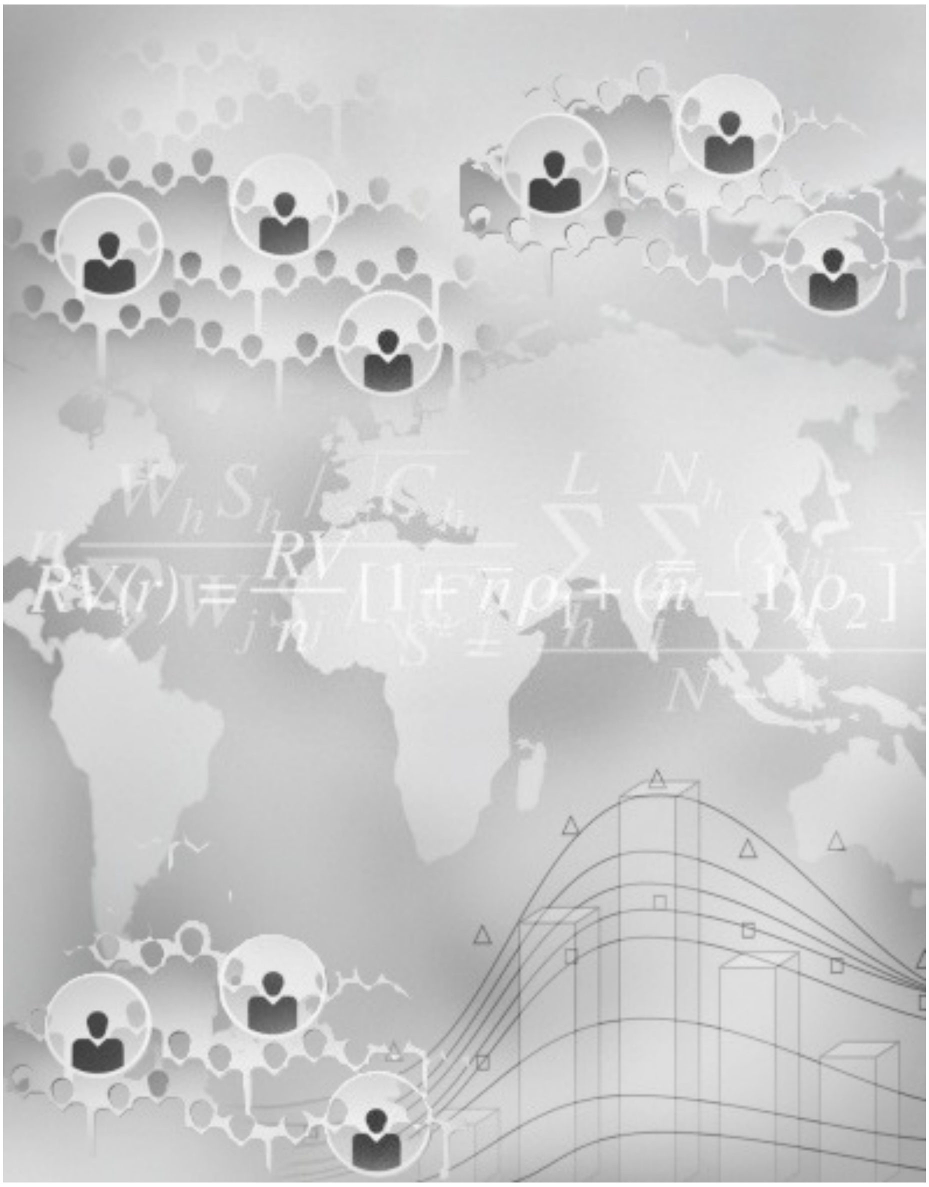
the Survey Statistician

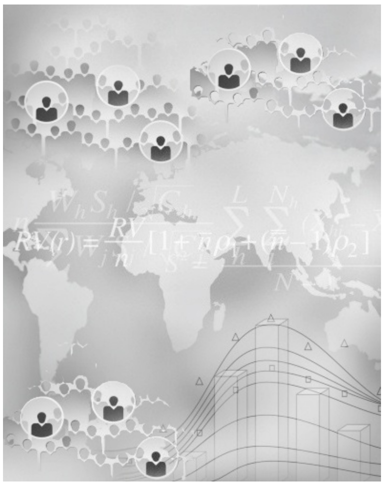
The Newsletter of the International Association of Survey Statisticians

No. 93 January 2026



International Association of Survey Statisticians

the Survey Statistician
The Newsletter of the International Association of Survey Statisticians



The Survey Statistician No. 93, January 2026

The Survey Statistician is published twice a year by the International Association of Survey Statisticians. Its members are informed about the new issue. The Survey Statistician is available on the IASS website at http://isi-iass.org/home/services/the-survey-statistician/

Enquiries for membership in the Association or change of address for current members should be addressed to: **isimembership@cbs.nl**

Comments on the contents or suggestions for articles in *The Survey Statistician* should be sent via e-mail to the editors Alina Matei (alina.matei@unine.ch) or Andrea Diniz da Silva (andrea.diniz@ibge.gov.br)

ISSN 2521-991X

In this Issue

3 Letter from the Editors

4 Letter from the President

5 Report from the Scientific Secretary

7 News and Announcement

9 Debate
- Are Non-probability Samples the Future of Surveys? By Changbao Wu and Li-Chun Zhang.

13 Ask the Experts
- Differential Privacy and its Application to Survey Data by Anne-Sophie Charest and Jörg Drechsler.

26 New and Emerging Methods
- Measures of Non-Ignorable Selection Bias for Non-Probability Samples by Sergio D. Martinez, Brady T. West, and Rebecca R. Andridge.

42 Early Career Survey Statistician
- Modeling Complex Survey Data: a Case Study of International Health Surveillance Surveys by Timothy Raxworthy, Yajuan Si, and Grace Chung.
- Variance Component Estimation Under a General Area-level Model by Yuting Chen and Hanqing Li.

62 Book & Software Review
- “Robust Small Area Estimation: Methods, Theory, Applications, and Open Problems” by Sylvia Harmening.
- R2BEAT: An R Package for Performing Optimal Allocation and Sample Selection by Giulio Barcaroli, Ilaria Bombelli, Andrea Fasulo, Alessio Guandalini, and Marco D. Terribili.

75 *Country Reports*
- Argentina, Australia, Brazil, Canada, Croatia, France, Hong Kong SAR, Lithuania, The Netherlands, Poland, United States, Uruguay.

86 Upcoming Conferences and Workshops

87 In Other Journals

93 Welcome New Members!

94 IASS Executive Committee Members

95 IASS Institutional Members

Letter from the Editors

Dear colleagues and readers,

We wish you a happy new year!

In this issue of *The Survey Statistician* (TSS), we bring you the latest news from the International Association of Survey Statisticians (IASS). This includes updates from the newly elected president, Partha Lahiri, and the appointed scientific secretary, Jenny Thompson. We also reveal the winner of the 2026 Waksberg Award.

TSS is privileged to publish papers by accomplished researchers and practitioners from around the world, as well as by young researchers. We are launching a new section called `Debate'. In it, Changbao Wu and Li-Chun Zhang answer `yes' and `no', respectively, to the question, `Are non-probability samples the future of surveys?'. In the `Ask the Experts' section, Anne-Sophie Charest and Jörg Drechsler review differential privacy and its application to survey data. In the `New and Emerging Methods' section, Sergio D. Martinez, Brady T. West and Rebecca R. Andridge discuss measures of non-ignorable selection bias for non-probability samples. Section `Early Career Survey Statistician' features two articles: `Modeling complex survey data: a case study of international health surveillance surveys' by Timothy Raxworthy, Yajuan Si and Grace Chung, and `Variance component estimation under a general area-level model' by Yuting Chen and Hanqing Li.Sylvia Harmening reviews the book `Robust Small Area Estimation: Methods, Theory, Applications, and Open Problems' by J. Jiang and J. S. Rao. The `Software Review' section is dedicated to an R package for optimal allocation and sample selection; the authors of this paper are Giulio Barcaroli, Ilaria Bombelli, Andrea Fasulo, Alessio Guandalini and Marco D. Terribili.

Country reports, a list of upcoming conferences and workshops, a list of articles published in other journals and other news about the IASS are also presented.

This issue of TSS marks a change in the editorial board, and we would like to thank the outgoing editors: Danutė Krapavickaitė, Annamaria Bianchi and Veronica Ballerini. We would particularly like to thank Danutė Krapavickaitė for her invaluable work as Editor-in-Chief of TSS in recent years. We would also like to welcome the new editors and express our gratitude to the current TSS editors for their rigour and dedication: Gaia Bertarelli, Mehdi Dagdoug, Francesco Pantalone, Jenny Thompson, Ton de Waal and Peter Wright. Finally, we would like to thank all the authors and contributors to this issue.

To help keeping TSS interesting, please share your knowledge and experience by presenting interesting topics and providing overviews of different areas of survey statistics and new ideas.

We hope you enjoy reading the January 2026 issue!

Alina Matei and **Andrea Diniz da Silva**

TSS Editors

Letter from the President

Dear Members,

Following a productive and inspiring 65th World Statistics Congress in The Hague, I am deeply honored to begin my term as President of the International Association of Survey Statisticians (IASS) for 2025–2027.

Stepping into this role, I am reminded of the immense dedication required to keep our global community vibrant. I want to personally thank our outgoing President, Natalie Shlomo. Her leadership and mentorship have been invaluable to me as President-Elect (2023-2025), and she leaves the Association with a clear and ambitious path forward. My thanks also go to our departing Vice-presidents Eric Rancourt, Jiraphan Suntornchost, Andres Gutierrez, and Annamaria Bianchi, as well as the ISI Executive Council and Permanent Office. Their tireless work behind the scenes is the reason our initiatives continue to thrive. Please refer to the Scientific Secretary's report for a list of our predecessors' recent accomplishments.

I am delighted to work alongside our newly approved Executive Committee: Ralf Münnich (President-Elect), and our Vice Presidents Gaia Bertarelli (Italy), Robert Clark (Australia), Haoyi Chen (China), and Katherine Jenny Thompson (USA). Together, we represent a diverse range of perspectives that will guide our strategy through 2027.

We are at a crossroads where traditional survey methods are meeting new data frontiers, and the IASS must be the home for that transition. To that end, we will continue to champion excellence through our flagship awards: the Hukum Chandra Prize (2026) and the Cochran-Hansen Prize (2027). These awards are more than just accolades; they are our way of ensuring that the next generation of survey statisticians, particularly from developing regions, has a seat at the global table.

Our Monthly Webinar Series will remain a cornerstone of our value to you. We are designing these sessions to be more than just lectures. We want them to be a space for:

- Discovery: Exploring frontier research in survey design and data integration.
- Problem Solving: Discussing the `messy' real-world challenges faced by National Statistical Offices.
- Practice: Providing hands-on training for the tools and software used daily.

Maintaining our financial health is a priority, and because membership fees currently do not support financial conference sponsorships, we are exploring a new seed partnership model. By funding specific deliverables, like short courses, in exchange for a share of the proceeds, we can continue to support 2026 conferences in a way that is sustainable for the Association. We also remain committed to providing no-cost co-sponsorships for mission-aligned events worldwide.

Please feel free to reach out to me directly at **plahiri@umd.edu**. I look forward to working with all of you to make the next two years a period of growth and innovation for the IASS.

With best wishes,

Partha Lahiri

IASS President 2025–2027

Report from the Scientific Secretary

Happy new year!

As the new scientific secretary of the IASS, I'll begin by thanking my predecessor Annamaria Bianchi (Italy) for her excellent work. This report primarily lists the efforts and achievements of the previous Executive Committee (EC), along with a few updates from the new EC, which has been meeting monthly since October 2025.

The new EC's first order of business was to discuss our communication. While we appreciate the timeliness of a monthly newsletter, it can be time-consuming to prepare and risks redundancy with the Scientific Secretary report in the January and July *The Survey Statistician* (TSS). Henceforth, these two reports serve as quarterly newsletters, with separate newsletters issued in April and in October. For timely updates and information, visit us at LinkedIn, Facebook, and Twitter (@iass_isi). To advertise events, seminars, or job opportunities through IASS social media, email **gaia.bertarelli@unive.it** with `IASS Social Media Post' in the subject line.

The IASS made several contributions to the 65th World Statistics Congress (WSC), held at The Hague from 5-9 October 2025. For details – and pictures – see the IASS October 2025 newsletter. Besides contributing to the WSC, the IASS provided financial support for three workshops: the workshop in honour of Professor Yves Tillé on the occasion of his retirement, Neuchâtel, Switzerland from 25-26 June 2025, the Baltic-Nordic-Ukrainian Network (BNU) Workshop in Vilnius, Lithuania from 25-29 August 2025 and the European Network for Better Establishment Statistics, 9th, biennial European Establishment Statistics Workshop (EESW25) in Rome, Italy from 5-7 November 2025. The IASS sponsored four invited talks the Fourth Workshop on Methodologies for Official Statistics in Rome, Italy from 1-2 December: `On adversarial risk analysis in official statistics' by F. Ruggeri (Italy), the keynote address; `Statistical Inference for a finite population mean with machine learning-based imputation for missing survey data' by D. Haziza (Canada) and M. Dagdoug (Canada); `A contamination model for multivariate zero-inflated data' by D. Di Cecco (Italy), D. Filipponi (Italy) and I. Guarnera (Italy); and `Small area estimation via spatio-temporal M-quantile modeling' by N. Salvati (Italy), F. Schirripa Spagnolo (Italy), M. Bugallo (Spain) and D. Morales (Spain).

As said in the President's report, the IASS is unable to offer financial funding for workshops or conferences in 2026 but continues to offer no-cost co-sponsorships to conferences of interest to IASS members. These conferences include the Survey Cost Conference to be held in Washington, D.C., on February 9-10, 2026, the 5th ISI Regional Statistics Conference (RSC) in Valletta, Malta, from 3-5 June 2026, and the Small Area Estimation Conference 2026 (SAE 2026) will be held in Bucharest, Romania, from 15-19 June 2026.

We have conducted five successful webinars since July 2025: `Debiased calibration estimation using generalized entropy under selection bias' by Jae Kwang Kim (U.S.A.), `Targeted designs to address survey nonresponse' by Peter Lynn (U.K.), `Sampling for business surveys at Statistics Canada' (2025 Waksberg Award Lecture) by M. A. Hidiroglou (Canada), `R-indicators for assessing representativeness for survey and non-survey data' by Natalie Shlomo (U.K.), and `Some history of the use of models in survey sampling' by Richard Valliant (U.S.A.). See our events page for updates in 2026. Expect to see a variety of topics ranging from theoretical to operational and presenters from developed and from emerging nations.

There are three open calls that should be of interest to the IASS membership:

1. Guest editors Maria Rosaria Ferrante and Natalie Shlomo seek contributions to a Special Issue of *Survey Methodology* on the theme `Shaping the future of survey statistics in the data-driven era'. The deadline for submission is 31 January 2026. Submit manuscripts through

https://mc04.manuscriptcentral.com/surveymeth and indicate in the cover letter that the submission is for the special issue.

2. Nominations are open for the 2027 Waksberg Award. This annual invited paper series honors the late Joseph Waksberg with a paper that reviews the development and current state of an important topic in the field of survey statistics and methodology. The recipient receives an honorarium and gives the 2027 Waksberg Invited Address. The paper will be published in an upcoming issue of *Survey Methodology* targeted for December 2027. Send nominations of individual candidates by 15 February 2026 to Paul Smith (**P.A.Smith@soton.ac.uk**). Nominations should include a CV and a letter of nomination and will remain active for 5 years.
3. Applications are open for the Hukum Chandra Memorial Prize. This prize is awarded by the IASS to a mid-career researcher. The recipient will receive an honorarium of 500 Euros and will be invited to present a special webinar with discussion in the IASS Webinar Series in October 2026. Nominations should include an extended abstract (maximum five pages) on the proposed webinar content, comprising original published or unpublished work. Each submission must also be accompanied by a short CV (max two pages). Send applications to Robert Clark (**robert.clark@anu.edu.au**) by 23:59 GMT on 22 May 2026.

Lastly, I am grateful for the opportunity to serve the IASS in this formal capacity. Please feel free to contact me with suggestions for monographs (preferably open access), special issues or edited books on topics of interest to IASS membership.

Jenny Thompson

IASS Scientific Secretary 2025-2027

Jennythompson731967@gmail.com

2026 Waksberg Award

We are pleased to announce that Dr. Frauke Kreuter is the 2026 recipient of the Waksberg Award.

About the Waksberg Award

The journal *Survey Methodology* established an annual invited paper series in 2001 in honor of Joseph Waksberg to recognize his 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 survey statistics and methodology, reflecting the mixture of theory and practice that characterized Joseph Waksberg's work.

Joseph Waksberg was a giant in survey sampling for nearly seven decades, beginning at the U.S. Census Bureau in 1940 and moving to Westat in 1973, where he served as Chairman of the Board from 1990 until his death in 2006. The award includes an honorarium made possible by a grant from Westat.

About Dr. Frauke Kreuter

For the past two decades, Dr. Kreuter has spearheaded novel research in survey methodology, especially at the interface with big data and large-scale computing. Her work on survey paradata served as foundation for the emerging field of adaptive and responsive survey designs, more recently she shaped the discipline's thinking about the connection between surveys and AI.

Through her service on NASEM committees, Dr. Kreuter contributed to establishing principles of privacy protection in federal statistical data products based on combined data sources. She has also directed and built training programs, producing a new generation of data-science-savvy survey researchers.

Dr. Kreuter will give the Waksberg Invited Address at the Statistics Canada Symposium in 2026 and will write a paper planned for publication in the December 2026 issue of Survey Methodology.

Selection Committee

The recipient of the 2026 Waksberg paper was selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*: Jae-Kwang Kim (chair), Kristen Olson, Paul Smith, and Alina Matei.

For more information on the Waksberg Award, please visit:

https://www150.statcan.gc.ca/n1/pub/12-001-x/award-prix-eng.htm

Jae-Kwang Kim

Chair of the 2026 Waksberg Award Committee

Other news

- The call for nominations for the 2027 Waksberg Award is open until February 15, 2026. Paul A. Smith, chair of the 2027 Committee, can be contacted at **P.A.Smith@soton.ac.uk** for further details. For more information on the Waksberg Award, please visit https://www150.statcan.gc.ca/n1/pub/12-001-x/award-prix-eng.htm

- The journal *Statistics in Transition New Series (*https://sit.stat.gov.pl/) has been selected for inclusion in *Web of Science*, which represents one of the most trusted, publisher-independent global citation databases.

Are Non-probability Samples the Future of Surveys?

YES

Changbao Wu
Department of Statistics and Actuarial Science
University of Waterloo, Canada, cbwu@uwaterloo.ca

Non-probability samples are an indispensable part of the future of surveys. It is not because non-probability samples are a preferred source of higher-quality data; rather, it is part of the evolving landscape in the field of survey sampling and official statistics. The ups and downs in the development of probability sampling methods over the past 80 years, the emergence of data from non-traditional sources, and recent methodological advances in dealing with non-probability survey samples have offered a glimpse into the future of the field.

There is no denying that the widespread pursuit of probability samples and the development of probability sampling theory have been part of the feel-good stories of the statistical sciences. Probability sampling and probability samples, however, are a fairy tale of a magic world that is often fractured in reality. There are more philosophical and practical issues with probability samples than steep declines in response rates, sky-rocketing costs, and the inability to meet timely demands. To quote Meng (2022),
"By the time the data arrive at our desk or disk, even the most carefully designed probability sampling scheme would be compromised by the imperfections in execution, from (uncontrollable) defects in sampling frames to non-responses at various stages and to measurement errors in the responses. In this sense, the notion of probability sample is always a theoretical one, much like efficient market theory in economics, which offers a mathematically elegant framework for idealization and for approximations, but should never be taken literally."

It is important to distinguish between a non-probability sample and an arbitrary dataset.

NO

Li-Chun Zhang
University of Southampton, UK & Statistisk sentralbyrå, Norway, L.Zhang@soton.ac.uk

Let me start by removing two potential confusions in order to discuss the "future of surveys". First, non-probability samples are *not* new. In fact, they are ancient — e.g. any population census yields none other than a non-probability sample due to the unknown over-/under-counting errors, and probability sampling (Neyman, 1934) was historically the fruit of scientific evolution from purposively selected non-probability samples (e.g. Kiær, 1896). Second, although "survey" may refer broadly to any purposeful examination of someone or something, for survey statisticians the term is restricted to an observation process that is based on a designed questionnaire (or instrument) which requires informed consent and participation of the data subjects. This may be contrasted to "non-survey" observational big data (Zhang and Haraldsen, 2022), such as administrative registers, transaction records, remote sensing signals, internet webpages (of products, businesses). Despite the lack of a probability design, statistical use of such non-survey big data is both a necessity and an opportunity to be embraced, e.g. in order to address the "official statistics Olympic challenge" (Holt, 2007). The key is *integration* of relevant sources (Zhang, 2012), such as frames of population units, non-survey datasets with complementary or overlapping information, and not least probability sample surveys.

So what I contest here is the value of survey data obtained from non-probability samples, typically web panels, contrary to survey data from probability samples.

Much can be said about the different quality dimensions related to non-probability surveys; but limited space demands focus. From a scien-

Non-probability samples refer to datasets with unknown inclusion mechanisms and/or an unknown sampled population but contain measurements on variables of interest. There needs to be a "design feature" for any non-probability sample to ensure that key study variables and auxiliary variables are included and that an appropriate population is defined. Probability samples with severe nonresponse and/or imperfect sampling frames, samples collected through commercial online or phone panels or through combinations of convenient tools, and incomplete administrative records with relevant information on file are all examples of non-probability samples.

