfinkel ..... 399
A Rejoinder to Garfinkel (2023) - Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option
Krishnamurty Muralidhar and Josep Domingo-Ferrer ..... 411
A Note on the Optimum Allocation of Resources to Follow up Unit Nonrespondents in Probability Surveys Siu-Ming Tam, Anders Holmberg and Summer Wang ..... 421

All inquires about submissions and subscriptions should be directed to jos@scb.se

## Volume 51, No. 1, March/mars 2023

Issue Information ..... 1
Research Articles
A new test for high-dimensional regression coefficients in partially linear models
Fanrong Zhao, Nan Lin and Baoxue Zhang .....  5
Economic variable selection
Koji Miyawaki and Steven N. MacEachern ..... 19
Doubly robust weighted composite quantile regression based on SCAD- $L_{2}$ Zhimiao Cao, Xiaoning Kang and Mingqiu Wang. ..... 38
Nonparametric confidence regions via the analytic wild bootstrap
Katherine L. Burak and Adam B. Kashlak ..... 77
Graphon estimation via nearest-neighbour algorithm and two-dimensional fused-lasso denoising
Oscar Hernan Madrid Padilla and Yanzhen Chen ..... 95
Multivariate online regression analysis with heterogeneous streaming data
Lan Luo and Peter X.-K. Song. ..... 111
Bayesian clustering for continuous-time hidden Markov models
Yu Luo, David A. Stephens and David L. Buckeridge. ..... 134
Zero-inflated Poisson model with clustered regression coefficients: Application to heterogeneity learning of field goal attempts of professional basketball players
Guanyu Hu, Hou-Cheng Yang, Yishu Xue and Dipak K. Dey ..... 157
A model-averaging treatment of multiple instruments in Poisson models with errors Xiaomeng Zhang, Xinyu Zhang and Yanyuan Ma. ..... 173
Penalized likelihood ratio test for a biomarker threshold effect in clinical trials based on generalized linear models Parisa Gavanji, Wenyu Jiang and Bingshu E. Chen. ..... 199
Regression analysis of multivariate current status data under a varying coefficients additive hazards frailty model Yanqin Feng, K.D. Prasangika and Guoxin Zuo ..... 216
Estimation of conditional cumulative incidence functions under generalized semiparametric regression models with missing covariates, with application to analysis of biomarker correlates in vaccine trials
Yanqing Sun, Fei Heng, Unkyung Lee and Peter B. Gilbert. ..... 235
Minimax A-, c-, and I-optimal regression designs for models with heteroscedastic errors
Hanan Abousaleh and Julie Zhou. ..... 258
Better experimental design by hybridizing binary matching with imbalance optimization
Abba M. Krieger, David A. Azriel and Adam Kapelnerr. ..... 275
Two-dimensional projection uniformity for space-filling designs
Sixu Liu, Yaping Wang and Fasheng Sun. ..... 293
Statistical data integration using multilevel models to predict employee compensation
Andreea L. Erciulescu, Jean D. Opsomer and Benjamin J. Schneider ..... 312
On asymptotic approximation of ratio models for weakly dependent sequences
Yi Wu, Wei Yu, Wenzhi Yang, Saisai Ding and Xuejun Wang ..... 327
Acknowledgement
Acknowledgement of Referees' Services/Remerciements aux membres des jurys ..... 344

## Volume 51, No. 2, June/juin 2023

Issue Information ..... 351
Research Article
Integrating information from existing risk prediction models with no model details
Peisong Han, Jeremy M.G. Taylor and Bhramar Mukherjee ..... 355
Distributed estimation with empirical likelihood
Qianqian Liu and Zhouping Li ..... 375
Divide and conquer for accelerated failure time model with massive time-to-event data Wen Su, Guosheng Yin, Jing Zhang and Xingqiu Zhao ..... 400
Optimal subsampling for large-sample quantile regression with massive data Li Shao, Shanshan Song and Yong Zhou ..... 420
Causal inference for multiple treatments using fractional factorial designs Nicole E. Pashley and Marie-Abèle C. Bind ..... 444
Dynamic treatment regimes with interference
Cong Jiang, Michael P. Wallace and Mary E. Thompson ..... 469
Reducing bias due to misclassified exposures using instrumental variables Christopher Manuel, Samiran Sinha and Suojin Wang ..... 503
Nonparametric tests for treatment effect heterogeneity in observational studies
Maozhu Dai, Weining Shen and Hal S. Stern ..... 531
Subgroup analysis for functional partial linear regression model
Haiqiang Ma, Chao Liu, Sheng Xu and Jin Yang ..... 559
Classified generalized linear mixed model prediction incorporating pseudo-prior information Haiqiang Ma and Jiming Jiang ..... 580
Efficient multiple change point detection for high-dimensional generalized linear models
Xianru Wang, Bin Liu, Xinsheng Zhang, Yufeng Liu, for the Alzheimer's Disease Neuroimaging Initiative ..... 596
Missing data analysis with sufficient dimension reduction
Siming Zheng, Alan T.K. Wan and Yong Zhou ..... 630
Weighted lens depth: Some applications to supervised classification
Alejandro Cholaquidis, Ricardo Fraiman, Fabrice Gamboa and Leonardo Moreno ..... 652
The EAS approach for graphical selection consistency in vector autoregression models Jonathan P. Williams, Yuying Xie and Jan Hannig ..... 674
Composite bias-reduced $L^{p}$-quantile-based estimators of extreme quantiles and expectiles Gilles Stupfler and Antoine Usseglio-Carleve ..... 704

## 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 $1 \frac{1}{2}$ 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 " $\log (\cdot)$ ", etc.
3.4 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in Section 4.
3.5 Bold fonts should normally be used to distinguish vectors and matrices from scalars.