Like it or not, non-probability samples are on the rise and will be a major part of the field's future. However, recent methodological advances unequivocally show that reliable auxiliary information from the target population is the most crucial ingredient of any defensible statistical analysis of non-probability samples. This is where probability samples or census data can fill the gap, and *"a few high-quality national probability surveys with carefully designed survey variables can play a pivotal role in the analysis of non-probability survey samples"* (Wu, 2022).

New data sources will continue to emerge, and the future of surveys will be a blended universe of probability and non-probability samples, with probability sampling theory remaining one of the pillars of statistical frameworks.

References

Meng, X.L. (2022). Comments on "Statistical Inference with Non-Probability Survey Samples" – Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples. *Survey Methodology*, 48, 339–360.

Wu, C. (2022). Statistical Inference with Non-Probability Survey Samples (with Discussion). *Survey Methodology*, 48, 283–311.

tific point of view, the core issue is the *initial* selection problem of non-probability samples, now that survey nonresponse and measurement error are present in probability and non-probability samples alike.

Now, there have been recently a flourish of techniques proposed for the so-called two-sample setting, where the target variable exists in a non-probability sample and some common covariates exist in a separate probability sample additionally. While it is necessary (and potentially helpful) to devise remedies given *incomplete* auxiliary information as such, one must not lose sight of the core selection problem. In fact, in many register-rich countries, it would be easy to replace the additional probability sample entirely by a population frame containing the same covariates. Stripping away the distraction caused by the incompleteness of auxiliary information, one would still be left to confront the initial non-probability selection problem.

In theory, as we know from the history of statistics, there are no guaranteed cures of the selection problem, such as in the context of treatment-control analyses or observational studies. The task-specific judgment required for useful generalisations from *any particular sample* to the population, if taken for granted unwittingly, is detrimental compared to the trust one can rightly place in transparent, target-agnostic inference from probability *sampling*. It serves well to remind us on this point that Neyman (1934) called "the method of sampling representative", not that any particular sample can ever be representative.

Moreover, practical speaking, any adjustment technique of non-probability selection may as well be considered for survey nonresponse in probability samples, and empirical studies so far have only evidenced increasing risks of bias when comparing "well built" non-probability samples to "low response rate" probability samples (Dutwin and Buskirk, 2017).

Of course, decreasing response rates in probability samples and increasing costs thereby are serious challenges that need to be handled by continuously improving the survey methodology. Multisource statistics based on non-survey big data have provided many alternatives in the past and will become even more important in future. But the transition has been and will be gradual, especially in official statistics due to the high quality requirements. Adopting design-based audit sampling as a standard for validation and quality assessment is attractive in this respect due to its transparent probability-inference basis (Zhang, 2021, 2023).

In other words, sample survey will remain a valuable method of statistical investigation in future, but only if it is based on probability sampling to start with.

References

Dutwin, D. and Buskirk, T.D. (2017). Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples. *Public Opinion Quarterly*, 81:213-239, `https://doi.org/10.1093/poq/nfw061`

Holt, T. (2007). The official statistics Olympic challenge: wider, deeper, quicker, better, cheaper. *The American Statistician* 61:1-8.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, IX(2):176-183.

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.

Zhang, L.-C. (2023). Audit sampling as a quality standard for multisource official statistics. *Spanish Journal of Statistics*, 5:67-83, `https://doi.org/10.37830/SJS.2023.1.05`

Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, 184:571-588, `https://doi.org/10.1111/rssa.12632`

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66:41-63.

Zhang, L.-C. and Haraldsen, G. (2022). Secure Big Data Collection and Processing: Framework, Means and Opportunities. *Journal of the Royal Statistical Society Series A*, 185:1541-1559, https://doi.org/10.1111/rssa.12836

Differential Privacy and its Application to Survey Data

Anne-Sophie Charest[1] and Jörg Drechsler[2,3,4]

[1] Université Laval, Canada, anne-sophie.charest@ulaval.ca
[2] Institute for Employment Research, Germany, joerg.drechsler@iab.de
[3] Ludwig-Maximilians-Universität (LMU), Munich, Germany
[4] University of Maryland, USA

Abstract

Differential privacy has emerged as a rigorous and broadly applicable framework for protecting confidential data, offering guarantees that do not depend on unverifiable assumptions. In this paper, we first present the definition of differential privacy and explain how it can be achieved in simple settings using standard mechanisms. We then examine the application of differential privacy to survey data and outline five key issues that complicate its use in this context.

Keywords: differential privacy, confidentiality, surveys, sampling, weighting, imputation.

1 Introduction

Confidentiality for survey and census data has long been a central concern. C*onfidentiality in this context means protecting participants' identities and data by keeping it private and secure, preventing unauthorized access, and only reporting results in aggregated forms to build trust and encourage honest answers, especially for sensitive topics. These goals are often achieved through a combination of techniques such as anonymization, encryption, strict access controls, and clear communication of data usage.* Focusing on anonymization, many techniques have been used to limit the risk of disclosing private information: data suppression, swapping, data perturbation, and, more recently, synthetic data (see for example Hundepool et al., 2012). These approaches aim to reduce disclosure risks but applying them effectively requires deciding when the risk is acceptable. A previous paper in this newsletter (Shlomo, 2022) reviewed traditional disclosure risks, namely identity, attribute, and inferential disclosure, and described how statisticians have estimated these risks over several decades. It also briefly introduced alternative privacy models proposed by computer scientists, such as differential privacy (DP).

DP provides guarantees that differ fundamentally from assumption-based risk estimates. Indeed, traditional disclosure risk metrics depend on unverifiable assumptions regarding the knowledge and capabilities of ill-intended users of the data, henceforth called attackers, that try to learn sensitive information about the units included in the data. Because of these assumptions traditional risk metrics can fail when new external data becomes available, whereas DP offers provable protections that do not rely on assumptions about the attacker's knowledge. Although the idea originated in computer science, an increasing number of statisticians are actively contributing to research in the field, motivated in part by the Census Bureau's adoption of DP to protect data from the 2020 U.S. Census (Abowd, 2018) and by the appeal of its clean theoretical guarantees.

However, the use of DP in the context of surveys is not straightforward. In fact, we will discuss in this paper five key issues that complicate this application. But, first, we present the DP framework in detail and outline a few methods to achieve this guarantee.

2 Differential Privacy

DP was first proposed by Dwork et al. (2006) and has been a very active research area since, particularly in the last ten years or so. The term now encompasses a broad family of definitions, such as (ε, δ)-DP, concentrated DP, and Rényi DP, each designed to address specific analytic or operational needs. We present the original definition in detail and provide references to a few important variants below. In the following section, we will explain how one can achieve DP, for example with the addition of carefully selected noise to a statistic of interest.

2.1 Pure DP

This is the original definition, now referred to as pure DP. It is important to understand that DP is the property of an algorithm, which takes as input a dataset to generate some output (for example, a statistic, a parameter estimate or even a synthetic dataset) and not the property of a specific output. This algorithm is usually called a randomized mechanism because satisfying the DP constraint generally requires the addition of randomness. This randomness plays a crucial role: it ensures that the mechanism's outputs cannot depend too heavily on any single individual's data. More precisely, a randomized mechanism M with output space S satisfies ε-differential privacy for a given privacy parameter $\varepsilon > 0$ if and only if for any two neighboring datasets D and D' and any $A \subseteq S$, we have that $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A]$.

To illustrate, we can look at a class of algorithms that achieve DP by adding independent discrete random noise to the statistic of interest, say the population total, i.e., instead of reporting the true total the algorithm would return a noisy total where the noise is chosen in such a way that the probability that the algorithm returns a specific value $t \in A$ if dataset D was used as the input is very close to the probability of returning the same value if dataset D' was used as the input. How close these two probabilities need to be is governed by the parameter ε.

Note that neighboring datasets can be defined in different ways, but the key idea is that they differ in the data of a single individual. For a classic tabular dataset where rows represent observations and columns represent variables, datasets D and D' are neighbors if they differ by exactly one row. More precisely, we talk about unbounded DP if D' is obtained by adding or removing one row from D, and bounded DP if D and D' have the same number of rows but differ in the values of one individual's record. There are subtle but important differences between these two definitions; see for example chapter 2 of Li et al. (2017). Other data structures require alternative notions of neighboring datasets. For example, in network data, one may define neighbors by removing or adding a single edge (Nissim et al., 2007) or a single node (Kasiviswanathan et al., 2013).

The guarantee offered by the pure DP definition can be interpreted in several ways. One is plausible deniability: an individual can claim that their data has any value, and the output of a DP mechanism cannot be used to refute that claim, even if an adversary holds as much information as the entire dataset except for that individual. This is because adding any row to this known dataset creates a neighboring dataset, and under DP the mechanism's output must be almost as likely under each of these possibilities. Consequently, no observer can reliably determine which specific values the individual contributed.

Another interpretation, given in Wasserman and Zhou (2010), is in the form of a hypothesis test. Pure DP implies a strict limit on how well any statistical test can distinguish whether the mechanism's output came from D or from D'. Specifically, for any test of level α, the power of that test must be smaller or equal to $e^{\varepsilon}\alpha$, so that the power of the test is very similar to its level. Thus, under ε-differential privacy with sufficiently small values of ε, even the most powerful test cannot reliably determine which of the two neighboring datasets produced the observed output, ensuring privacy for that individual.

Another important aspect of DP is the set of useful properties that follow directly from the definition. First, DP is immune to post-processing, meaning that any computation applied to the output of a DP mechanism will preserve the privacy guarantee. Second, DP composes in a straightforward way: when several DP mechanisms are applied to the same dataset, their privacy losses accumulate in a mathematically quantifiable manner, allowing to keep track of the privacy loss over multiple data releases. For example, k mechanisms that each satisfy ε-differential privacy jointly satisfy at most $k\varepsilon$-differential privacy. Because of this composition, the parameter ε is sometimes also referred to as a privacy budget; it defines the total amount of privacy leakage that is still considered acceptable. Based on the composition property one can then decide how this privacy budget should be spent across several outputs from a single dataset. More details on these properties can be found in Dwork and Roth (2014).

2.2 Approximate DP

Pure DP is a very strict guarantee, concerned with the worst-case scenario, because the inequality $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A]$ must hold for any possible D and D', even very implausible ones. A variant allows the guarantee not to hold when the probabilities of an output are small. More precisely, a randomized mechanism M is said to satisfy (ε, δ)-differential privacy with $\varepsilon > 0$ and $\delta > 0$ if for any two neighboring datasets D and D' and for any $A \subseteq S$ we have that $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A] + \delta$. This variant is more frequently used than pure DP, which is the special case where $\delta = 0$, and often is referred to as simply DP.

2.3 Other variants

Many other variants of differential privacy have been proposed over the years. These alternatives might modify the definition of neighbouring datasets or the way the privacy loss is measured. Desfontaines and Pejó (2020) surveys hundreds of such definitions inspired by DP. A few variants are worth mentioning: Rényi DP (Mironov, 2017), widely used in machine learning and in DP stochastic gradient descent, zero-concentrated DP (Bun and Steinke, 2016), which offers tighter composition bounds and Gaussian DP, which offers an analytically tractable, hypothesis-testing–based framework with tight composition rules (Dong et al., 2022).

Another active line of work focuses on settings with no trusted curator, where privacy must be guaranteed at the user level, that is before the data is stored in a central database (see for example. Kasiviswanathan et al., 2011). This local differential privacy model is used in practice, for example, in large-scale telemetry systems (Apple, 2017).

3 Achieving DP

DP is typically achieved through the addition of randomness. There are a few building block mechanisms, which are often combined to obtain mechanisms for more complex tasks. These are described in Dwork and Roth (2014) and summarized below.

3.1 Noise addition for numeric outputs

Noise addition is the basic building block of many differentially private algorithms. Under pure DP, the standard approach is the Laplace mechanism. Suppose you want to release the output of some function f applied to a dataset D, the Laplace mechanism will add Laplace noise to the value of $f(D)$. The variance of the added noise depends on the global sensitivity of the function, which is the maximum possible change in the value of f when computed on any two neighboring datasets D and D', that is, the largest difference $|f(D) - f(D')|$ over all such pairs. More precisely, the Laplace mechanism for a function f releases $f(D) + X$, where X is drawn from a Laplace distribution with mean 0 and scale equal to the global sensitivity of f divided by ε. The sensitivity must be computed for each function f. For instance, the sensitivity of a counting query is 1, while the sensitivity of the mean depends on the range of the possible values for the individual values. For a dataset for which the size n can be treated as public knowledge, each observation is in $[a, b]$, then the range is $R = b - a$ and the sensitivity of the mean is R/n. Note that if we cannot provide bounds for the individual values, then the sensitivity will be infinite, and thus it will not be possible to achieve DP. In practice, one may estimate the range from the observed data, but this will require spending some of the privacy budget.

For approximate DP, the standard mechanism is to add Gaussian noise, with variance determined by the global sensitivity of the function and the privacy parameters ϵ and δ. Other variants include adding geometric noise, or discrete or truncated noise distributions. Extensions also exist for multidimensional outputs, with mechanisms designed to handle vector-valued or high-dimensional functions.

3.2 Exponential mechanism for non-numeric outputs

Noise addition works well for numeric outputs, but many tasks require selecting from a set of categorical or structured outcomes. The exponential mechanism is a second fundamental building block that provides a general framework for releasing non-numeric outputs under differential privacy. It selects an output r with probability proportional to $\exp(\varepsilon u(D,r) \,/\, (2\Delta u))$, where $u(D,r)$ is a utility score for reporting r on dataset D and Δu is the sensitivity of this utility score. This mechanism is especially useful when the goal is to select the "best'' option according to a data-dependent criterion, such as choosing a model or a quantile, while ensuring that the choice does not reveal too much about the underlying data. For example, one may use the exponential mechanism to publish the mode of a categorical variable by using the number of observations in dataset D with value equal to r as utility function $u(D,r)$. This utility function has sensitivity $\Delta u = 1$.

3.3 More complicated mechanisms

Most DP mechanisms are constructed from these basic building blocks, together with the composition and post-processing properties of DP. For tasks such as regression, for example, one may add noise directly to the data, perturb the objective function, or add noise to the final output, or even decide to use more robust statistics to reduce the amount of noise required (see for example Alabi et al., 2022). The optimal strategy is problem-dependent, and in many settings remains an active area of research.

In machine learning, using differentially private versions of stochastic gradient descent (DP-SGD) has become the dominant approach for training models with differential privacy. The privacy guarantees of these algorithms rely on privacy accounting. Simple composition is far too loose when models are trained over tens of thousands of gradient steps. Privacy accounting methods provide tight bounds by exploiting subsampling amplification (Balle et al., 2018) and advanced composition frameworks such as Rényi DP (RDP), zero-concentrated DP (zCDP), and Gaussian DP (GDP). Without such accounting techniques, training would appear to consume impractically large privacy budgets, rendering DP-SGD useless in practice. Several accounting methods exist (Abadi et al., 2016; Bun and Steinke, 2016; Mironov, 2017; Dong et al., 2022; Koskela et al., 2020), each trading off accuracy, efficiency, and ease of implementation.

3.4 Practical challenges

Implementing DP in practice raises several important challenges. A first difficulty is choosing the privacy parameter ε. Although DP offers a formal privacy guarantee, it is really only meaningful if ε is relatively small. There is little consensus on what values are acceptable in applied contexts, and existing legal or regulatory frameworks such as the European Data Protection Regulation (GDPR), (Regulation (EU) 2016/679) offer only high-level guidance (Lee and Clifton, 2011).

A second challenge is that theoretical guarantees do not always translate cleanly to real-world implementations. Floating-point arithmetic, numerical clipping, and implementation-level randomness can all introduce deviations from the idealized model. Even subtle issues in pseudorandom number generators can weaken privacy guarantees, as demonstrated in early attacks on DP implementations (Mironov, 2012). Robust software engineering is therefore essential. Although mature libraries such as Google's Differential Privacy library (Google, 2019), OpenDP (OpenDP Project, 2021), and IBM's diffprivlib (Holohan et al., 2019) mitigate many of these risks, ensuring trustworthy and reproducible implementations remains an active area of work.

Finally, dependence within the data and adaptivity in the analysis process introduce additional complexities. Differential privacy is defined for datasets differing in one individual assuming independence between the units, but real datasets may exhibit strong correlations, for instance, between members of the same family, which can increase effective sensitivity and weaken protection (Kifer and Machanavajjhala, 2011). Adaptive analyses, where later queries depend on earlier outputs, also complicate privacy accounting. For example, model diagnostics such as residual checks or comparisons of fit statistics should be handled carefully, and obtaining them under DP will consume additional privacy budget (Dwork et al., 2015).

4 DP for surveys

The following discussions are excerpts from Drechsler and Bailie (2024) and we refer interested readers to this text for a more detailed discussion of DP in the survey context. When working with survey data, there are additional complexities which typically do not arise in other settings. Moreover, the implications of using DP in the context of surveys have received little attention in the DP literature until recently. Overall, there are (at least) five aspects that need to be considered when implementing DP in this context: (i) the multiple stages of the survey pipeline, (ii) limited privacy gains from complex sampling designs, (iii) challenges in computing the privacy guarantees of survey weighted estimates, and consequences of (iv) weighting adjustments and (v) imputations for missing data. We will discuss each of these aspects in the remainder of this section.

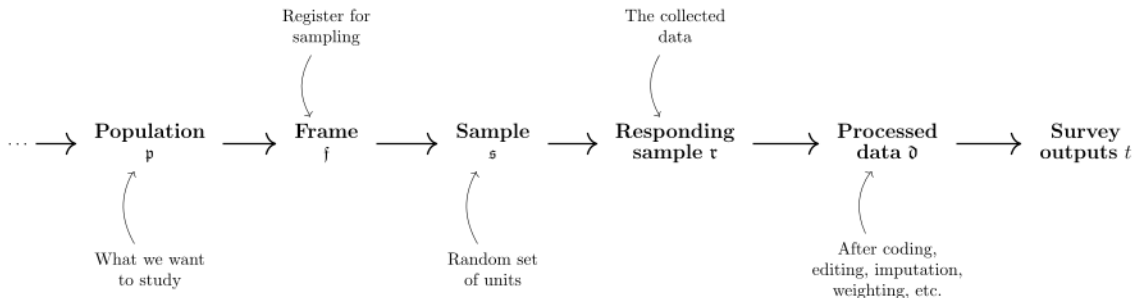


Figure 1: The main steps of the survey data pipeline

4.1 DP and the Survey Data Pipeline

As illustrated in Figure 1, the production of survey data is a complex multistage process. There are two important considerations when integrating a DP mechanism into a data pipeline. Firstly, at what point in the pipeline should the DP mechanism start? And secondly, which of the earlier stages of the data pipeline should be considered invariant? In the DP literature, invariants generally refer to aspects of the input data that remain fixed over neighboring dataset *D* and *D'*. For example, for the Decennial Census 2020, the U.S. Census Bureau decided that several counts must be released unaltered and thus treated them as invariant in their application of DP. With survey pipelines, there are several possible options with respect to the starting point of the mechanism and the decision on which of the earlier stages should be treated as invariant.

In the option most seen in the DP literature, the data-release mechanism starts at the end of the pipeline and performs just the last step – computing the survey outputs from the processed data – and none of the previous steps are taken as invariant. However, a mechanism could conceivably start at any point of the survey pipeline and incorporate all the steps that follow. Furthermore, any of the steps before the mechanism starts could conceivably be taken as invariant. Overall, this leads to up to 15 possible scenarios that need to be considered. Figure 2 highlights ten of these scenarios for illustration (the remaining five options would all start at the responding sample level).

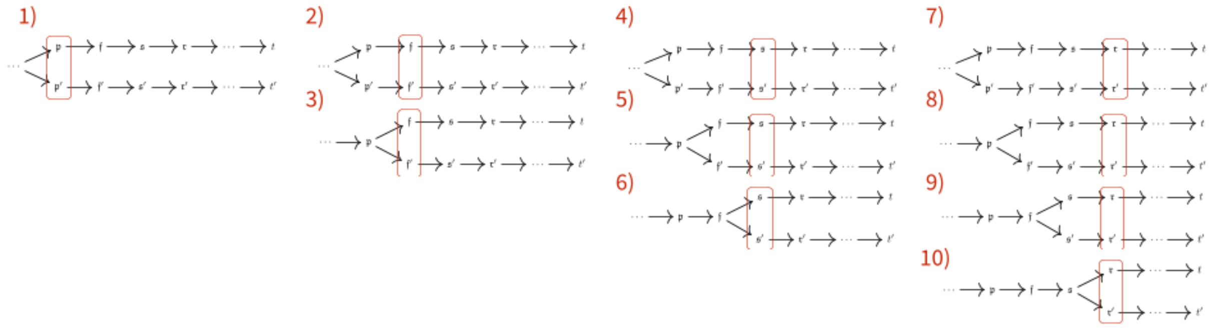


Figure 2: Ten out of fifteen possible settings for a DP mechanism in the survey context (the remaining five settings would all start at the level of the processed data). The red rectangles indicate the starting point of the mechanism.

A few general observations can be made regarding the advantages and disadvantages of the different scenarios (see Drechsler and Baile (2024) for a more detailed discussion): From a privacy perspective, it seems advantageous to start the DP mechanism as early as possible to benefit as much as possible from privacy amplification through subsampling (Balle et al., 2018), further discussed in Section 4.2 below. Note that the data production pipeline consists of three sampling steps: beyond the classical sampling step, nonresponse can be treated as another subsampling step and even the frame can be seen as a random sample from the target population, if we model the probability of inclusion in the frame as a random variable. However, this privacy amplification can be nullified if the attacker knows that the record they are attacking is in the sample, a scenario that statistical agencies often need to consider in practice and the additional privacy amplifications are either small or difficult to quantify (Bailie and Drechsler, 2024). On the other hand, any stage of the survey pipeline that should be part of the DP mechanism must first be fully "algorithmized" (that is, the process by which each of the stage's possible inputs is transformed into one of its outputs must be completely and programmatically specified). A survey pipeline often includes a number of complex, ill-defined and human-intensive tasks, such as building the frame, choosing a sampling design, coding and editing. Because these tasks all usually require a degree of human judgment, they would be difficult to algorithmize.

Another downside of starting the DP mechanism earlier is the fact that it can complicate the computation of the cumulative privacy loss across multiple data-release mechanisms because DP's composition theorems are not applicable when there is dependence between the mechanisms' noise terms (which can happen, for example, when their sampling designs are dependent or when two noisy statistics are computed from the same sample) (Bailie and Drechsler, 2024).

However, even if a data-release mechanism begins later in the survey pipeline so that some steps of the pipeline do not have to be incorporated in the mechanism, implementing DP still requires understanding those steps' effect on the mechanism's input data. For example, with hot deck imputation an individual survey respondent can contribute to multiple records in the post-imputation dataset. This complicates the appropriate definition of neighboring datasets: In the post-imputation dataset, changing a single record does not correspond to changing the data of one entity. In general, the later the DP mechanism begins, the more difficult it is to determine an appropriate notion of neighbors since steps earlier in the pipeline may introduce dependencies between dataset records.

These complexities demonstrate that there can be conflicting demands in deciding where a DP mechanism should start within the survey pipeline. (See Drechsler and Bailie (2024) for further aspects that need to be considered.) We now return to the question of which steps of the survey pipeline should DP take as invariant. DP assesses the privacy of a data-release mechanism by comparing the survey outputs' distribution under pairs of counterfactual input datasets. By taking some of the steps of the survey pipeline as invariant, DP's counterfactual comparisons are reduced to only those pairs of input datasets which share the same realization of the invariant steps. For example, suppose the steps in the survey pipeline which generate the population and the frame are taken as invariant and the data-release mechanism starts with the responding sample. Then DP only compares those responding samples (i.e., those counterfactual input datasets) which could have come from the same frame. Adding invariants will weaken the privacy guarantees provided by DP (Kifer et al., 2022, Abowd et al., 2022). In general, the later the stage of the pipeline that is kept invariant, the greater the reduction in privacy. However, invariants may be justifiable when the output of the invariant steps can be considered public knowledge (such as if the frame was sourced commercially rather than constructed from confidential information). Moreover, constraining some steps to be invariant has the advantage of reducing the sensitivity of survey weighted estimators and thereby decreasing the noise which must be added for privacy protection as discussed in Section 4.3.

4.2 DP with Complex Sampling Designs

Statistical agencies have been aware for decades that sampling can be a simple and effective strategy to reduce disclosure risks simply because an attacker can no longer be sure whether a specific target record is included in the sample or not. This is the main reason why most statistical agencies only release samples from their censuses as public use micro datasets (they typically also apply additional measures to further increase the level of protection). This idea has been formalized in several papers in the context of DP (Kasiviswanathan et al., 2011, Wang et al., 2016, Bun et al., 2015, Balle et al., 2018, Wang et al., 2019). The authors show that the level of privacy is amplified through sampling, i.e., the actual privacy guarantees are higher than those implied by the chosen privacy loss parameters when protecting the sample output. Specifically, for small sampling rates r and small privacy loss parameters ε, applying certain simple sampling designs (simple random sampling with and without replacement, and Poisson sampling) before running an ε-DP mechanism reduces the privacy loss to approximately $r\varepsilon$. However, most surveys conducted by statistical agencies use complex multistage sampling designs, potentially with different sampling strategies at the different stages. Bun et al. (2022) study the amplification effects for complex designs and find that amplification is small for most of the sampling designs used in practice. Their findings can be summarized as follows:

- Cluster sampling using simple random sampling without replacement to draw the clusters offers negligible amplification in practice except for small ε and very small cluster sizes.
- With minor adjustments, stratified sampling using proportional allocation can provide privacy amplification. For small ε, the amplification is still linear in the sampling rate up to a constant factor.

- Data dependent allocation functions such as Neyman allocation for stratified sampling will likely result in privacy degradation. (The effects will depend on the sensitivity of the allocation function.)
- With PPS sampling at the individual level, the privacy amplification will linearly depend on the maximum probability of inclusion (for small ε).
- Systematic sampling will only offer amplification if the ordering of the population is truly random. In all other cases, systematic sampling will suffer from the same effects as cluster sampling, leading to no amplification (assuming the ordering is known to the attacker).

In practice this implies that for many multistage sampling designs, which typically start with (multiple stages of) stratified cluster sampling, amplification effects can generally only be expected from those stages at which individual units or households are selected (typically the last stage of selection).

4.3 DP for Weighted Estimates

As discussed in Section 3.1, the amount of noise that needs to be added to achieve a specific privacy loss ε directly depends on the sensitivity of the statistic of interest. From a utility perspective, this implies that more reliable (less noisy) DP outputs can be expected from statistics with low sensitivity.

When analyzing survey data, it is generally important to take the sampling design into account since the probabilities of selection vary between the units included in the sample. To simplify this task for data users, statistical agencies typically provide survey weights. In practice, these survey weights will also account for nonresponse and other data deficiencies such as undercoverage. (We will address this extra layer of complexity in the next section.)

Using survey weighted estimates raises the question: how (if at all) does the sensitivity of a statistic change when the survey design is taken into account? To illustrate the possible impacts, let us assume the analyst is interested in estimating the mean of some variable Y in the population using the sampled values $y_i,\ i = 1, \ldots, n$, where n denotes the sample size. If the probabilities of selection were equal for all units, the sample mean would be an unbiased estimate for the population mean and, as indicated in section 3.1, its sensitivity would be R/n, with R denoting the range of possible values for y_i.

When dealing with unequal probabilities of selection, a popular estimator for the population mean is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952): $\widehat{\mu_Y^{HT}} = \sum w_i y_i/N$ where w_i is the weight of unit i, for $i = 1, \ldots, n$ and N is the size of the population. Note that we assume for simplicity that N is known and does not need to be protected and w_i is the design weight, i.e., it only accounts for the sampling design.

If we can treat the weights as fixed, the sensitivity of $\widehat{\mu_Y^{HT}}$ is $\max(w_i)\, R/N$. Whether the maximum is over all units in the frame, over all units in the population, or over all possible counterfactual units, depends on which stages of the survey pipeline are treated as invariant as discussed in Subsection 4.1. Note that for equal-probability designs all $w_i = N/n$ and thus the sensitivity of the Horvitz-Thompson estimator is the same as for the unweighted estimator. If $\max(w_i) > N/n$, the Horvitz-Thompson estimator will have larger sensitivity than the unweighted estimator.

However, these discussions assume that the weights can be treated as fixed, that is, they do not change if a record changes in the database. For most sampling designs used in practice, such an assumption is unrealistic. For example, with sampling proportional to size (PPS), the i-th record's probability of inclusion is given by $\pi_i = (nx_i)/(N\bar{x})$, where x_i is the value for unit i of the measure-of-size variable X that is used to improve the efficiency of the sampling design, and $\overline{x} = \sum_{i=1}^{N} x_i / N$ is the population mean of X. Changing the value of X for a single record will change the probabilities of inclusion and thus the survey weights for all other records in the sampling frame. Therefore, the sensitivity will be larger compared to the setting with fixed weights as we no longer only need to consider the maximum possible change in a single record's value for Y. We also need to consider the impact of the weight change for all the other records even if their values for Y don't change.

A recently proposed strategy to mitigate this potentially substantial increase in sensitivity is to regularize the weights, as explored by Seeman et al. (2024). (An extreme version of this strategy would set all weights to be equal; this could be justifiable if the increase in the privacy noise due to the weights dwarfs the bias introduced by ignoring the sampling design.) Another possible strategy is to treat the frame or at least the design variables within the frame as invariant as discussed in Figure 2. Frame invariance assumes any two neighboring datasets must always originate from the same frame and so can only differ at the sample level (or later). However, treating the frame as invariant has two additional implications that need to be considered. First, fixing the frame implies that privacy amplification from sampling is no longer possible (we would need to have neighboring datasets at the frame level in order to achieve amplification). However, given the results of Bun et al. (2022), this amplification is likely small in practice and thus the positive effects of reducing the sensitivity will tend to outweigh the negative effects of losing the amplification effect. On the other hand, fixing the frame will restrict the possible counterfactual input datasets to those which are consistent with the realized frame. Because this restriction will fix the survey weights, it might introduce strong constraints on the possible neighboring datasets, depending on the sampling design. As a consequence, the actual privacy guarantees for a frame invariant setting could be significantly weaker than the guarantees under a non-frame-invariant setting even for the same privacy loss parameter. How problematic this reduction in privacy is in real settings is currently an open question for research.

In contrast, if only the design variables are treated as fixed, the data-release mechanism could still start at the frame level, strengthening the privacy guarantees.

4.4 DP and Weighting Adjustments

In practice, two adjustment steps are commonly applied to the design weights to correct for unit nonresponse and other data deficiencies such as over- or undercoverage in the sampling frame: nonresponse adjustments and calibration. How these adjustment steps interfere with differential privacy has not been studied so far. However, both steps are data dependent, that is, they use information from the survey units for the adjustments. This implies that these steps cannot be ignored from a privacy perspective as the adjusted weights leak some personal information. Looking at the impacts on the sensitivity of the final statistic of interest (which uses the adjusted weights), similar problems as those discussed in the previous section will arise: changing one record in the database can potentially change the weight-adjustment factors for all other units in the survey. Thus, it seems imperative to already account for these adjustment steps during data pre-processing. Better results in terms of the privacy-accuracy trade-off might be achieved if the weight-adjustment steps were carried out in a differentially private way. More research is needed to better understand this trade-off. For example, it seems beneficial to identify robust adjustment strategies as less noise would be required to satisfy DP for these strategies.

In the particular case of post-stratification (which is a simple type of calibration), one such robust adjustment strategy has been proposed by Clifton et al. (2023). Another strategy would be to regularize the nonresponse and calibration weight adjustments. (This would be similar to the survey weight regularization strategy of Seeman et al. (2024) discussed in the previous section.)

4.5 DP and Imputation

All survey data are plagued by item nonresponse as survey respondents are often unwilling or unable to respond to all survey questions especially if they request sensitive information. A common strategy to deal with this problem is to impute the missing values before analyzing the data to avoid biases that might arise when simply discarding incomplete records before the analysis. However, imputations are always data dependent as they typically build a model based on the observed data and use this model to impute the missing values. As a consequence, the implications of imputation on the DP guarantees need to be considered regardless of whether or not the imputation procedure is included inside the data-release mechanism. Some preliminary results for this problem are discussed in Das et al. (2022).

Similar to the problem of weighting adjustments, there are two possible strategies to account for imputation under DP. The first strategy only considers the effects when analyzing the imputed data. The second strategy modifies the imputation routines to ensure that the imputations already satisfy DP. As Das et al. (2022) have shown, the first strategy implies that in the worst case the sensitivity increases linearly with the number of imputed observations. This substantial increase of the sensitivity arises because changing one record in the database can potentially impact all of the imputed values. Whether the worst case applies depends on the analysis of interest and on the selected imputation procedure. Still, for statistical agencies offering pre-imputed datasets for accredited researchers, this strategy is not an option since they cannot anticipate which analyses might be performed on the imputed data.

The second strategy can break the dependence on the number of imputed records at least for certain imputation strategies. The key requirement for breaking the dependence is that the imputation model m can be written as $D_{imp}^{(i)} \sim m\left(D_{obs}^{(i)}, \hat{\theta}\right)$, where $D_{imp}^{(i)}$ and $D_{obs}^{(i)}$ contain the imputed and observed variables for record i and $\hat{\theta}$ denotes the model parameters estimated on the complete data. The model implies that, given $\hat{\theta}$, the imputed values of record i only depend on the observed values of that record and not on any other record. If these requirements are met and the parameters θ of the imputation model are estimated using any suitable differentially private mechanism with privacy loss parameter ε_1, then, given any ε_2 differentially private mechanism used for analyzing the data, the overall privacy loss is given by $\varepsilon_1 + \varepsilon_2$ by the composition property

We note that the conditional independence assumption of the imputation model holds for many imputation methods, for example, parametric imputation models based on linear regression. However, it does not hold for hot-deck imputation, an imputation method commonly applied at statistical agencies.

5. Discussion

Differential privacy provides a formal, elegant framework with strong theoretical guarantees and several appealing properties such as post-processing immunity, clean composition rules, and a clear interpretation of privacy loss. In simple settings, these guarantees are straightforward to compute and the required noise is easy to calibrate. However, real data-analysis workflows are rarely this tidy, and its application to survey data highlights just how complex differential privacy can become in practice. The presence of complex sampling designs, weighting adjustments, imputation steps, and data-dependent decisions means that calibrating privacy loss is rarely straightforward. Each of these operations can interact with DP in subtle ways, and determining how much noise is needed, or even what the appropriate sensitivity should be, poses challenges that haven't been fully addressed in the literature.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L. (2016, October). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308-318.

Abowd, J. M. (2018). The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2867-2867.

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022). The 2020 Census disclosure avoidance system TopDown Algorithm. In: Harvard Data Science Review (Special Issue 2).

Alabi, D., McMillan, A., Sarathy, J., Smith, A. and Vadhan, S. (2022). Differentially Private Simple Linear Regression. *Proceedings on Privacy Enhancing Technologies*.

Apple Differential Privacy Team. (2017). *Differential privacy technical overview*. Apple Inc. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Bailie, J. and Drechsler, J. (2024). Whose Data Is It Anyway? Towards a Formal Treatment of Differential Privacy for Surveys. In: National Bureau of Economic Research (2024): Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences, Spring 2024, Washington, 33 p. URL: https://conference.nber.org/conf_papers/f194306.pdf.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In: Advances in Neural Information Processing Systems, 31.

Bun, M., and Steinke, T. (2016) Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Theory of cryptography conference. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. (2015). Differentially private release and learning of threshold functions. In: Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science, FOCS '15, Washington, DC, USA, IEEE Computer Society, 634–649.

Bun, M., Drechsler, J., Gaboardi, M., McMillan, A., and Sarathy, J. (2022). Controlling privacy loss in sampling schemes: An analysis of stratified and cluster sampling. In: Foundations of Responsible Computing (FORC 2022), 24 p.

Clifton, C., Dajani, A. N., Hanson, E. J., Clark, S., Merrill, K., Merrill, S., and Rodriguez, R. (2023). Preliminary report on differentially private post-stratification. Working Paper ced-wp-2023-004, U.S. Census Bureau. URL: https://www.census.gov/library/working-papers/2023/adrm/ced-wp-2023-004.html

Das, S., Drechsler, J., Merrill, K., and Merrill, S. (2022). Imputation under differential privacy. Preprint: arXiv:2206.15063.

Desfontaines, D. and Pejó, B. (2020). SoK: Differential privacies. *Proceedings on Privacy Enhancing Technologies*.

Dong, J., Roth, A. and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(1), 3-37.

Drechsler, J., and Bailie J. (2024). The Complexities of Differential Privacy for Survey Data. NBER working paper / National Bureau of Economic Research 32905. To appear in: R. Gong, V. J. Hotz, I. M. Schmutte (Eds.), Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and Their Consequences, 18 p. doi: 10.48550/arXiv.2408.07006.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, *9*(3–4), 211-407.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636-638.

Google. (2019). Google differential privacy library [Computer software]. GitHub. https://github.com/google/differential-privacy

Holohan, N., Braghin, S., Mac Aonghusa, P. and Levacher, K. (2019). Diffprivlib: the IBM differential privacy library. arXiv preprint arXiv:1907.02444.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. In: Journal of the American Statistical Association, 47(260), 663–685.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and De Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith. A. (2011). What can we learn privately? In: SIAM Journal on Computing, 40(3), 793–826.

Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S. and Smith, A. (2013). Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, Berlin, Heidelberg: Springer, 457-476.

Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., and Zhuravlev, P. (2022). Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census. Technical Report: arXiv:2209.03310.

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 193-204.

Koskela, A., Jälkö, J. and Honkela, A. (2020). Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, PMLR, 2560-2569.

Lee, J. and Clifton, C. (2011). How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, Springer Berlin Heidelberg, 325-340.

Li, N., Lyu, M., Su, D. and Yang, W. (2017). *Differential privacy: From theory to practice*. Morgan & Claypool.

Mironov, I. (2012). On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, 650-661.

Mironov, I. (2017) Rényi differential privacy. In: Proceedings of the 30th IEEE computer security foundations symposium (CSF). IEEE, Santa Barbara, 263-275.

Nissim, K., Raskhodnikova, S. and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 75-84.

OpenDP Project. (2021) The OpenDP Library (version 0.3) Harvard University. https://opendp.org

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4 May 2016, p. 1–88.

Seeman, J., Si, Y., and Reiter, J. P. (2024). Differentially private population quantity estimates via survey weight regularization. In: National Bureau of Economic Research (2024): Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences, Spring 2024, Washington, 33 p. URL: https://www.nber.org/system/files/chapters/c15023/c15023.pdf

Shlomo, N. (2022). How to measure disclosure risk in microdata?. *The Survey Statistician*, *86*(2), 13-21.

Wang, Y., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled R´enyi differential privacy and analytical moments accountant. In: The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan, PMLR, 89, 1226–1235. URL: http://proceedings.mlr.press/v89/wang19b.html.

Wang, Y.-X., Lei, J., and Fienberg, S. E. (2016). Learning with differential privacy: Stability, learnability, and the sufficiency and necessity of ERM principle. In: Journal of Machine Learning Research, 17(183),1–40.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, *105*(489), 375-389.

Measures of Non-Ignorable Selection Bias for Non-Probability Samples

Sergio D. Martinez[1], **Brady T. West**[2], **and Rebecca R. Andridge**[3]

[1]PhD(c) in Survey and Data Science, University of Michigan, sergiomm@umich.edu
[2]Research Professor, Survey Research Center of the Institute for Social Research, University of Michigan, bwest@umich.edu
[3]Professor, Division of Biostatistics in the College of Public Health, The Ohio State University, andridge.1@osu.edu

Abstract

Non-probability samples are increasingly used in applied research, raising concerns about non-ignorable selection bias in estimates based on these samples that traditional diagnostics cannot adequately assess. Conventional diagnostics and inferential approaches for these samples offer limited insight because they ignore the link between selection mechanisms and outcomes. This paper reviews variable-dependent measures for non-ignorable selection bias based on the proxy pattern–mixture model (PPMM), with emphasis on the Standardized Measure of Unadjusted Bias (SMUB) for means and the Measure of Unadjusted Bias for Proportions (MUBP). Both indices are grounded in the PPMM framework, which uses auxiliary variables with reliable population benchmarks to construct a single proxy and summarize departures from ignorability through a single sensitivity parameter. Evidence from simulation studies shows that the SMUB and MUBP can accurately capture the direction and magnitude of bias when auxiliary variables are at least moderately predictive of outcomes, outperforming traditional diagnostics. Empirical applications in health surveys, political polling, pandemic surveillance, and administrative data confirm their practical value while emphasizing the need for strong, harmonized auxiliary covariates. We conclude with guidance on implementation and a brief discussion of ongoing research. Our aim is to motivate broader adoption of these indices as practical and interpretable diagnostics for assessing selection bias in applied survey research, aided by accessible `R` software that facilitates their implementation in practice.

Keywords: selection bias, non-probability samples, proxy pattern–mixture model, sensitivity analysis.

1 Introduction

The cornerstone of survey inference has long rested on a fundamental assumption: that the mechanism by which units are selected into a sample does not depend on the values of the variables being measured. This assumption of ignorable selection, formalized by Rubin (1976), provides the theoretical justification for design-based inference from probability samples. However, the contemporary landscape of survey research presents mounting challenges to this ideal. Declining response rates across all survey modes (Brick and Williams, 2013; Williams and Brick, 2018; de Leeuw, Hox, and Luiten, 2018; Luiten, Hox, and de Leeuw, 2020; Daikeler, Bošnjak, and Lozar Manfreda, 2020; Lundmark and Backström, 2025), escalating costs of maintaining probability samples (Presser and McCulloch, 2011), and the proliferation of readily available non-probability data sources (Baker et al., 2013; Mercer et al., 2017; Cornesse et al., 2020) have created an environment where ignorability can no longer be taken for granted.

Non-probability samples, which lack a formal randomization mechanism, present particular challenges for inference. Unlike probability samples where design-based methods can in principle account for known selection probabilities, non-probability samples require model-based approaches. Elliott and Valliant (2017) outlined two broad approaches for making inferences under this setting: quasi-randomization and superpopulation modeling. Both approaches critically depend on the assumption of ignorable selection, which in practice is unlikely to hold precisely, yet existing adjustment methods provide little guidance on the magnitude of bias that may result from its violation. This gap motivates the development of sensitivity analysis tools that explicitly parameterize departures from ignorability and quantify their impact on estimates of interest.

The methodological response to this challenge has centered on developing model-based approaches that explicitly parameterize departures from ignorability. A particularly influential framework emerged from the work of Andridge and Little, 2011, who introduced the proxy pattern-mixture model (PPMM) as a principled method for sensitivity analysis in the presence of non-ignorable survey non-response. The PPMM compresses auxiliary information observed for both respondents and nonrespondents into a single proxy variable that is predictive of the outcome of interest. By modeling the joint distribution of the proxy and the outcome under different assumptions about the missing data mechanism, the PPMM provides a structured sensitivity analysis for non-response bias.

Building on the PPMM framework, Little et al. (2020) introduced the Standardized Measure of Unadjusted Bias (SMUB), a family of interpretable indices that quantify the degree of departure from ignorable selection in estimated means. Boonstra et al. (2021) later demonstrated that the SMUB correlates more strongly with true bias than traditional diagnostics. Extending this framework to binary outcomes, Andridge et al. (2019) developed the Measure of Unadjusted Bias for Proportions (MUBP), which reformulates the PPMM within a latent variable setting. More recently, West et al. (2021) generalized this framework to regression coefficients in both linear and probit models.