## 4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the top of tables or figures. Use a two-level numbering system based on the section of the paper. For example, Table 3.1 is the first table in Section 3.
4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

## 5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
5.2 The first time a reference is cited in the text, the name of all authors must be written. For subsequent occurrences, the names of all authors can again be written. However, if the reference contains three or more authors, the names of the second and subsequent authors can be replaced with "et al.".
5.3 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## 6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words, including tables, figures and references.


[^0]:    Number 2

[^1]:    Notes: Independent SRSWOR samples were taken in each of $D=40$ areas with sample sizes ranging between 5 and 40 .
    ELE $=$ Exchangeable linkage errors; MIP $=$ Missing information principle; RMSE $=$ Root mean square error; SRSWOR $=$ Simple random sampling without replacement.

[^2]:    1. Camelia Goga, LMB, Université de Franche-Comté. E-mail: camelia.goga@univ-fcomte.fr; Anne Ruiz-Gazen, TSE, Université Toulouse

    Capitole. E-mail: anne.ruiz-gazen@tse-fr.eu.

[^3]:    1. Alina Matei, Institute of Statistics, University of Neuchâtel, Switzerland. E-mail: alina.matei@unine.ch; Paul A. Smith, S3RI \& Dept. of Social Statistics and Demography, University of Southampton, UK. E-mail: p.a.smith@soton.ac.uk; Marc J.E. Smeets, Statistics Netherlands, Netherlands. E-mail: mje.smeets@cbs.nl; Jonas Klingwort, Statistics Netherlands, Netherlands. E-mail: j.klingwort@cbs.nl.
[^4]:    1. Estelle Medous, Toulouse School of Economics, Université Toulouse Capitole, 1 Esplanade de l'Université, 31000 Toulouse, Université de Franche-Comté, Laboratoire de Mathématiques de Besançon and La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex. E-mail: estelle.medous@tse-fr.eu; Camelia Goga, Université de Franche-Comté, Laboratoire de Mathématiques de Besançon. E-mail: camelia.goga@univ-fcomte.fr; Anne Ruiz-Gazen, Toulouse School of Economics, Université Toulouse Capitole, 1 Esplanade de l’Université, 31000 Toulouse. E-mail: anne.ruiz-gazen@tse-fr.eu; Jean-François Beaumont, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada. E-mail: jean-francois.beaumont@statcan.gc.ca; Alain Dessertaine and Pauline Puech, La Poste, 3 rue Jean Richepin, 93192 Noisy le Grand cedex. E-mail: alain.dessertaine@laposte.fr and pauline.puech@laposte.fr.
[^5]:    1. Vincent Loonis, Insee (Institut national de la statisitque et des études économiques), Paris, France. E-mail: vincent.loonis@insee.fr.
[^6]:    Notes: PSU = primary sampling unit; NRFU = nonresponse followup; Ftf = face-to-face.

[^7]:    Notes: For the CIL, values for each variable divided by the mean of the CIL across scenarios. $\mathrm{RB}=$ relative bias; $\mathrm{PS}=$ primary sampling; $\mathrm{CV}=$ coefficient of variation; $\mathrm{RRMSE}=$ relative root mean squared error; $\mathrm{CI}=$ confidence interval; $\mathrm{ABS}(\mathrm{RB})=$ absolute standardized bias (relative bias).

[^8]:    1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. E-mail: balnan@wpi.edu and balgobin.nandram@usda.gov; Nathan B. Cruze, NASA Langley Research Center, Mail Stop 290, Hampton, VA 23681. E-mail: nathan.b.cruze@nasa.gov; Andreea L. Erciulescu, Westat, 1600 Research Blvd., Rockville, MD 20850, USA. E-mail: AndreeaErciulescu@westat.com.
[^9]:    Notes: MFSA is the new benchmarking model with FSA values as lower bounds for the model estimates. MFSA-DS refers to the double shrinkage model with benchmarking and inequality constraint, $\mathrm{CV}(0.08-0.93)$ and $d=0.95$.
    ASD = Agricultural Statistics Districts; CV = Coefficient of variation; DE = Direct estimates; DS = Double shrinkage; FSA = Farm Service Agency; ME = Bayesian Fay-Herriot model; MERB = Bayesian Fay-Herriot model with random benchmarking; MFSA = Bayesian FayHerriot model with inequality constraint and benchmarking; NDS $=$ Not including double shrinkage.

[^10]:    1. Yanzhe Li, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands; Sander Scholtus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: s.scholtus@cbs.nl; Arnout van Delden, Statistics Netherlands, PO Box 24500 , 2490 HA The Hague, The Netherlands.