To compute the SMUB and MUBP, three ingredients are required: (1) microdata from a non-probability sample containing the outcome Y (continuous for SMUB or binary for MUBP) and a set of auxiliary variables Z that are predictive of Y; (2) reliable population-level summaries of Z, including means and, when available, variances and covariances, obtained from high-quality data sources such as administrative registers, large probability surveys, or other external benchmarks; and (3) an assumed sensitivity parameter ϕ that governs the degree of non-ignorability. In general terms, the estimation proceeds as follows. A proxy variable X is first constructed for the outcome Y by regressing Y on the auxiliary variables Z using the non-probability sample data, with linear regression used for SMUB and probit regression for MUBP. This proxy represents the best available predictor of the outcome based on the observed covariates, reducing a multidimensional set of auxiliaries to a single composite predictor. Population-level summaries of Z are then used to compute the corresponding population mean and variance of the proxy X. Finally, SMUB(ϕ) or MUBP(ϕ) can be obtained either by fixing specific values for ϕ (commonly 0, 0.5, and 1) or, under a Bayesian framework, by assigning ϕ a noninformative prior distribution that reflects the absence of prior knowledge about the degree of non-ignorability.

The predictive strength of the auxiliary variables plays a central role in this framework. Weakly predictive Z variables may yield highly uncertain bias estimates, making it difficult to assess the direction or magnitude of potential selection bias, reducing the diagnostic value of the indices. Correlations above approximately 0.3–0.4 between the outcome and the proxy are desirable (Andridge et al., 2019; Little et al., 2020). As discussed in Section 3, both simulations and empirical results reinforce the importance of having a strongly predictive set of covariates to ensure reliable inference.

Empirical studies have demonstrated the versatility of these indices across diverse survey domains.

West and Andridge (2023) applied the MUBP to pre-election polling data, showing improved alignment with certified election results; Andridge (2024) used the MUBP to assess bias in COVID-19 vaccine uptake estimates, finding results consistent with non-ignorable selection; Hammon and Zinn (2024) validated the MUBP using population data from the German General Social Survey; and Schroeder and West (2025) applied the MUBP to evaluate potential selection bias in the 2019 Health Survey Mailer (HSM), an off-wave supplement to the longitudinal Health and Retirement Study (HRS). Using harmonized demographic and health covariates shared across waves, they found that MUBP adjustments were small, indicating largely ignorable selection but highlighting the method's value for diagnosing bias in longitudinal survey contexts. Most recently, Gómez-Echeverry et al. (2025) applied the SMUB framework to short-term economic indicators, highlighting the benefits of incorporating historical auxiliary data to improve adjustment accuracy.

This article provides an overview of the theoretical foundations, empirical performance, and practical implementation of the SMUB and MUBP indices. We begin by outlining the PPMM framework that underlies both measures. We then synthesize evidence from simulation studies that systematically vary proxy strength, selection mechanisms, and outcome distributions, together with validation exercises and empirical applications spanning public health surveillance, demographic surveys, political polling, and administrative data used for economic indicators. Finally, we summarize practical guidance on proxy construction and sensitivity analysis, and discuss methodological extensions already available as well as ongoing research aimed at refining these indices and broadening their applicability. The overarching goal is to show that the SMUB and MUBP constitute accessible, interpretable, and empirically validated tools for diagnosing selection bias, and to motivate their broader adoption in contemporary survey research, where evaluating data quality under potential non-ignorable selection has become increasingly critical.

2 Measures of selection bias

This section summarizes the methodological framework and key formulations introduced by Little et al. (2020) and Andridge et al. (2019). The following exposition outlines the main components, assumptions, and analytical expressions underlying these measures. Readers interested in full derivations and implementation details are referred to the original papers for comprehensive discussions.

2.1 Indices of Non-Ignorable Selection Bias for Means

Little et al. (2020) developed an index-based sensitivity analysis framework that explicitly quantifies potential selection bias under varying assumptions about the degree of non-ignorability. Their approach embeds the PPMM into a tractable parametric framework that enables researchers to bound the range of plausible bias values and assess the robustness of substantive conclusions to departures from ignorable selection.

Suppose the non-probability sample provides data $\mathcal{D} = \{(y_i, z_i) : i = 1, \ldots, n\}$, where y_i denotes the continuous survey outcome for unit i and z_i is a p-dimensional vector of auxiliary variables predictive of y_i and for which summary statistics are available for the population. Let $S_i \in \{0, 1\}$ indicate selection into the non-probability sample, with $S_i = 1$ for selected units and $S_i = 0$ otherwise.

The first step constructs an auxiliary proxy for the unobserved outcome values among non-selected units. Formally, we regress Y on Z using data from selected units ($S = 1$) to obtain the fitted linear predictor $X = Z^\top \hat{\beta}$ where $\hat{\beta}$ denotes the least-squares coefficient vector. This proxy X represents the best linear predictor of Y based on the available auxiliaries and serves as a surrogate for Y in the non-selected population. To facilitate model specification and interpretation, X is rescaled to match the variance of Y within the non-probability sample $X^* = X\sqrt{\frac{s_{YY}^{(1)}}{s_{XX}^{(1)}}}$ where $s_{YY}^{(1)}$ and $s_{XX}^{(1)}$ denote sample variances among selected units. For notational convenience, we denote the rescaled proxy from now on as X, with the rescaling implicit.

The PPMM assumes that the joint distribution of (Y, X) follows a bivariate normal distribution conditional on selection status S:

$$(Y, X) \mid S = j \sim \mathcal{N}_2\left(\begin{pmatrix}\mu_Y^{(j)} \\ \mu_X^{(j)}\end{pmatrix}, \begin{pmatrix}\sigma_{YY}^{(j)} & \sigma_{XY}^{(j)} \\ \sigma_{XY}^{(j)} & \sigma_{XX}^{(j)}\end{pmatrix}\right), \quad j \in \{0, 1\}. \tag{1}$$

Some parameters governing the distribution of (Y, X) among nonselected units ($j = 0$) are not identified from the observed data $(\mu_Y^{(0)}, \sigma_{YY}^{(0)}, \sigma_{XY}^{(0)})$. Identification is achieved by assuming that the probability of selection depends on (X, Y) through a scalar index formed as a convex combination of the two variables:

$$\Pr(S = 1 \mid X, Y) = g\big((1-\phi)X + \phi Y\big),$$

where $g : \mathbb{R} \to (0, 1)$ is an unspecified monotonic function and $\phi \in [0, 1]$ is a scalar sensitivity parameter.

The parameter ϕ quantifies the *degree of non-ignorability* and admits an intuitive interpretation. When $\phi = 0$, selection depends only on the observed proxy X, corresponding to *selection at random* (SAR) conditional on Z, which represents ignorable selection. When $\phi = 1$, selection depends entirely on the outcome Y, representing fully non-ignorable selection where the auxiliary variables provide no direct information about the selection mechanism. For intermediate values $0 < \phi < 1$, selection depends on both X and Y, with larger values indicating stronger dependence on the unobserved outcome.

The maximum likelihood estimator for the population mean of Y as a function of ϕ is given by:

$$\hat{\mu}_Y(\phi) = \bar{y}^{(1)} + \frac{\phi + (1-\phi)\hat{\rho}_{XY}^{(1)}}{\phi\hat{\rho}_{XY}^{(1)} + (1-\phi)}\sqrt{\frac{s_{YY}^{(1)}}{s_{XX}^{(1)}}}\,(\bar{x}^{(1)} - \bar{X}),$$

where $\bar{y}^{(1)}$ and $\bar{x}^{(1)}$ denote sample means among selected units, $\bar{X}$ is the known population mean of X (computed from population-level summaries of Z and the estimated coefficients $\hat{\beta}$), and $\hat{\rho}_{XY}^{(1)}$ is the sample Pearson correlation between Y and X among selected units.

The *Measure of Unadjusted Bias* (MUB) is defined as the difference between the naive sample mean and this model-based adjustment:

$$\text{MUB}(\phi) = \bar{y}^{(1)} - \hat{\mu}_Y(\phi).$$

Because MUB depends on the measurement scale of Y, hindering comparisons across outcomes, Little et al. (2020) recommend standardizing by the sample standard deviation of Y, obtaining the Standardized Measure of Unadjusted Bias (SMUB):

$$\text{SMUB}(\phi) = \frac{\text{MUB}(\phi)}{\sqrt{s_{YY}^{(1)}}} = \frac{\phi + (1-\phi)\hat{\rho}_{XY}^{(1)}}{\phi\hat{\rho}_{XY}^{(1)} + (1-\phi)} \cdot \frac{\bar{x}^{(1)} - \bar{X}}{\sqrt{s_{XX}^{(1)}}}.$$

Critically, ϕ cannot be estimated from the observed data, as there is no information about the distribution of Y among non-selected units. The strategy adopted is therefore to conduct a *sensitivity* *analysis*, computing bias estimates across a range of plausible ϕ values to assess the robustness of conclusions to departures from ignorability. Three particular values of ϕ provide intuitive benchmarks:

$$\text{SMUB}(0) = \hat{\rho}_{XY}^{(1)} \frac{\bar{x}^{(1)} - \bar{X}}{\sqrt{s_{XX}^{(1)}}}, \qquad \text{SMUB}(0.5) = \frac{\bar{x}^{(1)} - \bar{X}}{\sqrt{s_{XX}^{(1)}}}, \qquad \text{SMUB}(1) = \frac{1}{\hat{\rho}_{XY}^{(1)}} \frac{\bar{x}^{(1)} - \bar{X}}{\sqrt{s_{XX}^{(1)}}}.$$

To reflect sensitivity to the choice of ϕ, Little et al. (2020) recommend reporting the *sensitivity interval* $[\text{SMUB}(0), \text{SMUB}(1)]$ to bound the range of plausible bias values under the PPMM assumptions, with SMUB(0.5) serving as a central point estimate when no prior information about ϕ is available. If this interval excludes zero and is substantively meaningful in magnitude, it provides evidence that selection bias may threaten the validity of conclusions drawn from the non-probability sample.

To isolate the component of bias attributable specifically to departures from ignorability (i.e., $\phi > 0$), Little et al. (2020) define the *Standardized Measure of Adjusted Bias* (SMAB) as:

$$\text{SMAB}(\phi) = \text{SMUB}(\phi) - \text{SMUB}(0) = \frac{\phi\{1 - (\hat{\rho}_{XY}^{(1)})^2\}}{\phi\hat{\rho}_{XY}^{(1)} + (1-\phi)} \cdot \frac{\bar{x}^{(1)} - \bar{X}}{\sqrt{s_{XX}^{(1)}}}.$$

While SMUB quantifies the total bias in the unadjusted sample mean $\bar{y}^{(1)}$, SMAB captures the portion of the overall bias in an unadjusted estimate that exists after adjustment for the known auxiliary variables (given a choice of ϕ), under an assumption that selection is only a function of X (or ignorable).

We note that SMUB(0), SMUB(0.5) and SMUB(1) can be computed without access to microdata for population elements excluded from the non-probability sample. A key advantage of these indices is that they require only the aggregate population mean of the proxy X, which in turn depends on the population means of the auxiliary variables Z. However, these point estimates of bias do not account for sampling variability in constructing the proxy X, that is, in estimating $\hat{\beta}$ from the regression of Y on Z, and may therefore understate total uncertainty. To address this limitation, Little et al. (2020) proposed a fully Bayesian approach that yields posterior draws of SMUB(ϕ), allowing uncertainty to be fully propagated and producing point estimates and credible intervals that can assess whether the estimated bias differs meaningfully from zero or exceeds a substantively important threshold.

Specifically, under a fully Bayesian approach, prior distributions are placed on the regression coefficients β defining the proxy X, the pattern-specific parameters in Equation (1), and the sensitivity parameter ϕ, which can either be fixed or assigned a prior distribution. A common default specification assigns relatively noninformative priors to β and the pattern–mixture parameters and a $\mathrm{Uniform}(0,1)$ prior to ϕ, reflecting complete ignorance about the degree of non-ignorability. Markov chain Monte Carlo methods then yield posterior draws of $\mathrm{SMUB}(\phi)$ that fully propagate uncertainty, producing credible intervals that can be used to assess whether estimated bias is meaningfully different from zero or exceeds a substantively important threshold. This approach requires the sample mean and variance of X for the non-sampled population, which depend on the sample mean and covariance matrix of Z among non-sampled units. When only the means of Z are available, as is often the case, it can be assumed that the population covariance matrix of Z is the same for sampled and non-sampled units, allowing it to be estimated from the sampled cases.

2.2 Indices of Non-Ignorable Selection Bias for Proportions

The SMUB framework presented in Section 2.1 assumes normally distributed outcomes, limiting its direct applicability to binary variables. To address this limitation, Andridge et al. (2019) extended the proxy pattern–mixture model to binary outcomes by introducing a latent variable formulation, building on earlier developments by Andridge and Little (2020). This extension preserves the intuitive interpretation of the sensitivity parameter ϕ while accommodating the discrete nature of proportions, yielding the Measure of Unadjusted Bias for Proportions (MUBP).

Let Y be a binary variable taking values 0 or 1, representing, for instance, the presence or absence of a particular characteristic in the target population. Following standard probit model conventions, Y is assumed to arise from an underlying continuous latent variable U via the threshold mechanism

$$Y = \begin{cases} 1 & \text{if } U > 0, \\ 0 & \text{if } U \leq 0. \end{cases}$$

The latent variable formulation facilitates the specification of a tractable joint model for Y and auxiliary predictors, enabling the application of normal pattern-mixture modeling techniques analogous to those used for continuous outcomes.

As in the continuous case, let $S \in \{0, 1\}$ denote selection into the non-probability sample, with Y observed only when $S = 1$. The proxy X is constructed by regressing the binary outcome Y on the auxiliaries Z using a probit model fitted to the non-probability sample. In this case, Z must be available for all units in the non-probability sample, and either sufficient statistics (means, variances and covariances) or microdata for Z must be available for the non-selected units. A probit regression model is used for the binary indicator of interest because this model assumes that the observed indicator arises from an underlying, unobserved latent variable that follows a normal distribution.

Following the same pattern-mixture framework used for continuous outcomes, the joint distribution of the latent variable U and proxy X is assumed to follow a bivariate normal distribution conditional on selection status:

$$(U, X \mid S = j) \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_U^{(j)} \\ \mu_X^{(j)} \end{pmatrix}, \begin{pmatrix} \sigma_{uu}^{(j)} & \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} \\ \rho_{ux}^{(j)} \sqrt{\sigma_{uu}^{(j)} \sigma_{xx}^{(j)}} & \sigma_{xx}^{(j)} \end{pmatrix} \right), \quad j \in \{0, 1\}.$$

Here $\mu_U^{(j)}$ and $\mu_X^{(j)}$ denote the means of the latent variable and proxy in selection pattern j, $\sigma_{uu}^{(j)}$ and $\sigma_{xx}^{(j)}$ are their variances, and $\rho_{ux}^{(j)}$ is their correlation. As in the continuous-outcome case, some parameters governing the distribution of (U, X) among non-selected units ($j = 0$) are not identified without additional assumptions. To achieve identification, the same structural assumption is used in SMUB, namely that selection depends on (U, X) through a scalar index, $\Pr(S = 1 \mid U, X) = g\big((1-\phi)X^* + \phi U\big)$, where $X^* = X\sqrt{\frac{\sigma_{uu}^{(1)}}{\sigma_{xx}^{(1)}}}$, $g(\cdot)$ is an unspecified monotonic function and $\phi \in [0, 1]$ is the sensitivity parameter.

We note that the effectiveness of the auxiliary proxy X in predicting the binary outcome Y is quantified by the *biserial correlation*, which measures the association between a continuous variable (the proxy X) and a binary variable (Y). In the latent variable framework, this is equivalent to the Pearson correlation between U and X among selected units, $\rho_{ux}^{(1)} = \text{Corr}(U, X \mid S = 1) = \text{Biserial Corr}(Y, X \mid S = 1)$.

As is customary with latent variables, $\sigma_{uu}^{(1)} = 1$, since the mean and variance cannot be separately estimated. Under this model specification, the marginal probability that $Y = 1$ in the target population can be expressed by:

$$\mu_Y = \Pr(Y=1) = \pi\Phi(\mu_U^{(1)}) + (1-\pi)\Phi\left(\frac{\mu_U^{(0)}(\phi)}{\sigma_{uu}^{(0)}(\phi)}\right),$$

where $\pi = \Pr(S=1)$ is the proportion of selected cases in the population, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and $(\mu_U^{(0)}(\phi), \sigma_{uu}^{(0)}(\phi))$ are the mean and variance of U among non-selected units, which depend on the assumed value of ϕ. These unidentified parameters for a specific choice of ϕ are given by:

$$\mu_U^{(0)}(\phi) = \mu_U^{(1)} + \frac{\phi + (1-\phi)\rho_{ux}^{(1)}}{\phi\rho_{ux}^{(1)} + (1-\phi)} \cdot \frac{\mu_X^{(0)} - \mu_X^{(1)}}{\sigma_{xx}^{(1)}}, \tag{2}$$

$$\sigma_{uu}^{(0)}(\phi) = 1 + \left[\frac{\phi + (1-\phi)\rho_{ux}^{(1)}}{\phi\rho_{ux}^{(1)} + (1-\phi)}\right]^2 \cdot \frac{\sigma_{xx}^{(0)} - \sigma_{xx}^{(1)}}{\sigma_{xx}^{(1)}}. \tag{3}$$

The difference of the proportion for selected cases from the overall proportion is therefore

$$\mu_y^{(1)} - \mu_y = \mu_y^{(1)} - \left\{\pi\Phi(\mu_u^{(1)}) + (1-\pi)\Phi\left(\frac{\mu_u^{(0)}}{\sqrt{\sigma_{uu}^{(0)}}}\right)\right\}.$$

For a given choice of ϕ, a Measure of Unadjusted Bias for Proportions, MUBP(ϕ), is then defined as the difference between the proportion observed in the non-probability sample and the estimated population proportion:

$$\begin{aligned}
\mathrm{MUBP}(\phi) &= \hat{\mu}_y^{(1)} - \hat{\mu}_y \\
&= \hat{\mu}_y^{(1)} - \hat{\pi}\Phi(\hat{\mu}_u^{(1)}) - (1-\hat{\pi}) \\
&\quad \times \Phi\left(\left\{\hat{\mu}_u^{(1)} + \frac{\phi + (1-\phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1-\phi)} \frac{\hat{\mu}_x^{(0)} - \hat{\mu}_x^{(1)}}{\sqrt{\hat{\sigma}_{xx}^{(1)}}}\right\} \Big/ \right. \\
&\qquad \left. \sqrt{1 + \left\{\frac{\phi + (1-\phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1-\phi)}\right\}^2 \frac{\hat{\sigma}_{xx}^{(0)} - \hat{\sigma}_{xx}^{(1)}}{\hat{\sigma}_{xx}^{(1)}}}\right).
\end{aligned}$$

where $\hat{\mu}_Y^{(1)} = \bar{y}^{(1)}$ is the sample proportion among selected units, and $\hat{\mu}_Y(\phi)$ is computed by replacing the parameters by estimates into Equations (2) and (3). Estimation of MUBP(ϕ) requires computing the sampling fraction π, which may be close to 0 for larger populations, the biserial correlation $\rho_{ux}^{(1)}$ between the latent variable U and the proxy X among selected units, and sufficient statistics for the proxy variable X for both the selected and the non-selected portions of the target population. This last requirement is stronger than that for the SMUB, which only requires the population mean of X, not its variance. Maximum likelihood (ML) estimates of these sufficient statistics for the selected cases can be computed using the observed data from the non-probability sample. Andridge et al. (2019) estimate $\rho_{ux}^{(1)}$ using the two-step approach of (Olsson, Drasgow, and Dorans, 1982), while the remaining parameters are obtained via ML. They refer to the resulting estimates as ‘modified’ ML (MML). To prevent overfitting in the construction of the proxy X and in the estimation of $\rho_{ux}^{(1)}$, they recommend multifold cross-validation.

As with SMUB, extreme and intermediate values of ϕ provide interpretable benchmarks for sensitivity analysis. When $\phi = 0$ (ignorable selection), selection depends only on the observed proxy X. When $\phi = 1$ (fully non-ignorable selection), selection depends entirely on the latent outcome U, and MUBP(1) provides an upper bound on potential bias under the model assumptions. The midpoint MUBP(0.5) represents a compromise assumption of equal dependence on the proxy and latent outcome. Andridge et al. (2019) recommend reporting the sensitivity interval $[\text{MUBP}(0), \text{MUBP}(1)]$ to bound the plausible range of bias values.

Finally, as with the SMUB, the maximum likelihood estimation of the MUBP treats the coefficients in the probit model and therefore the proxy X as fixed, potentially understating total uncertainty in the bias estimates. To address this limitation, Andridge et al. (2019) proposed a fully Bayesian implementation that propagates uncertainty through all levels of estimation. Under this approach, prior distributions are placed on the regression coefficients defining X, on the parameters of the pattern–mixture model, and on the sensitivity parameter ϕ, which can either be fixed or assigned a prior distribution such as $\mathrm{Uniform}(0, 1)$. The Gibbs sampler alternates between imputing the latent variable U from a truncated normal distribution, updating the regression coefficients for the probit model, regenerating the proxy X, and drawing the parameters of the pattern–mixture model. These steps yield posterior draws of $\mathrm{MUBP}(\phi)$ that fully incorporate parameter and model uncertainty.

3 Evidence from Simulations and Empirical Applications of SMUB and MUBP

The development of the SMUB and MUBP has been followed by a series of simulation studies and empirical applications designed to evaluate how well these indices perform in realistic survey conditions. These studies examined their ability to detect and quantify non-ignorable selection bias for continuous and binary outcomes, proxy strengths, and selection mechanisms, providing a clear picture of their practical strengths and limitations.

Andridge et al. (2019) conducted a simulation study comparing the MUB with the MUBP. The simulation design generated a binary outcome from a latent variable framework, allowing direct comparison of both approaches while varying the correlation between the proxy and the latent outcome and the degree of non-ignorability. Results showed that the MUBP more accurately captured bias in proportions, avoiding implausible estimates outside the [0,1] range that could arise under the linear-normal MUB formulation. Its performance was strong when the proxy was at least moderately predictive, producing well-calibrated sensitivity intervals. The ML-based intervals tend to be wider and to have higher coverage for the normal model than the MML-based intervals for the probit model. Coverage of the Bayesian intervals is higher than that of the MML-based intervals for both models.

The subsequent simulation work of Boonstra et al. (2021) offered a systematic evaluation of SMUB and related diagnostics in settings with continuous outcomes. The authors simulated finite populations where the relationship between outcome, auxiliary variables, and selection could be controlled, manipulating parameters such as the correlation between the outcome and its proxy, the overlap between outcome and selection predictors, and the strength of non-ignorability. Across these conditions, SMUB showed the strongest and most consistent correlation with the realized bias in estimated means, outperforming traditional diagnostics. The SMAB index effectively captured the portion of bias due to non-ignorability, remaining accurate when model assumptions were satisfied. However, as noted by Boonstra et al. (2021), performance declined when the auxiliary variable was only weakly correlated with the outcome, confirming that the usefulness of outcome-based diagnostics depends critically on having a sufficiently informative proxy.

Empirical applications further validated these insights. In the studies introducing these indices, Little et al. (2020) and Andridge et al. (2019) applied them to data from the National Survey of Family Growth (NSFG), treating smartphone owners as a non-probability sample. This design allowed for direct comparison between sample-based estimates and population benchmarks. In Little et al. (2020), the authors demonstrated that SMUB effectively identified survey variables most vulnerable to selection bias, performing well when the proxy–outcome correlation exceeded roughly 0.4. When this relationship was weak, they noted that any diagnostic based solely on auxiliary variables would likely be uninformative. Building on this framework, Andridge et al. (2019) applied the MUBP, showing that it produced narrower and more interpretable sensitivity intervals than its continuous counterpart (MUB) for proportions. The MUBP accurately captured the true bias for most binary outcomes when the proxy was at least moderately predictive of the latent outcome (Pearson correlation above about 0.3) and achieved improved coverage when uncertainty in the probit coefficients was incorporated through Bayesian credible intervals.

Subsequent research has demonstrated the practical value of these indices in diverse real-world settings. West and Andridge (2023) applied the MUBP to evaluate bias in pre-election polling for the 2020 U.S. presidential election. The main case study drew on the ABC/Washington Post polls conducted by Abt Associates in September and October 2020, focusing on likely voters in key states including Wisconsin, Michigan, and Pennsylvania. The estimand of interest was the proportion intending to vote for Donald Trump. Concerns about non-ignorable selection arose from the possibility that Trump supporters were systematically less likely to participate in pre-election polls. Population benchmarks were drawn from three major sources: the November 2020 CPS Voter Supplement, the 2020 ANES pre-election survey, and the AP/NORC VoteCast 2020 data. Each source offered advantages and limitations: CPS lacked direct measures of ideology and party identification, ANES had relatively small state samples, and VoteCast was not entirely probability-based. Covariates harmonized across sources included sex, age, education, race/ethnicity, political ideology, and party identification. Results showed that MUBP-adjusted estimates of Trump support were consistently higher than those produced by standard weighting alone. In many cases, the adjusted estimates narrowed the gap between poll results and the certified election outcomes. At the same time, the authors emphasized the practical challenges of implementing the MUBP, particularly the difficulties of aligning covariates across benchmark datasets.

Applications have also extended beyond political polling. Andridge (2024) investigated estimates of COVID-19 vaccine uptake from the Census Household Pulse Survey (HPS) and the Delphi-Facebook COVID-19 Trends and Impact Survey. Both surveys overestimated uptake relative to CDC benchmarks—by 14 and 17 percentage points, respectively—despite their very large sample sizes. The HPS was treated as a non-probability survey due to its extremely low response rate (6–7%). Auxiliary covariates included sex, age, education, race/ethnicity, and state, harmonized with the American Community Survey. MUBP analysis indicated that the observed overestimation was consistent with non-ignorable selection, especially if unvaccinated individuals were less likely to respond.

Validation studies have reinforced the empirical patterns and limitations observed in earlier applications. Hammon and Zinn (2024) conducted a validation study using the German General Social Survey (GGSS) as a finite population. Ten binary outcomes, including unemployment, union membership, and religious affiliation, were analyzed by comparing the full GGSS population to an artificial non-probability sample defined by internet use and political interest. They concluded that the MUBP performs well in detecting selection bias in estimated proportions when the assumptions of the underlying PPMM are satisfied. They emphasized that a moderate difference in the proxy distributions between sampled and non-sampled cases is crucial for correctly indicating the true bias. In their analysis, this condition was even more relevant than a very high correlation between X and Y, although a strong correlation is an important condition to avoid ineffective and very wide intervals of potential selection bias. The same study applied the MUBP to a large river-sampled online survey in Germany, where the authors demonstrated the practical utility of the MUBP for assessing the robustness of estimated proportions under different assumptions about the selection mechanism.

An evaluation of potential non-ignorable selection bias was conducted using the 2019 Health Survey Mailer (HSM), an off-wave supplement to the Health and Retirement Study (HRS) with an 83% response rate. Despite this high participation, eligibility restrictions raised concerns about systematic exclusion. Using demographic and health covariates common to the HSM and the HRS core, Schroeder and West (2025) estimated MUBP-adjusted proportions for ten binary health outcomes. Weighted and MUBP-adjusted estimates were generally consistent, with overlapping confidence and credible intervals for most outcomes. Larger MUBP shifts were observed only when auxiliary proxies were strong (biserial correlations above 0.5), while weaker proxies yielded wider credible intervals and smaller adjustments. Benchmark analyses treating common covariates as outcomes confirmed that both methods moved estimates toward the population truth. The study also compared results using the National Health Interview Survey (NHIS) as an alternative population source, finding lower biserial correlations and wider credible intervals when fewer common covariates were available. Beyond these empirical findings, the authors emphasized broader implications for survey researchers: MUBP can be especially valuable in panel studies that include informative covariates shared across waves, where population-level information is easier to obtain and proxy correlations tend to be higher. They also highlighted the importance of identifying strong auxiliary predictors to ensure efficiency and interpretability of bias adjustments. Overall, the results suggested that selection bias in the HSM was likely ignorable given the available covariates, and that standard weighting sufficed, while the MUBP provided reassurance and diagnostic insight into potential non-ignorable selection bias.

While the HSM study examined a traditional survey application, subsequent research has adapted these indices for use in administrative and short-term estimation contexts. Gómez-Echeverry et al. (2025) applied the MUB to flash estimates constructed from gradually filling non-probability samples, such as administrative data used for short-term economic indicators. The authors proposed three practical implementations that differ in how the sensitivity parameter ϕ is handled: MUB(C.5) fixes $\phi = 0.5$, MUB(C) uses plausible values of ϕ to approximate the range of potential bias values that are consistent with the observed data, and MUB(M) estimates ϕ using lagged and current information. Simulation results showed that MUB(M) achieved the best overall performance, particularly when selection bias was substantial, demonstrating that anchoring ϕ to historical data improves adjustment accuracy. The study also found that the correlation between the target variable and the selection mechanism was more influential than the specific distributional shape of the target variable in determining bias. A case study using turnover data from Statistics Netherlands confirmed these findings, with MUB(M) producing the lowest estimation errors across several economic sectors.

Across simulation studies and empirical applications, several consistent patterns emerge regarding the performance and practical utility of the SMUB and MUBP as diagnostics for non-ignorable selection bias. When auxiliary variables are at least moderately predictive of the outcome, typically with correlations above 0.3 to 0.4, these outcome-aware indices reliably capture both the direction and magnitude of bias, outperforming traditional representativeness diagnostics that ignore outcome distributions. Their performance declines predictably when proxies are weak, signaling insufficient auxiliary information rather than masking uncertainty. The treatment of the sensitivity parameter ϕ plays a key role: fixed values such as $\phi = 0.5$ offer simple summaries but can be less accurate than analyses spanning the full $\phi \in [0, 1]$ range, while approaches that estimate ϕ from historical or lagged data yield the most precise adjustments. Bayesian formulations, which propagate uncertainty from both proxy construction and model estimation, tend to produce better-calibrated intervals than maximum likelihood estimation. Applications across health, political, and administrative domains confirm that the SMUB and MUBP can uncover non-ignorable selection risks overlooked by conventional diagnostics, provided that proxies are strong and covariates are harmonized across data sources.

4 Discussion

The goal of this paper was to review and synthesize recent developments in diagnostic measures of selection bias for non-probability samples, focusing on the SMUB and the MUBP. Both indices provide accessible and interpretable tools for quantifying the sensitivity of survey estimates to non-ignorable selection. Their main strengths lie in their parsimony and practical feasibility. The use of a single sensitivity parameter ϕ captures the continuum between ignorable and non-ignorable selection, while the construction of a proxy variable summarizes the influence of multiple auxiliary covariates into a single dimension, simplifying implementation. Moreover, these indices can be computed even in the absence of population microdata, provided that sufficient population-level summary statistics for the auxiliary variables are available. Compared to traditional approaches, the SMUB and MUBP have been shown to detect non-ignorable bias in both simulations and empirical applications more effectively.

Building on this foundation, selection of the auxiliary variables of Z should be guided by both predictive power for the outcome of interest and availability of reliable population benchmarks. In the PPMM framework, Z is used to construct the proxy X and to supply population summaries that anchor identification, so variables that are strongly related to Y and measured consistently across surveys are preferred. These covariates must be predictive of Y but also collected in comparable form across data sources to ensure valid application of the SMUB and MUBP. Weak correlations inflate the sensitivity of results to ϕ and may produce wide, uninformative sensitivity intervals. Researchers should report the estimated correlation to communicate the strength of the proxy.

Reliable population statistics are usually obtained from large probability surveys. However, due to the increasing challenges faced by these methods, some government agencies are turning to administrative data sources to produce official statistics (Berzofsky et al., 2025), making them a promising source of auxiliary information to implement PPMM-based indices. It is worth noting that administrative data also face issues related to quality and coverage, and because of their nonprobabilistic nature, a non-ignorable missing data mechanism can cause systematic biases that standard adjustment methods may not fully correct. Nonetheless, the PPMM-based measures discussed here could also be applied to evaluate bias in administrative datasets and support their use in the production of official statistics.

In some applications, the auxiliary variables Z needed to construct the proxy may not be directly available in the non-probability sample. In such cases, these variables can be obtained by linking the sample to external sources. Little et al. (2020) suggest that, when suitable auxiliary variables are unavailable, data fusion techniques can be used to integrate variables with the required properties from another independent dataset. Linking to administrative register data can be particularly advantageous, as these sources often provide rich and reliable information. However, when such linking procedures are employed, uncertainty arising from potential mismatch errors should be properly accounted for. Recent work by Slawski et al. (2025) has developed a general framework for valid post-linkage inference in the presence of mismatch error. Incorporating these ideas into the estimation of the indices discussed in this paper represents a promising direction for future research.

Once auxiliary variables are selected and population benchmarks are identified, estimation can proceed directly. Closed-form expressions for SMUB and SMAB permit straightforward maximum likelihood (ML) estimation using sample statistics and external population summaries. Little et al. (2020) and Andridge et al. (2019) provide accompanying `R` functions `github.com/bradytwest/IndicesOfNISB`) that implement both ML and Bayesian estimation for these indices. A preliminary `R` package is also available at `github.com/randridge/ppmm`. Together, these open-source tools facilitate replication of published results, illustrate practical implementation of the indices, and, importantly, are designed to lower the barrier to their application across a wide range of research domains.

The next step involves assessing how sensitive the conclusions are to different assumptions about the selection process. Because the true selection mechanism is rarely known, sensitivity analysis provides a transparent way to evaluate robustness. A practical approach is to report sensitivity intervals, which bound the plausible range of bias under varying assumptions. The midpoint $\phi = 0.5$ offers a convenient single-number summary corresponding to equal dependence on X and Y, and has simple closed-form expressions in the continuous case. A Bayesian formulation further extends this approach by propagating uncertainty across all model components.

The PPMM framework relies on the assumption that (Y, X) follows a bivariate normal distribution. Gómez-Echeverry et al. (2025) reported that deviations of Y from normality exert a weaker influence on the performance of the MUB than changes in the strength of the non-ignorable selection mechanism or the predictive power of the proxy. Although the distributional shape plays a secondary role, it can marginally affect estimator accuracy when the selection mechanism is strongly non-ignorable or when the auxiliary variables are only moderately informative. Scenarios that combine current auxiliary variables with lagged information on the target variable appear to offer some protection against departures from normality. Overall, results obtained under normality tend to perform better than those with larger deviations. Further research examining the robustness of the SMUB and MUBP to their distributional assumptions is needed; the gamma-based extension proposed by Andridge and Thompson (2015) could provide additional insights.

Recent work has extended this framework beyond its original focus on means and proportions. West et al. (2021) extended the approach to regression coefficients in both linear and probit models, while ongoing research by Andridge and colleagues is adapting the method to ordinal and nominal outcomes through versions of the MUBP based on ordinal and multinomial probit models. Complementary methodological developments have re-expressed the PPMM as a selection model (Yiadom and Andridge, 2024) and extended the framework to subgroup estimation. Together, these efforts reinforce the conceptual foundation of the indices and expand their applicability across a broader range of survey estimation problems.

Ongoing research aims to refine the SMUB and MUBP by accounting for sampling uncertainty from finite probability survey benchmarks. This refinement is especially relevant when benchmarks are drawn from moderately sized reference samples rather than large-scale data sources, where sampling error can materially affect the accuracy of the indices. Work in progress is also focused on improving proxy construction using machine learning methods such as Bayesian additive regression trees (BART), which can capture complex non-linear relationships between auxiliary variables and outcomes, potentially yielding stronger proxies and tighter sensitivity intervals.

The original articulation of proxy pattern–mixture models emphasized that auxiliary covariates are indispensable for evaluating bias, a principle that equally applies to traditional methods for inference from non-probability samples. Valid estimation ultimately depends on the availability of high-quality auxiliary information. While the SMUB and MUBP can be computed using summary-level rather than microdata, their effectiveness still hinges on the accuracy and relevance of the covariates used.

In general, best practices for drawing valid inferences from non-probability samples (or from probability samples with low response rates) call for the identification of a large reference probability survey targeting the same population. Such a reference data source supplements the non-probability sample by providing auxiliary information on population characteristics that are essential for bias adjustment. Both data sources must include a set of common, harmonized covariates measured in the same way for individuals from the same target population. These shared covariates should be strong predictors of the key variables observed only in the non-probability sample. In practice, this requires identifying a large, representative probability survey such as the ACS, CPS, NHIS, or ANES that includes comparable measures, allocating sufficient time for harmonization when variables differ across surveys, and verifying after data collection that the chosen covariates are indeed predictive of the outcomes of interest. These steps are by no means trivial but remain indispensable, as methods for non-probability inference, whether based on microdata modeling or on population-level sufficient statistics, all rely on the consistency and predictive strength of harmonized covariates.

In this regard, we strongly support the argument made by Elliot (2022), who noted that the growing reliance on non-probability samples creates an urgent need for well-supported probability surveys to provide reliable benchmark information. Sustained investment in government-funded probability surveys is critical not only to preserve their role as independent data sources but also to strengthen their capacity to serve as analytical partners for non-probability survey inference, ensuring coverage of key covariates across the many domains where inference from non-probability samples is needed.

Acknowledgments

This work was supported in part by the National Cancer Institute of the National Institutes of Health through grant R03CA280007 (PI: Andridge).

References

Andridge, R. R. (Aug. 2024). “Using proxy pattern-mixture models to explain bias in estimates of COVID-19 vaccine uptake from two large surveys”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 187.3, pp. 831–843. ISSN: 0964-1998. DOI: 10.1093/jrsssa/qnae005. URL: https://doi.org/10.1093/jrsssa/qnae005 (visited on 09/25/2025).

Andridge, R. R. and R. J. A. Little (2011). “Proxy Pattern-Mixture Analysis for Survey Nonresponse”. In: *Journal of Official Statistics* 27, pp. 153–180.

Andridge, R. R. and R. J. A. Little (July 2020). “Proxy Pattern-Mixture Analysis for a Binary Variable Subject to Nonresponse”. en. In: *Journal of Official Statistics* 36.3, pp. 703–728. DOI: 10.2478/jos-2020-0035. URL: https://sciendo.com/article/10.2478/jos-2020-0035 (visited on 09/29/2025).

Andridge, R. R. and K. J. Thompson (Dec. 2015). “Assessing nonresponse bias in a business survey: Proxy pattern-mixture analysis for skewed data”. In: *The Annals of Applied Statistics* 9.4, pp. 2237–2265. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/15-AOAS878. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-9/issue-4/Assessing-nonresponse-bias-in-a-business-survey--Proxy-pattern/10.1214/15-AOAS878.full (visited on 09/25/2025).

Andridge, R. R., B. T. West, R. J. A. Little, P. S. Boonstra, and F. Alvarado-Leiton (Nov. 2019). “Indices of Non-Ignorable Selection Bias for Proportions Estimated from Non-Probability Samples”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 68.5, pp. 1465–1483. ISSN: 0035-9254. DOI: 10.1111/rssc.12371. URL: https://doi.org/10.1111/rssc.12371 (visited on 09/25/2025).

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau (Nov. 2013). “Summary Report of the AAPOR Task Force on Non-probability Sampling”. In: *Journal of Survey Statistics and Methodology* 1.2, pp. 90–143. ISSN: 2325-0984. DOI: 10.1093/jssam/smt008. URL: https://doi.org/10.1093/jssam/smt008 (visited on 10/21/2025).

Berzofsky, M. E., D. Liao, C. Barnett-Ryan, and E. L. Smith (June 2025). “A total data quality paradigm for official statistics based on administrative data”. EN. In: *Statistical Journal of the IAOS* 41.2, pp. 205–215. ISSN: 1874-7655. DOI: 10.1177/18747655251337671. URL: https://doi.org/10.1177/18747655251337671 (visited on 10/30/2025).

Boonstra, P. S., R. J. Little, B. T. West, R. R. Andridge, and F. Alvarado-Leiton (Sept. 2021). “A Simulation Study of Diagnostics for Selection Bias”. EN. In: *Journal of Official Statistics* 37.3, pp. 751–769. ISSN: 0282-423X. DOI: 10.2478/jos-2021-0033. URL: https://journals.sagepub.com/action/showAbstract (visited on 09/25/2025).

Brick, J. M. and D. Williams (Jan. 2013). “Explaining Rising Nonresponse Rates in Cross-Sectional Surveys”. EN. In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 36–59. ISSN: 0002-7162. DOI: 10.1177/0002716212456834. URL: https://doi.org/10.1177/0002716212456834 (visited on 10/13/2025).

Cornesse, C., A. G. Blom, D. Dutwin, J. A. Krosnick, E. D. De Leeuw, S. Legleye, J. Pasek, D. Pennay, B. Phillips, J. W. Sakshaug, B. Struminskaya, and A. Wenz (Feb. 2020). “A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research”. In: *Journal of Survey Statistics and Methodology* 8.1, pp. 4–36. ISSN: 2325-0984. DOI: 10.1093/jssam/smz041. URL: https://doi.org/10.1093/jssam/smz041 (visited on 10/21/2025).

Daikeler, J., M. Bošnjak, and K. Lozar Manfreda (June 2020). “Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates”. In: *Journal of Survey Statistics and Methodology* 8.3, pp. 513–539. ISSN: 2325-0984. DOI: 10.1093/jssam/smz008. URL: https://doi.org/10.1093/jssam/smz008 (visited on 10/21/2025).

De Leeuw, E. d., J. Hox, and A. Luiten (Dec. 2018). “International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data”. en-US. In: *Survey Methods: Insights from the Field (SMIF)*. ISSN: 2296-4754. DOI: 10.13094/SMIF-2018-00008. URL: https://surveyinsights.org/?p=10452 (visited on 10/21/2025).

Elliot, M. R. (2022). “Comments on “Statistical inference with non-probability survey samples””. en. In: *Survey Methodology, Statistics Canada* 48.2. URL: https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00004-eng.htm (visited on 09/30/2025).

Elliott, M. R. and R. Valliant (May 2017). “Inference for Nonprobability Samples”. In: *Statistical Science* 32.2, pp. 249–264. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/16-STS598. URL: https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Inference-for-Nonprobability-Samples/10.1214/16-STS598.full (visited on 10/13/2025).

Gómez-Echeverry, S., A. Van Delden, T. De Waal, D. Pavlopoulos, and R. Stoel (Sept. 2025). “Quantifying Selection Bias in Flash Estimates Derived from Cumulative Non-Probability Samples”. In: *Journal of Survey Statistics and Methodology* 00, pp. 1–26. ISSN: 2325-0992. DOI: 10.1093/jssam/smaf016. URL: https://doi.org/10.1093/jssam/smaf016 (visited on 09/25/2025).

Hammon, A. and S. Zinn (2024). “Validating an Index of Selection Bias for Proportions in Non-Probability Samples”. en. In: *International Statistical Review*. ISSN: 1751-5823. DOI: 10.1111/

insr.12590. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12590 (visited on 09/25/2025).

Little, R. J. A., B. T. West, P. S. Boonstra, and J. Hu (Nov. 2020). "Measures of the Degree of Departure from Ignorable Sample Selection". In: *Journal of Survey Statistics and Methodology* 8.5, pp. 932–964. ISSN: 2325-0984. DOI: 10.1093/jssam/smz023. URL: https://doi.org/10.1093/jssam/smz023 (visited on 09/25/2025).

Luiten, A., J. Hox, and E. de Leeuw (Sept. 2020). "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys". EN. In: *Journal of Official Statistics* 36.3, pp. 469–487. ISSN: 0282-423X. DOI: 10.2478/jos-2020-0025. URL: https://doi.org/10.2478/jos-2020-0025 (visited on 10/21/2025).

Lundmark, S. and K. Backström (Oct. 2025). "Predicting Survey Nonresponse with Registry Data in Sweden between 1993 and 2023: Cohort Replacement or a Deteriorating Survey Climate?" en. In: *Survey Research Methods* 19.3, pp. 247–265. ISSN: 1864-3361. DOI: 10.18148/srm/2025.v19i3.8278. URL: https://ojs.ub.uni-konstanz.de/srm/article/view/8278 (visited on 10/21/2025).

Mercer, A. W., F. Kreuter, S. Keeter, and E. A. Stuart (Apr. 2017). "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference". In: *Public Opinion Quarterly* 81.S1, pp. 250–271. ISSN: 0033-362X. DOI: 10.1093/poq/nfw060. URL: https://doi.org/10.1093/poq/nfw060 (visited on 10/21/2025).

Olsson, U, F Drasgow, and N. Dorans (1982). "The polyserial correlation coefficient". In: *Psychometrika* 47.3, pp. 337–347.

Presser, S. and S. McCulloch (July 2011). "The growth of survey research in the United States: Government-sponsored surveys, 1984–2004". In: *Social Science Research* 40.4, pp. 1019–1024. ISSN: 0049-089X. DOI: 10.1016/j.ssresearch.2011.04.004. URL: https://www.sciencedirect.com/science/article/pii/S0049089X11000718 (visited on 10/13/2025).

Rubin, D. B. (1976). "Inference and Missing Data". In: *Biometrika* 63.3, pp. 581–592. ISSN: 0006-3444. DOI: 10.2307/2335739. URL: https://www.jstor.org/stable/2335739 (visited on 10/10/2025).

Schroeder, H. M. and B. T. West (Feb. 2025). "Analyzing Potential Non-Ignorable Selection Bias in an Off-Wave Mail Survey Implemented in a Long-Standing Panel Study". In: *Journal of Survey Statistics and Methodology* 13.1, pp. 100–127. ISSN: 2325-0992. DOI: 10.1093/jssam/smae039. URL: https://doi.org/10.1093/jssam/smae039 (visited on 09/25/2025).

Slawski, M., B. T. West, P. Bukke, Z. Wang, G. Diao, and E. Ben-David (July 2025). "A general framework for regression with mismatched data based on mixture modelling". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 188.3, pp. 896–919. ISSN: 0964-1998. DOI: 10.1093/jrsssa/qnae083. URL: https://doi.org/10.1093/jrsssa/qnae083 (visited on 11/02/2025).

West, B. T. and R. R. Andridge (Aug. 2023). "Evaluating Pre-election Polling Estimates Using a New Measure of Non-ignorable Selection Bias". In: *Public Opinion Quarterly* 87.S1, pp. 575–601. ISSN: 0033-362X. DOI: 10.1093/poq/nfad018. URL: https://doi.org/10.1093/poq/nfad018 (visited on 09/25/2025).

West, B. T., R. J. Little, R. R. Andridge, P. S. Boonstra, E. B. Ware, A. Pandit, and F. Alvarado-Leiton (Sept. 2021). "Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys". In: *The Annals of Applied Statistics* 15.3, pp. 1556–1581. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/21-AOAS1453. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-

15/issue-3/Assessing-selection-bias-in-regression-coefficients-estimated-from-nonprobability-samples/10.1214/21-AOAS1453.full (visited on 09/25/2025).

Williams, D. and J. M. Brick (June 2018). “Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort”. In: *Journal of Survey Statistics and Methodology* 6.2, pp. 186–211. ISSN: 2325-0984. DOI: 10.1093/jssam/smx019. URL: https://doi.org/10.1093/jssam/smx019 (visited on 10/13/2025).

Yiadom, S. A. and R. Andridge (Sept. 2024). *Formulating the Proxy Pattern-Mixture Model as a Selection Model to Assist with Sensitivity Analysis*. arXiv:2409.18117. DOI: 10.48550/arXiv.2409.18117. URL: http://arxiv.org/abs/2409.18117 (visited on 09/25/2025).

Modeling Complex Survey Data: a Case Study of International Health Surveillance Surveys

Timothy Raxworthy[1], Yajuan Si[2] and Grace Chung[3]

[1]University of Michigan, Ann Arbor, MI 48104 USA, raxworth@umich.edu
[2]Institute for Social Research, 426 Thompson St, Ann Arbor, MI 48104 USA, yajuan@umich.edu
[3]University of New Mexico, 2300 Tucker Ave NE, Albuquerque, NM 87131 USA, gchung@salud.unm.edu

Abstract

Model-based approaches to inference are common in the presence of complex survey data. Although statistical modeling is an often necessary approach for analyzing data, there is no firm consensus as to how these analyses should handle sampling weights. Using a case study of international health surveillance surveys, this paper examines the roles of weights in the generalized linear models (GLMs) and generalized linear mixed-effect models (GLMMs). We considered two different ways of including weights with model estimates: using weighted likelihood functions for model fitting and weighted average values of individual predictions. We compared GLM and GLMM estimates as well as unweighted and weighted variants of these models. We found that including weights in the model fitting processes does not substantially change the model parameter estimates and predictions. The difference between weighted and unweighted descriptive statistics is more pronounced than that of the model parameter estimates. We recommend comparing the weighted and unweighted descriptive summaries as a standard analysis routine in practice.

Keywords: multilevel modeling, complex sample survey, weights, diabetes prevalence.

1 Introduction

Most researchers in the social sciences and public health use sample survey data for finite population inference and account for complex sample design features if they are related to the survey outcome (Si, Lee, and Heeringa, 2024). When the sampling design features are informative, e.g., the selection leads to the sample distribution of the quantity of interest deviating from the underlying population distribution, appropriate analysis methods are necessary to adjust for the sample discrepancy. Despite this general recommendation, there have been historical debates on whether survey weights are necessary when fitting statistical models. Survey weights often require special treatments to meet researchers' analytic goals and are sometimes considered to be a nuisance when they only inflate the variance of an estimate without changing the point estimate itself. In addition to model specification, the various ways handling weights can also be due to factors such as researchers' familiarity with using survey weights, the quality of documentation on how weights are constructed, and the availability of auxiliary information about the population. Using a case study of international health surveillance surveys, this paper examines the roles of weights in the generalized linear models (GLMs) and generalized linear mixed-effect models (GLMMs).

Motivated by the need of meeting the World Health Organization's (WHO) recommended diabetes targets (Gregg et al., 2023), we aim to estimate and compare diabetes prevalence between countries. We use the WHO Stepwise Approach to Surveillance (STEPS) surveys (Riley et al., 2016), which are cross-sectional probability sample surveys conducted in more than 100 countries and collect various health risk indicators to provide population level estimates. Most STEPS surveys have implemented multistage stratified data collections. First, enumeration areas are selected as primary sampling units (PSUs) and then households are selected as secondary sampling units. The final stage randomly selects eligible individuals from each household. Following the design of the STEPS survey with multistage stratified sampling, every individual i in the sample was assigned a survey weight w_i. To properly estimate the sampling variance, it is necessary to account for stratification, PSU clustering, and weights in the analysis. Further, we are also interested in assessing whether the sample survey design features affect the estimation of diabetes prevalence across multiple countries.

The paper structure is organized as follows. Section 1.1 describes the data and measures used in this study in detail. Section 2 introduces the model-based inference approach. The results from different methods in Section 3 are then compared. Section 4 summarizes the main takeaways from the study.

1.1 Data and Measures

We used a subset of the WHO's STEPS data in this analysis, which were collected from 10 different countries between 2015 and 2016. This dataset was processed, resulting in 26,752 individuals. We define a case with diabetes as having any of the following: (1) a fasting plasma glucose of 7.0 mmol/L or higher, (2) hemoglobin A1c (HbA1c) level of 6.5% or higher, or (3) self-reported use of glucose-lowering medication or use of insulin or oral hypoglycemic drugs. This definition of diabetes is used in the Global Monitoring Framework for Non Communicable Diseases (Gregg et al., 2023). Once these cases are computed, a binary indicator variable for diabetes is then defined and used as the dependent variable in the analysis. Eligible participants who have been assigned survey weights are included. The person level covariates used in predicting our outcome are body mass index (BMI), age, sex, and highest completed education level.

Table 1: Description of Each Country's Sample and Sampling Weights

Country	Sample Size	Sum of Weights	Mean	SE
Algeria	6,393	25,888,236	4,049.47	1,316.80
Benin	5,073	2,441,103	481.20	1,686.59
Brunei	2,018	102,824	50.95	43.13
Ethiopia	9,800	34,097,395	3,479.33	3,707.59
Guyana	1,178	203,440	172.70	144.86
Iraq	4,071	14,942,707	3,670.53	3,569.90
Kiribati	2,156	57,561	26.70	59.80
Nauru	1,387	4,212	3.04	0.33
Solomon Is.	2,522	341,164	135.28	123.54
Vietnam	3,758	78,831,165	20,976.89	17,094.49

Table 1 includes further detailed information regarding each country sample. The 'Sample Size' column refers to the total number of eligible participants assigned a sampling weight and the 'Sum of Weights' is the calculated total sum of the sampling weights. The 'Mean' column is the mean value of the sampling weights and 'SE' column is the standard error of the weights.

There were missing values for some respondents' BMI and education. The amount of item nonresponse is less than 5% of the records. We assumed missing at random and used multiple imputation to fill in missing item values with a proportional odds model accounting for the order of the BMI and education categories. Along with these variables, other covariates used in multiple imputation were the indicator variable for a diabetic case, age, and sex. We used one imputed dataset for simplicity, even though multiple completed datasets could have been pooled for analysis via combining rules.

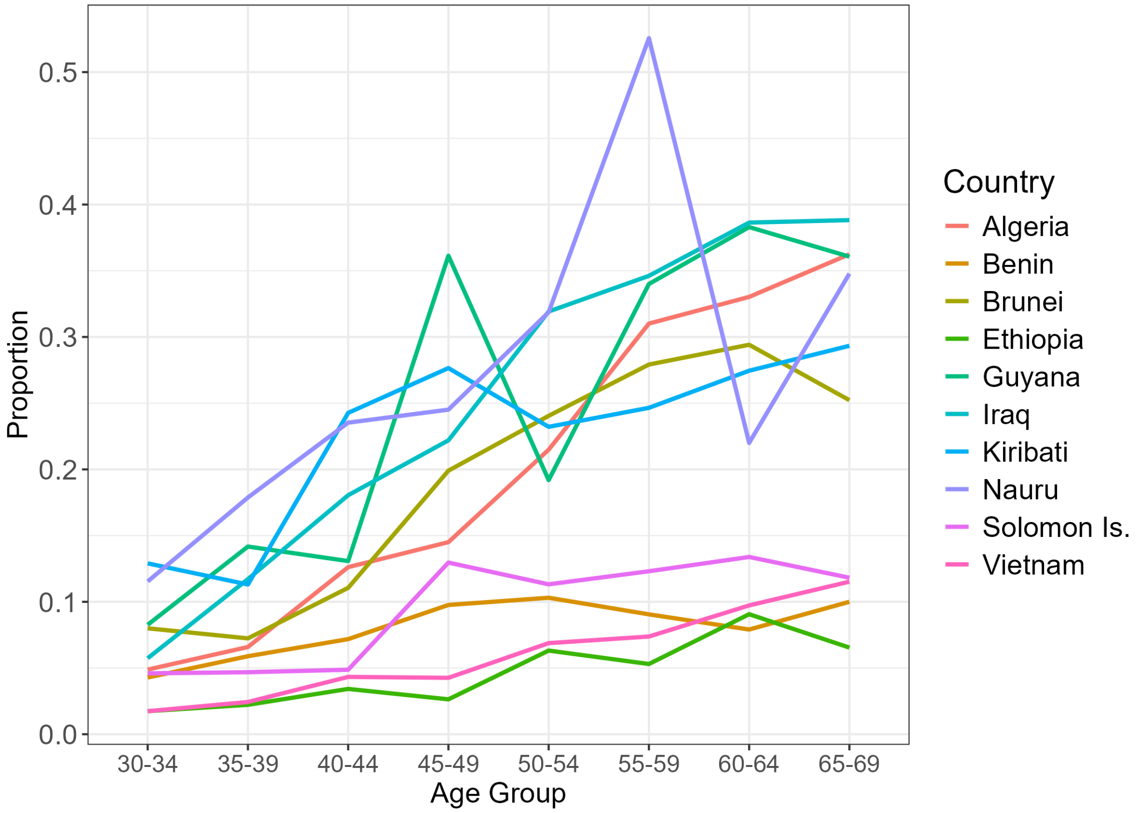


Figure 1: Observed proportions of individuals with diabetes by age group and country.

Figure 1 displays how the observed proportions of individuals with diabetes change across age groups for each country included in our analysis. The generally increasing trends are similar between countries but the changing rates over time are accentuated for Algeria, Brunei, Guyana, Iraq, Kiribati and Nauru. Table 1 shows that the sample sizes across countries also varies largely. Further, the likely cause for the spikes shown at specific age cohorts is due to small sample size for those ages.

2 Modeling Approach

We use the individual level data to model the probability of having diabetes in a logistic regression. First, we define the outcome variable as:

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ in country } j \text{ has the disease,} \\ 0 & \text{if individual } i \text{ in country } j \text{ does not have the disease,} \end{cases}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, where n_j is the total number of individuals, and J is the total number of countries. Our GLM model is specified as

$$log\left\{\frac{Pr(y_{ij} = 1|\boldsymbol{X'_{ij}})}{Pr(y_{ij} = 0|\boldsymbol{X'_{ij}})}\right\} = \boldsymbol{X'_{ij}\beta}, \tag{1}$$

where $\boldsymbol{X'_{ij}}$ denotes the person-level covariates, including BMI, age, sex, and education, and country indicators.

In the GLMM, we include country-level and PSU-level random intercepts to borrow information across countries and clusters, stabilize estimates for small countries and account for the PSU clustering effects with the following specification:

$$log\left\{\frac{Pr(y_{ij}=1|\boldsymbol{X'},u_j,v_{k[i]})}{Pr(y_{ij}=0|\boldsymbol{X'},u_j,v_{k[i]})}\right\}=\boldsymbol{X'\beta}+u_j+v_{k[i]}, \tag{2}$$

where $\boldsymbol{X'}$ denotes the person-level covariates, u_j's are the country-varying effects, and $v_{k[i]}$'s are the PSU-varying effects, where $k[i]$ is the PSU index k that individual i is assigned to, $u_j \overset{iid}{\sim} N(0,\sigma_u^2)$, $v_k \overset{iid}{\sim} N(0,\sigma_v^2)$, with both random effects assumed to be independent, identically and normally distributed with a mean of 0 and variance σ_u^2 and σ_v^2, respectively. The intra-country correlation (ICC) is also measured:

$$\rho=\frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2+\frac{\pi^2}{3}}.$$

Although we considered different interactions between the categorical variables and country level predictors, we did not include additional covariates because of model estimation problems. We consider two different ways to account for weights in estimating the country-specific diabetes prevalence as the proportion of people with the characteristics described in 1.1.

1) Using a weighted average of the estimated predicted probability $\hat{p}_{ij}=Pr(y_{ij}=1|\boldsymbol{X'_{ij}})$ of having diabetes for individual i in country j based on Model (1) conditional on $\boldsymbol{X'_{ij}}$. For Model (2) this is expressed as $Pr(y_{ij}=1|\boldsymbol{X'},u_j,v_{k[i]})$. The weighted average of the predicted probabilities is given based on the Hájek estimator (Hájek, 1971). For country j, the weighted prevalence is given by

$$\hat{\theta}_j=\frac{\sum_{i\in s_j} w_i * \hat{p}_{ij}}{\sum_{i\in s_j} w_i}, \tag{3}$$

where s_j is the sample of individuals in country j.

2) Including weights in the model fitting processes of either the GLM in (1) or the GLMM in (2). We use the pseudo maximum likelihood (PML) estimation to obtain the parameter estimates that maximum the weighted likelihood function, where each individual's likelihood is powered by the corresponding weight value (Skinner, 1989). The weighted GLM likelihood $l_{WGLM}(\cdot)$ and weighted GLMM likelihood $l_{WGLMM}(\cdot)$ are as below.

$$l_{WGLM}(y,\boldsymbol{X'},\beta,w)=\prod_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{X'}_{ij}\beta)}{1+\exp(\boldsymbol{X'}_{ij}\beta)}\right)^{y_{ij}}\left(\frac{1}{1+\exp(\boldsymbol{X'}_{ij}\beta)}\right)^{(1-y_{ij})}\right]^{w_i} \tag{4}$$

$$l_{WGLMM}(y,\boldsymbol{X'},\beta,u,v,w)=\prod_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{X'\beta}+u_j+v_{k[i]})}{1+\exp(\boldsymbol{X'\beta}+u_j+v_{k[i]})}\right)^{y_{ij}}\left(\frac{1}{1+\exp(\boldsymbol{X'\beta}+u_j+v_{k[i]})}\right)^{(1-y_{ij})}\right]^{w_i}. \tag{5}$$

We normalize the weights when fitting the GLMM. Rabe-Hesketh and Skrondal, 2006 show that the weighted likelihood function requires weights at each level of the data hierarchy. Based on the STEPS survey design, different countries independently conducted the surveys, and there was no random selection of countries. To effectively pool estimates across countries, we scaled the person-level weights using method 2 described in Pfeffermann et al., 1998. The weight adjustment scales the weights w_i of individuals in country j, for $i \in s_j$, by adjusting the sum to be equal to the sample size of each country n_j. The adjustment factor a_j for individuals in country j can be expressed as: $a_j=\frac{n_j}{\sum_{i\in s_j} w_i}$, the product of which and the weight w_i will be used in the pseudo maximum likelihood (PML) estimation based on the GLMM. This scaling method has been described as performing better in simulations where the design is considered informative (Pfeffermann et al., 1998).

3 Model Inference

We fit GLM and GLMM models, both unweighted and weighted, to predict the diabetes prevalence with age (30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, and 65-69), BMI ($<$25, $\leq$25 & $<$30, and $\geq$30), education (no education, primary school, high school [HS] & above), sex (male, and female), and country indicators (Algeria, Benin, Brunei, Ethiopia, Guyana, Iraq, Kiribati, Nauru, Solomon Islands, and Vietnam).

We account for stratification and clustering in the standard error estimation for both models. We use analytic variance estimation via Taylor series linearization by default in the R *survey* package (Lumley, 2024) and Stata (StataCorp, 2025), i.e., defining the complex survey design object including strata codes, PSU codes, and sampling weights (for weighted estimates). To obtain country-specific estimates, we apply the unconditional approach for variance estimation and takes the full complex sample design into account when analyzing subpopulations (Heeringa, West, and Berglund, 2017). When summarizing the model predictions, the complex survey design features (PSUs, strata and weights) are accounted for to obtain design-based estimates using expression (3).

3.1 Model Estimation

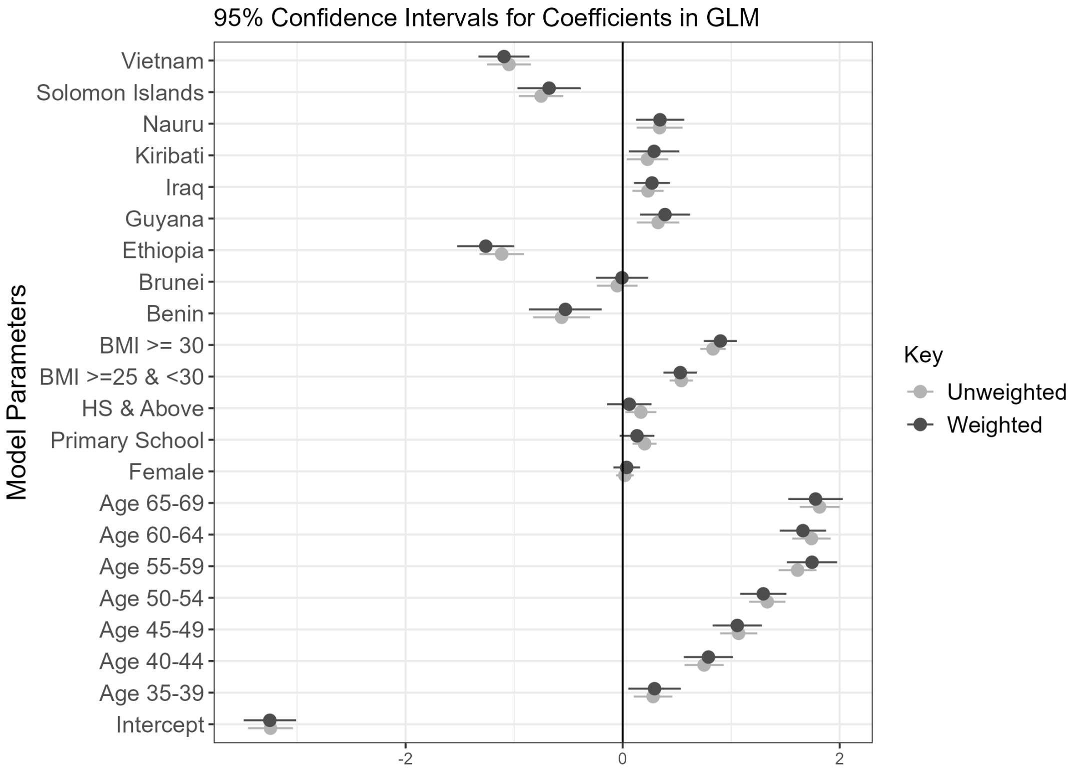


Figure 2: Comparison of coefficient estimates and 95% confidence intervals for the true coefficients between unweighted GLM 1 and weighted GLM 4.

First, we fit the unweighted GLM in (1) and weighted GLM in (4). The reference categories for each of the predictors are those persons who are aged 30-34, male, have received no formal education, have a BMI $<$ 25 and reside in Algeria which is the first country in alphabetical order. The model coefficient estimates together with the corersponding 95% confidence intervals for the true coefficients are presented in Figure 2 . Including weights in model fitting tends to increase the variance of coefficient estimation, resulting in wider confidence intervals as shown in Figure 2.

When comparing the coefficient point estimates of the weighted model versus the unweighted model, the weights slightly change our interpretation of the coefficients associated with each country. For example, across all countries, the weighted estimate's 95% confidence interval for the effect of primary school contain zero when its corresponding unweighted estimate does not. For all countries, the age coefficient estimates in the weighted model are either lower or higher than those in the unweighted model, changing without any apparent pattern. The opposite trend is the case for coefficients associated with education where the weights are reducing the magnitude for the effects of both age categories. These subtle changes are due to the differences in the maximum likelihood estimates when the weights are added in the GLM model fitting. The role of weights depends on the model specification and its dependency on the sample design features. Therefore, fitting both unweighted and weighted models is helpful when analyzing complex survey samples to gain more insight into how the weights interact with the model fitting process.

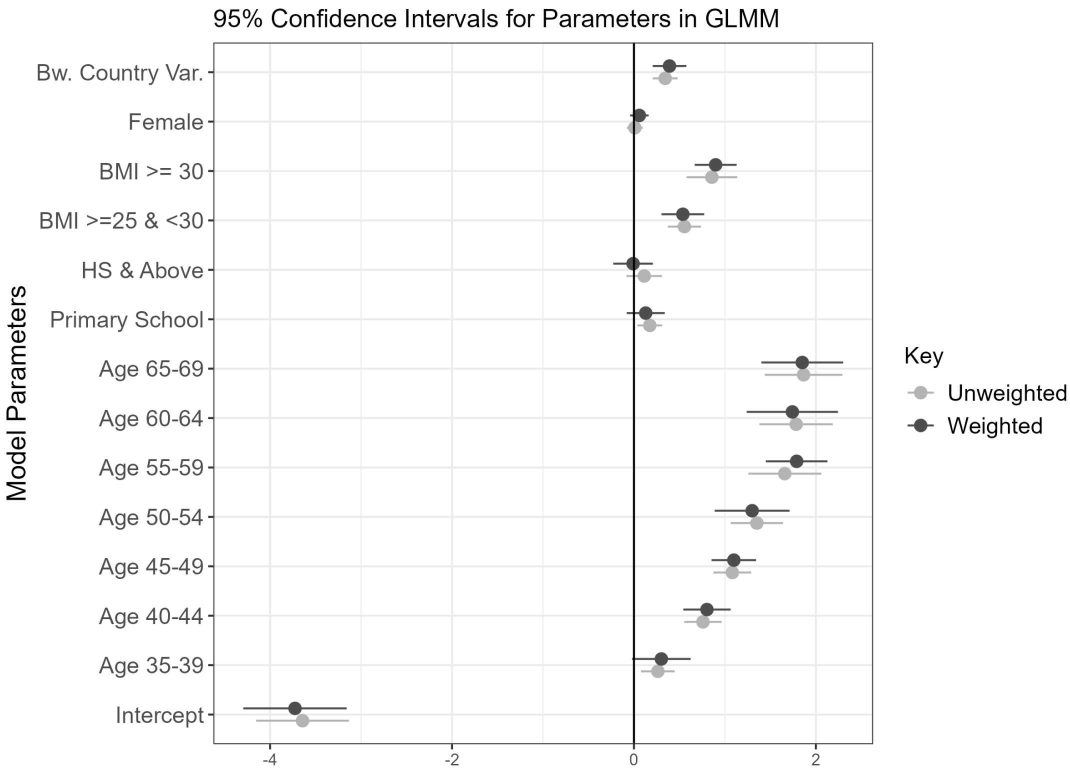


Figure 3: Comparison of coefficient estimates and 95% confidence intervals between unweighted and weighted generalized linear mixed-effect models.

Next we fit the unweighted GLMM in (2) and weighted GLMM in (5). The ICC measuring the intra-country similarity is estimated to be 0.084 for the unweighted model and 0.094 in the weighted model. Figure 3 compares the estimated model coefficients for the weighted and unweighted GLMM models. In general, the model coefficients are similar between both models, although the weighted GLMM model estimates have larger variances.

In sum, with either GLM or GLMM, including the weights in model fitting does not substantially change the model parameter estimates.

Table 2: Model predicted (Pred), weighted model predicted (Weighted pred), observed (Obs) and weighted observed (Weighted Obs.) diabetes prevalence in 10 countries with standard error values reported in parentheses.

Country	Obs	Pred		Pred	
		GLM	Weighted GLM	GLMM	Weighted GLMM
Algeria	.166 (.007)	.166 (.002)	.163 (.002)	.158 (.004)	.154 (.004)
Benin	.073 (.008)	.073 (.001)	.075 (.001)	.066 (.003)	.067 (.003)
Brunei	.184 (.011)	.184 (.003)	.182 (.003)	.175 (.006)	.168 (.008)
Ethiopia	.036 (.003)	.036 (.001)	.032 (.001)	.032 (.001)	.027 (.001)
Guyana	.234 (.014)	.234 (.004)	.239 (.004)	.219 (.005)	.223 (.005)
Iraq	.215 (.009)	.215 (.002)	.220 (.002)	.204 (.004)	.205 (.004)
Kiribati	.208 (.013)	.208 (.003)	.213 (.003)	.201 (.009)	.201 (.008)
Nauru	.233 (.015)	.233 (.004)	.232 (.004)	.228 (.013)	.227 (.013)
Solo. Is.	.082 (.007)	.082 (.001)	.086 (.001)	.076 (.004)	.078 (.006)
Vietnam	.053 (.004)	.053 (.001)	.050 (.001)	.047 (.001)	.044 (.001)

Country	Weighted Obs.	Weighted pred		Weighted pred	
		GLM	Weighted GLM	GLMM	Weighted GLMM
Algeria	.153 (.007)	.156 (.002)	.153 (.002)	.147 (.004)	.144 (.004)
Benin	.069 (.011)	.067 (.003)	.069 (.003)	.062 (.005)	.063 (.007)
Brunei	.164 (.014)	.164 (.003)	.164 (.003)	.159 (.006)	.154 (.008)
Ethiopia	.030 (.003)	.035 (.001)	.030 (.001)	.029 (.001)	.026 (.001)
Guyana	.220 (.017)	.215 (.005)	.220 (.005)	.201 (.005)	.204 (.007)
Iraq	.217 (.011)	.212 (.003)	.217 (.003)	.204 (.005)	.205 (.005)
Kiribati	.211 (.014)	.207 (.007)	.211 (.007)	.199 (.007)	.206 (.011)
Nauru	.235 (.015)	.236 (.005)	.235 (.005)	.231 (.013)	.230 (.013)
Solo. Is.	.079 (.010)	.076 (.001)	.079 (.001)	.071 (.004)	.073 (.006)
Vietnam	.046 (.004)	.049 (.001)	.046 (.001)	.043 (.001)	.040 (.001)

3.2 Model Prediction

Using both the weighted and unweighted variants of the GLM and GLMM described above, we predict the response probability for each individual and estimate prevalence by country. The top table in Table 2 compares the observed prevalence as the simple proportion of diabetic cases divided by the total sample size with the predictions from the GLM in (1), weighted GLM in (4), GLMM in (2), and weighted GLMM in (5). The bottom table in Table 2 applies weights to the individual probabilities based on (3) and presents the weighted observation using expression (3) and prediction values. Based on Table 2, model-based predictions have lower standard errors than the observed prevalence of each country. We have omitted the model-based error in predicting the probabilities, but the prediction variability is smaller than the sampling error. The country-level prevalence values calculated as the average of individual predictions from the GLM are the same as the observed summaries, which is as expected because the GLM model includes the fixed effects of countries. Nevertheless, the GLMM includes the random effects of countries, and partial pooling across countries yields predicted summaries different from the observed values. Comparing GLM and GLMM before and after weighting, we see that using weighted likelihood estimation does not substantially change the predicted values or standard errors. The predictions based on GLMM have more variability than those based on GLM, probably due to the inclusion of random effects.

Applying weights to the individual probabilities based on (3) does change the point estimates across countries and increases the standard errors. This is true for both observed and predicted values from all models. The weighted predictive averages based on the weighted GLM are the same as the weighted observed summaries but the standard errors are lower. The reason for the weighted predictive averages being the same is because the weights have been normalized to equal the sum of the sample size within each country. Overall, this shows that using a weighted average of the predicted probabilities generates larger influences than including weights in the model fitting processes. Weights are more influential for descriptive summaries than model estimates, which is consistent with the literature findings, e.g., Si, Lee, and Heeringa, 2024.

4 Conclusion

In this study, we used a case study with international health surveys to assess the role of survey weights in model inference and prevalence estimation. We considered two different ways of including weights with model estimates: using weighted likelihood functions for model fitting and weighted average values of individual predictions. We compared GLM and GLMM estimates as well as unweighted and weighted variants of these models. We found that including weights in the model fitting processes does not substantially change the estimated model coefficients and predictions. The difference between weighted and unweighted prevalence summaries is more pronounced than that of the model parameter estimates. We recommend comparing the weighted and unweighted descriptive summaries as a standard analysis routine in practice.

Finally, our empirical comparisons cannot be validated without knowing the gold-standard or true values. When comparing these diabetes estimates to those published in other sources, there may be some cases where the unweighted estimates are closer to estimates found by experts in these fields, while other cases have weighted estimates with closer comparisons. This dilemma demonstrates the need to work with experts in the topic of analysis who can properly evaluate survey estimates beyond the statistical component. As pointed out above, we only use one set of predicted probabilities for the point estimates and omitted the prediction error due to model fitting. The model-based error is smaller than the sampling error. Future work would be needed to develop practical methods that account for both modeling and sampling error, such as using Monte Carlo simulation.

5 Data and Software

This paper uses data from the Algeria 2016 (Ministry of Health (Algeria), Population and Hospital Reform, and World Health Organization (WHO), 2017), Benin 2015 (Ministry of Health (Benin), and World Health Organization (WHO), 2015), Brunei Darussalam 2015-2016 (Ministry of Health (Brunei), and World Health Organization (WHO), 2016), Ethiopia 2015 Ethiopia Public Health Institute, Federal Ministry of Health (Ethiopia) and World Health Organization (WHO), 2016, Guyana

2016 (Pan American Health Organization (PAHO), Ministry of Public Health of Guyana, and the Bureau of Statistics (Guyana), 2019), Iraq 2015 (Ministry of Health (Iraq), Ministry of Planning (Iraq), World Health Organization (WHO), 2015), Kiribati 2015-2016 (Ministry of Health and Medical Services (Kiribati), World Health Organization (WHO), 2015), Nauru 2015 (Ministry of Health (Nauru), World Health Organization (WHO), 2016), Solomon Islands 2015 (Ministry of Health (Solomon Islands), World Health Organization (WHO), 2020), and Viet Nam 2015 (Ministry of Health (Vietnam), World Health Organization (WHO), 2016) STEPS surveys. These surveys were implemented by the agencies listed in each citation along with support by the World Health Organization. These datasets are available upon request from the https://extranet.who.int/ncdsmicrodata/index.php/homeWHO NCD Microdata Repository.

The svyglm function in the https://cran.r-project.org/package=surveysurvey package for R software was used to fit GLMs (Lumley, 2010). The https://www.stata.com/manuals/memelogit.pdfmelogit function within Stata software was used to fit the GLMMs.
The https://cran.r-project.org/package=micemice package of R software was used for multiple imputation (Van Buuren and Groothuis-Oudshoorn, 2011).

References

Ethiopia Public Health Institute, Federal Ministry of Health (Ethiopia) and World Health Organization (WHO) (2016). *Ethiopia Steps Report On Risk Factors for Chronic Noncommunicable Diseases and Prevalence of selected NCDs*. World Health Organization STEPS Survey Repository.

Gregg, E. W., J. Buckley, M. K. Ali, J. Davies, D. Flood, and et al (2023). “Improving health outcomes of people with diabetes: target setting for the WHO Global Diabetes Compact”. In: *Lancet* 401.10384, pp. 1302–1312.

Hájek, J. (1971). “Comment on ”An Essay on the Logical Foundations of Survey Sampling, Part One””. In: *The Foundations of Survey Sampling*.

Heeringa, S. G., B. T. West, and P. A. Berglund (2017). *Applied Survey Data Analysis*. 2nd. CRC Press.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley and Sons.

— (2024). *survey: analysis of complex survey samples*. R package version 4.4.

Ministry of Health (Algeria), Population and Hospital Reform, and World Health Organization (WHO) (2017). *Algeria STEPS Noncommunicable Disease Risk Factors Survey 2016–2017*. World Health Organization STEPS Survey Repository.

Ministry of Health and Medical Services (Kiribati), World Health Organization (WHO) (2015). *Kiribati NCD Risk Factors STEPS REPORT 2015-2016*. World Health Organization STEPS Survey Repository.

Ministry of Health (Benin), and World Health Organization (WHO) (2015). *Benin STEPS Noncommunicable Disease Risk Factors Survey 2015*. World Health Organization STEPS Survey Repository.

Ministry of Health (Brunei), and World Health Organization (WHO) (2016). *Brunei STEPS Noncommunicable Disease Risk Factors Survey 2015-2016*. World Health Organization STEPS Survey Repository.

Ministry of Health (Iraq), Ministry of Planning (Iraq), World Health Organization (WHO) (2015). *Noncommunicable Diseases Risk Factors STEPS Survey Iraq 2015 Report*. World Health Organization STEPS Survey Repository.

Ministry of Health (Nauru), World Health Organization (WHO) (2016). *Nauru NCD RIsk Factors STEPS Report 2015-2016*. World Health Organization STEPS Survey Repository.

Ministry of Health (Solomon Islands), World Health Organization (WHO) (2020). *Solomon Islands NCD Risk Factors STEPS REPORT 2015*. World Health Organization STEPS Survey Repository.

Ministry of Health (Vietnam), World Health Organization (WHO) (2016). *National Survey on the Risk Factors of Non-Communicable Diseases (STEPS) Viet Nam 2015*. World Health Organization STEPS Survey Repository.

Pan American Health Organization (PAHO), Ministry of Public Health of Guyana, and the Bureau of Statistics (Guyana) (2019). *Pan American STEPS Risk Factor Surveillance Survey Guyana 2016 Report*. World Health Organization STEPS Survey Repository.

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash (1998). “Weighting for unequal selection probabilities in multilevel models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1, pp. 23–40.

Rabe-Hesketh, S. and A. Skrondal (2006). “Multilevel modelling of complex survey data”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.4, pp. 805–827.

Riley, L., R. Guthold, M. Cowan, S. Savin, L. Bhatti, T. Armstrong, and R. Bonita (2016). “The World Health Organization STEPwise Approach to Noncommunicable Disease Risk-Factor Surveillance: Methods, Challenges, and Opportunities”. In: *American Journal of Public Health* 106.1, pp. 74–78.

Si, Y., S. Lee, and S. G. Heeringa (2024). “Population weighting in statistical analysis”. In: *JAMA Internal Medicine* 184.1, pp. 98–99.

Skinner, C. J. (1989). “Domain means, regression and multi-variate analysis”. In: *Analysis of Complex Surveys*. Ed. by C. J. Skinner, D. Holt, and T. M. F. Smith. Wiley, pp. 59–88.

StataCorp (2025). *Stata Statistical Software: Release 19*.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3, pp. 1–67.

Variance Component Estimation Under a General Area-level Model

Yuting Chen[1] **and Hanqing Li**[2]

[1]Eastern Kentucky University, USA, yuting.chen@eku.edu
[2]University of Calgary, Canada, hanqing.li@ucalgary.ca

Abstract

Small area estimation (SAE) provides reliable inference for domains with limited survey sample data by borrowing strength across areas through modeling. The typical area-level model assumes normally distributed random effects, an assumption that may not hold in practice. This paper empirically examines the performance of residual maximum likelihood (REML) and adjusted REML estimators under general area-level models with non-normal random effects. Using simulations with heavy-tailed and asymmetric distributions, we evaluate point estimation and prediction interval performance. REML remains reasonably robust in estimating the variance component and supporting reliable predictions, but zero boundary estimates can degrade interval performance when the number of domains is small. Adjusted REML reduces boundary issues and yields more reliable interval coverage while maintaining competitive estimation accuracy. These results highlight adjusted likelihood methods as a practical and robust option even when the normality assumption is uncertain.

Keywords: data linkage, survey statistics, uncertainty quantification.

1 Introduction

In survey sampling, researchers often aim to estimate population parameters such as totals, means, or proportions based on data from a representative sample. In many practical settings, however, it is also of interest to estimate similar characteristics for specific subpopulations or domains (e.g., regions, demographic groups, or institutions). Large-scale surveys are typically designed to yield reliable estimates for large domains, but for smaller domains, the sample sizes may be too small or even zero, to produce direct estimates with acceptable precision. This situation gives rise to the small area problem.

To address this challenge without increasing sample sizes, small area estimation (SAE) techniques have been developed to “borrow strength” across related areas. Model-based SAE methods achieve this by linking data from different areas through statistical models that include area-specific random effects and auxiliary information. These approaches enable more precise and reliable estimation of small area parameters.

Suppose that the population of interest, U, is partitioned into m areas (or subpopulations), denoted by $U_1, \cdots, U_m$ and that we are interested in estimating the corresponding area means $\{\theta_i,\ i = 1, \cdots, m\}$. Let s_i denote the sample drawn from area U_i. When the sample size n_i is small, we may encounter the small area issue. A widely used framework in SAE is the two-level area-level model, which for area $i = 1, \cdots, m$, can be expressed as:

$$
\begin{aligned}
&\text{Level 1: (Sampling model): } \hat{y}_i|\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, D_i);\\
&\text{Level 2: (Linking model): } \theta_i \overset{\text{ind}}{\sim} \mathcal{G}(\mathbf{x}_i'\beta, A, \phi).
\end{aligned} \tag{1}
$$

The Level 1 model represents the sampling distribution of the direct estimator $\hat{y}_i$, which may be a weighted or unweighted estimate for area i. For example, $\hat{y}_i$ could be the sample mean based on n_i observations from area i with sampling variance $D_i = \sigma^2/n_i$, where σ^2 is known or reliably estimated from all areas (Fay and Herriot, 1979; Otto and Bell, 1995; Hawala and Lahiri, 2018). The Level 2 model links the true small area means θ_i to a vector of known auxiliary variables $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$, often obtained from administrative records, census data, or other external sources. We assume that the Level 2 distribution $\mathcal{G}$ is a fully parametric distribution, not necessarily normal, with mean $\mathsf{E}(\theta_i) = \mathbf{x}_i'\beta$, variance $\mathsf{Var}(\theta_i) = A \geq 0$, and any additional parameters ϕ. The coefficient vector $\beta \in \mathcal{R}^p$ and the variance component A are unknown and must be estimated from the data.

The classical area-level model proposed by Fay and Herriot (1979) assumes normality at both levels. The normality assumption at Level 1 may not be considered as restrictive as the normality of θ_i, due to the central limit theorem's effect on direct estimator $\hat{y}_i$ (Rao and Molina, 2015; Jiang and Torabi, 2022). To relax this assumption, recent studies have explored non-normal alternatives for the Level 2 distribution $\mathcal{G}$ (Chen, Hirose, and Lahiri, 2024). For instance, Bell and Huang (2006) used a t-distribution to mitigate the influence of outliers; Fabrizi and Trivisano (2010) proposed exponential power and skewed exponential power distributions to handle heavy-tailed or asymmetric effects; and Jiang and Torabi (2022) employed a skewed normal distribution.

The above two-level model can equivalently be expressed as the linear mixed model:

$$\hat{y}_i = \theta_i + e_i = \mathbf{x}_i'\beta + u_i + e_i, \quad i = 1, \cdots, m, \tag{2}$$

where random effects u_i's and sampling errors e_i's are independent with $u_i \overset{\text{i.i.d}}{\sim} \mathcal{G}(0, A, \phi)$ and $e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, D_i)$. The small area parameter of interest is $\theta_i = \mathbf{x}_i'\beta + u_i,\ i = 1, \cdots, m$. When A is known, the best linear unbiased predictor (BLUP) of θ_i that minimize the mean squared prediction error (MSPE) among all linear unbiased predictors, is given by

$$\tilde{\theta}_i = (1 - B_i)\hat{y}_i + B_i\mathbf{x}_i'\tilde{\beta}, \tag{3}$$

where $B_i = D_i/(A + D_i)$ is the shrinkage factor, $\tilde{\beta} = \tilde{\beta}(A)$ is the standard weighted least squares estimator of β. The BLUP effectively shrinks the direct estimator y_i toward the regression synthetic estimator $\mathbf{x}_i'\tilde{\beta}$, with the degree of shrinkage determined by B_i. In this paper, we assume $A > 0$. In practice, since A is unknown, it must be estimated from the data, leading to the empirical BLUP (EBLUP):

$$\hat{\theta}_i = (1 - \hat{B}_i)y_i + \hat{B}_i\mathbf{x}_i'\hat{\beta}, \tag{4}$$

where $\hat{B}_i = D_i/(\hat{A} + D_i)$ and $\hat{\beta} = \tilde{\beta}(\hat{A})$.

When $\mathcal{G}$ is normal, several methods have been proposed to estimate A, including the Fay-Herriot method-of-moments (FH) estimator (Fay and Herriot, 1979), the Prasad-Rao simple method-of-moments (PR) estimator (Prasad and Rao, 1990), the maximum likelihood (ML) estimators and the residual maximum likelihood (REML) estimators (Datta and Lahiri, 2000). When the number of areas m is small, standard variance estimation methods, particularly the PR estimator, often produce boundary estimate $A = 0$, leading to $\hat{B}_i = 1$ for all i, even when some of the true B_i are not close to 1 (Li and Lahiri, 2010; Chen, Hirose, and Lahiri, 2024). This causes an overshrinkage problem in EBLUP, since now the EBLUP of θ_i reduces to the regression synthetic estimator. Moreover, with $\hat{A} = 0$, it also causes the problem of degenerate distribution and prevents the use of parametric bootstrap methods for uncertainty quantification, such as estimating MSPE or constructing prediction intervals.

To address this issue under normal random effects, several adjusted likelihood methods have been developed to guarantee positive estimates of A (Li, 2007; Yoshimori and Lahiri, 2014; Hirose and Lahiri, 2018). These methods solve the two problems above simultaneously in SAE applications. In addition, they show that the biases of the adjusted ML and REML estimators are of order $O(m^{-1})$ (Li and Lahiri, 2010), and those of the parametric bootstrap MSPE being $o(m^{-1})$ (Hirose and Lahiri, 2018). However, the performance of these adjusted estimators when the random effects are non-normal remains largely unexplored.

In this study, we investigate methods for estimating variance components under a general area-level model that allows for possibly non-normal random effects. Laird and Ware, 1982 and Cressie, 1990, among others, have favored the REML method over the ML method for variance component estimation in complex small area models. This preference was later supported by Datta and Lahiri (2000), in which they showed that the REML estimator has a lower order of bias than the ML estimator. Therefore, in this paper, we focus on the REML approach. Following Jiang (1996), we define the REML estimator of variance components as the solution to the REML equations, which we introduce in the next section. Although Jiang (1996) theoretically showed that REML estimates are consistent under certain identifiability and information conditions, their empirical performance under non-normal random effects has not been well studied in SAE. We therefore (i) empirically evaluate the performance of REML estimators under various non-normal settings, and (ii) extend the adjusted REML methods of Li and Lahiri (2010) to the general area-level model, assessing their performance through Monte Carlo simulations.

The remainder of this paper is organized as follows. Section 2 provides the list of notations and regularity conditions. Section 3 reviews the estimation methods for variance components, including REML and adjusted REML estimators. Section 4 presents Monte Carlo simulation results comparing different estimators under various model settings. Section 5 concludes with a summary and discussion.

2 A list of notations and regularity conditions

We introduce the following notations that will be used throughout the paper:

$\mathbf{y} = (\hat{y}_1, \cdots, \hat{y}_m)'$, a $m \times 1$ column vector of direct estimates;

$X' = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$, a $p \times m$ known matrix of rank p;

$\Sigma = \text{diag}(A + D_1, \cdots, A + D_m)$, a $m \times m$ diagonal matrix;

$\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{y}$, weighted least square estimator of β with known A;

$\mathbf{P} = \Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$.

We assume the following regularity conditions throughout the paper:

r.1 rank$(X) = p$ is fixed;

r.2 $\sup_{i\geq 1} h_{ii} = O(m^{-1})$, where $h_{ii} = \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i$;

r.3 $0 < \inf_{i\geq 1} D_i \leq \sup_{i\geq 1} D_i < \infty$.

3 REML and adjusted REML estimators

The REML approach introduced by Patterson and Thompson (1971), eliminates dependence on nuisance parameters by basing inference on linear transformations of the data that remove the fixed effects. Under normality at both levels, the restricted likelihood function is given by:

$$L_{\mathrm{RE}}(A) = c|X'\Sigma^{-1}X|^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}\right) \tag{5}$$

where c is a constant independent of A. Let $l_{\mathrm{RE}}(A)$ denote the corresponding restricted log-likelihood. The REML estimator $\hat{A}_{\mathrm{RE}}$ satisfies:

$$\begin{aligned}\frac{\partial l_{\mathrm{RE}}(A)}{\partial A} &= \frac{1}{2}\left[\mathbf{y}'\mathbf{P}^2\mathbf{y} - \mathrm{tr}(\mathbf{P})\right] \\ &= 0\end{aligned} \tag{6}$$

In general (without assuming normality), the REML estimate $\hat{A}_{\mathrm{RE}}$ is defined as solution of (6).

Following Li and Lahiri (2010), we also consider the same adjusted restricted likelihood under the general area-level model:

$$L_{\mathrm{adj}}(A) = A \times L_{\mathrm{RE}}(A). \tag{7}$$

The adjusted maximum likelihood estimator $\hat{A}_{\mathrm{adj}}$ is obtained by maximizing $L_{\mathrm{adj}}(A)$ or its logarithmic form, $l_{\mathrm{adj}}(A)$.

Since $L_{\mathrm{RE}}(A)$ is a continuous positive function of A and $\lim_{A\to\infty} A \times L_{\mathrm{RE}}(A) = 0$ for $m > p + 2$, it follows from Lemma 2.1 of Li (2007) that the maximizer $\hat{A}_{\mathrm{adj}}$ is strictly positive.

Specifically, because $L_{\mathrm{RE}}(A) > 0$ for all A, we have $A \times L_{\mathrm{RE}}(A) \leq 0$ for $A \leq 0$ and $A \times L_{\mathrm{RE}}(A) > 0$ for $A > 0$. Moreover, since $A \times L_{\mathrm{RE}}(A) \to 0$ as $A \to \infty$, there exists some $A_0 > 0$ such that

$$A_0 \times L_{\mathrm{RE}}(A_0) = \max_A \{A \times L_{\mathrm{RE}}\},$$

which ensures that the maximizer A_0 is positive.

3.1 Parametric bootstrap prediction intervals

A traditional prediction interval for θ_i is of the form $\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\mathrm{mspe}}$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th standard normal percentile and mspe is an estimate of the mean squared prediction error of $\hat{\theta}_i$. However, such intervals have coverage errors of order $O(m^{-1})$, which may be inadequate for small area applications. Chatterjee, Lahiri, and Li, 2008 proposed a parametric bootstrap method that constructs intervals from the bootstrap distribution approximation of $\hat{\sigma}_1^{-1}(\theta_i - \hat{\theta}_i)$ under a normal linear mixed model, where $\hat{\sigma}_1^2 = D_i(1 - \hat{B}_i)$. This method achieves improved coverage error of order $O(m^{-3/2})$.

Chen, Hirose, and Lahiri (2024) extended this method to the general area-level model (1) with non-normal level-2 distributions, and interestingly found that the bootstrap intervals can exhibit overcoverage under certain conditions. Their simulations also showed that there was high percentage of zero estimates in $\hat{A}_{\mathrm{PR}}$ estimator which affects the performance of associated bootstrap intervals. The result is consistent with the findings in Li and Lahiri, 2010.

In this paper, we assess the performance of the similar parametric bootstrap procedures under non-normal models using the REML and adjusted REML estimators of A. Specifically, let

$$\hat{y}_i^* = \mathbf{x}_i'\hat{\beta} + u_i^* + e_i^*$$

where $u_i^* \overset{\text{iid}}{\sim} \mathcal{G}(0, \hat{A}, \hat{\phi})$ and $e_i^* \overset{\text{ind}}{\sim} N(0, D_i)$ for $i = 1, \cdots, m$. Denote by $\hat{\beta}^*$, $\hat{A}^*$, $\hat{\theta}_i^*$, and $\hat{\sigma}_1^*$ the quantities computed from bootstrap samples $\mathbf{y}^* = \{\hat{y}_i^*,\ i = 1, \cdots, m\}$, and let $\theta_i^* = \mathbf{x}_i'\hat{\beta} + u_i^*$. The bootstrap distribution of $\hat{\sigma}_1^{*-1}(\theta_i^* - \hat{\theta}_i^*)$ is then used to approximate the distribution of $\hat{\sigma}_1^{-1}(\theta_i - \hat{\theta}_i)$. For a given significance level α, let q_l and q_u denote the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap distribution, respectively. The parametric bootstrap prediction interval for θ_i is then given by $(\hat{\theta}_i + q_l\hat{\sigma}_1, \hat{\theta}_i + q_u\hat{\sigma}_1)$.

4 Monte Carlo Simulations

To empirically evaluate the performance of various variance estimators and their associated prediction intervals in small m settings, we consider $m = 10$ and $m = 15$. Following Li and Lahiri, 2010, we use an unbalanced pattern for the sampling variances (D_i), consisting of five groups of small areas with common D_i values within each group. Specifically, we set $D_i \in \{4.0, 0.6, 0.5, 0.4, 0.2\}$ and fix $A = 1$. Without loss of generality, we take $\mathbf{x}_i'\beta = 0$. To reflect practical conditions, we still estimate the mean even when it is theoretically zero. Since areas within each group are exchangeable, we summarize results by group means in the tables.

We consider two non-normal Level 2 distributions in the area-level model (1): (i) a t-distribution with 5 degrees of freedom (symmetric case), and (ii) a shifted exponential (SE) distribution (asymmetric case). For each distributional scenario, we generate $N = 1,000$ independent datasets $\{y_i,\ i = 1, \ldots, m\}$ and use 1,000 bootstrap samples to construct the parametric bootstrap prediction intervals.

We examine three estimators of A: the PR estimator $\hat{A}_{\text{PR}}$ which does not rely on distributional assumptions, the REML estimator $\hat{A}_{\text{RE}}$ and the adjusted REML estimator $\hat{A}_{\text{AR}}$. We use both bias and mean squared error to compare different estimators. Let $\hat{A}^{(j)}$ be the estimate for the jth simulation run. We compute the following Monte Carlo measures:

$$\text{Bias}(\hat{A}) = \frac{1}{N}\sum_{j=1}^{N}(\hat{A}^{(j)} - A), \quad \text{RMSE}(\hat{A}) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{A}^{(j)} - A)^2}.$$

Table 1 shows the percentages of zero estimates in $\hat{A}$ and $\hat{A}^*$. For $m = 10$, the PR estimator yields the highest rate of zero estimates in both $\hat{A}$ and $\hat{A}^*$. Under the shifted exponential distribution, REML also result in a zero estimate in $\hat{A}$ although the percentage of 0 is relatively low (about 0.1%). All methods can produce zero estimates in $\hat{A}^*$, and the adjusted REML estimator exhibits the lowest percentage in all cases. As m increases to 15, the chance of zero estimate decreases across all methods.

Table 2 summarizes the small-sample performance of the three variance estimators in terms of bias and RMSE. Both PR and REML generally show smaller bias than adjusted REML. Overall, REML achieves the best performance in terms of both bias and RMSE under both distributions. The performance of adjusted REML estimator improves as m increases in terms of both bias and RMSE.

In SAE applications, prediction is often the primary objective. To investigate prediction accuracy of EBLUP with different plug-in variance estimates, we approximate the true MSPE through Monte Carlo

Table 1: Percentages of zero estimates in $\hat{A}$ and $\hat{A}^*$ for different estimation methods.

m	$\hat{A}_{\mathrm{PR}}$	$\hat{A}_{\mathrm{RE}}$	$\hat{A}_{\mathrm{AR}}$	$\hat{A}^*_{\mathrm{PR}}$	$\hat{A}^*_{\mathrm{RE}}$	$\hat{A}^*_{\mathrm{AR}}$
			t $\{u_i\}$			
10	21.900	0	0	33.411	0.015	<0.001
15	11.700	0	0	25.120	0.001	0
			Shifted exponential $\{u_i\}$			
10	26.200	0.100	0	36.673	0.022	0.001
15	16.100	0	0	27.714	0.001	<0.001

Table 2: Comparison of different estimators of A for $m = 10$ and $m = 15$ with true value of $A = 1$.

	Monte Carlo Bias			**Monte Carlo RMSE**		
m	PR	RE	AR	PR	RE	AR
			t $\{u_i\}$			
10	0.035	-0.015	0.720	1.225	0.901	1.409
15	0.074	0.011	0.437	1.031	0.762	1.002
			Shifted exponential $\{u_i\}$			
10	0.053	-0.043	0.690	1.418	1.119	1.653
15	0.086	0.014	0.443	1.148	0.917	1.173

simulations. Let $\theta_i^{(j)}$ and $\hat{\theta}_i^{(j)}$ be simulated true value and the EBLUP for area i in the jth simulation respectively, $i = 1, \cdots, m;\ j = 1, \cdots, N$. We also compute the Monte Carlo mean squared prediction error of $\hat{\theta}_i$:

$$\text{MSPE}(\hat{\theta}_i) = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_i^{(j)} - \theta_i^{(j)})^2.$$

Figure (1) shows the simulated MSPE results. When $m = 10$, $\hat{\theta}_i(\hat{A}_{\mathrm{RE}})$ tends to have the smallest MSPE when the sampling variance is large ($D_i = 4$), and $\hat{\theta}_i(\hat{A}_{\mathrm{RE}})$ and $\hat{\theta}_i(\hat{A}_{\mathrm{AR}})$ outperform $\hat{\theta}_i(\hat{A}_{\mathrm{PR}})$ in the remaining groups. When m = 15, $\hat{\theta}_i(\hat{A}_{\mathrm{RE}})$ and $\hat{\theta}_i(\hat{A}_{\mathrm{AR}})$ perform similarly across all groups and better than $\hat{\theta}_i(\hat{A}_{\mathrm{PR}})$.

For interval estimation, we compare two traditional intervals of the form $\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\mathrm{mspe}}$ based on $\hat{A}_{\mathrm{PR}}$ and $\hat{A}_{\mathrm{RE}}$, and three parametric bootstrap intervals based on $\hat{A}_{\mathrm{PR}}$, $\hat{A}_{\mathrm{RE}}$, and $\hat{A}_{\mathrm{AR}}$. Derivations of $\mathrm{mspe}(\hat{\theta}_i)$ using $\hat{A}_{\mathrm{PR}}$ and $\hat{A}_{\mathrm{RE}}$ appear in Prasad and Rao (1990) and Datta and Lahiri (2000), respectively.

Tables 3 and 4 present the empirical coverage probabilities and average lengths for nominal 95% intervals. When $m = 10$, the parametric bootstrap method using $\hat{A}_{AR}$ (PB-AR) performs the best in terms of the coverage probabilities and the average lengths. The PR-based traditional interval (PR) and PB-PR show severe undercoverage across all groups. The traditional REML interval also undercovers, especially for group 1. PB-RE achieves good coverage but yields substantially longer intervals than PB-AR. This may be because the REML method sometimes produces zero estimates. Since the estimate $\hat{A}^*$ appears in the denominator of the term $\hat{\sigma}_1^{*-1}(\theta_i^* - \hat{\theta}_i^*)$ used in our parametric bootstrap method, this quantity becomes undefined whenever $\hat{A}^*_{\mathrm{RE}} = 0$. To address this issue, we replaced those zero estimates with 0.01. In such cases, the resulting values can be extremely large, which may in turn lead to overly wide prediction intervals. As m increases, all methods improve, although PR, RE, and PB-PR still exhibit undercoverage. Overall, PB-AR provides competitive cov-

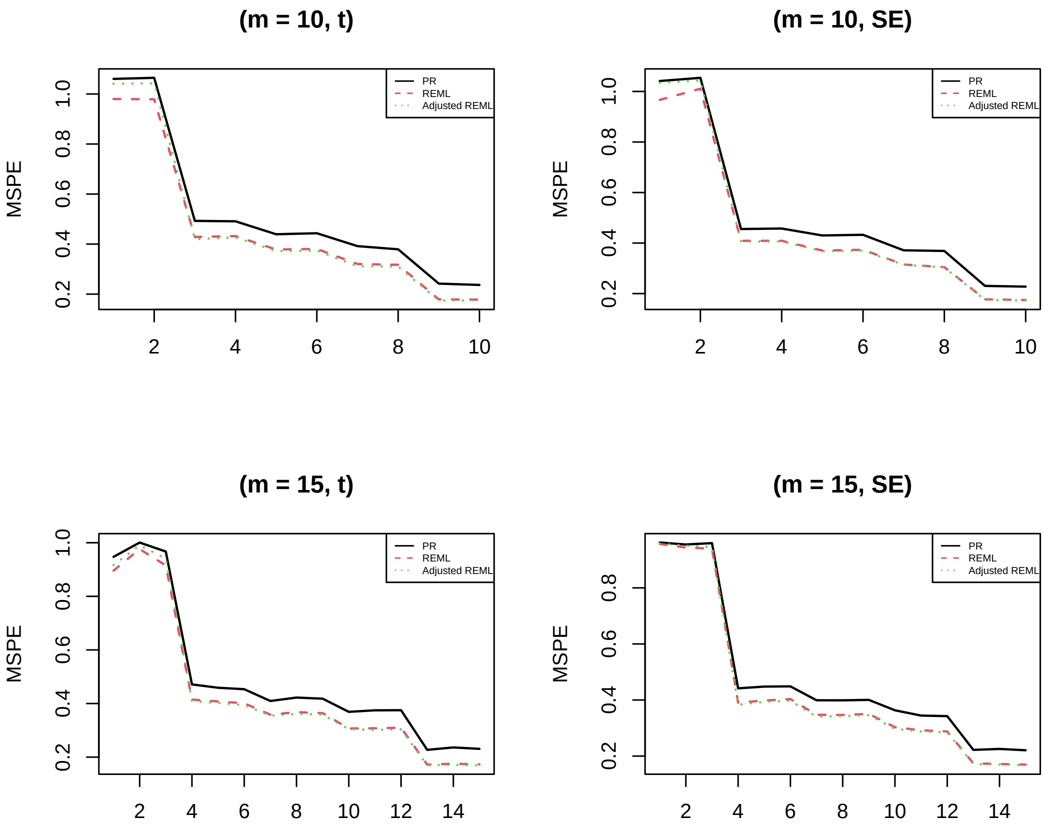


Figure 1: Simulated mean squared prediction error of $\hat{\theta}_i(\hat{A})$.

erage probabilities and interval lengths, showing only slight undercoverage for Group 1 under the shifted exponential distribution.

Table 3: Average Monte Carlo coverage and average length of different intervals for $m = 10$ with nominal coverage = 95% under t-distribution and shifted exponential distribution.

	PR	RE	PB-PR	PB-RE	PB-AR
			$t\ \{u_i\}$		
G1	84.10 (12.78)	88.70 (3.52)	83.90 (11.18)	96.25 (9.12)	95.10 (4.26)
G2	85.65 (11.22)	93.90 (2.53)	85.65 (6.74)	96.75 (5.39)	95.15 (2.56)
G3	85.65 (11.22)	93.90 (2.53)	85.65 (6.74)	96.75 (5.39)	95.15 (2.56)
G4	86.00 (10.76)	94.60 (2.24)	86.10 (5.69)	96.70 (4.42)	94.65 (2.19)
G5	86.25 (9.65)	96.30 (1.77)	86.60 (3.97)	97.25 (2.94)	94.70 (1.63)
			Shifted exponential $\{u_i\}$		
G1	84.20 (12.06)	87.85 (3.31)	84.10 (10.41)	94.55 (9.52)	94.85 (4.18)
G2	83.85 (10.52)	93.70 (2.44)	83.90 (6.19)	95.50 (5.41)	94.80 (2.48)
G3	83.85 (10.52)	93.70 (2.44)	83.90 (6.19)	95.50 (5.41)	94.80 (2.48)
G4	84.30 (10.14)	95.45 (2.20)	84.75 (5.26)	95.95 (4.42)	95.00 (2.14)
G5	87.05 (9.08)	96.80 (1.80)	87.35 (3.69)	96.00 (3.02)	94.70 (1.61)

Table 4: Average Monte Carlo coverage and average length of different intervals for $m = 15$ with nominal coverage = 95% under t-distribution and shifted exponential distribution.

	PR	RE	PB-PR	PB-RE	PB-AR
			$t\ \{u_i\}$		
G1	90.40 (10.76)	90.60 (3.52)	90.13 (9.52)	97.97 (6.46)	95.00 (4.07)
G2	90.67 (9.82)	93.50 (2.46)	90.70 (6.02)	97.60 (3.98)	95.17 (2.50)
G3	89.93 (9.74)	93.87 (2.33)	90.00 (5.65)	97.30 (3.69)	94.80 (2.34)
G4	90.37 (9.61)	93.73 (2.17)	90.20 (5.18)	97.13 (3.34)	94.43 (2.15)
G5	91.10 (8.99)	95.27 (1.68)	91.17 (3.75)	97.57 (2.35)	94.67 (1.62)
			Shifted exponential $\{u_i\}$		
G1	87.73 (10.63)	89.23 (3.40)	87.73 (9.36)	95.83 (7.14)	93.73 (4.09)
G2	89.13 (9.68)	93.60 (2.40)	89.20 (5.85)	96.50 (4.24)	94.97 (2.45)
G3	88.07 (9.56)	93.63 (2.28)	88.10 (5.47)	96.27 (3.94)	94.40 (2.29)
G4	89.03 (9.44)	94.03 (2.13)	88.67 (5.02)	95.97 (3.56)	94.60 (2.11)
G5	88.83 (8.85)	95.60 (1.69)	88.97 (3.64)	96.10 (2.51)	94.37 (1.60)

5 Discussion

This study provides empirical evidence on variance component estimation in general area-level models that allow non-normal random effects. The results indicate that the REML estimator can remain reasonably robust to deviations from normality, even when the number of areas is relatively small (for example, $m = 10$). Under both heavy-tailed and asymmetric random effect distributions, according to our simulation results, the bias of the REML estimator is similar to the PR estimator and its RMSE is smaller than both PR and adjusted REML estimators. Moreover, associated EBLUP based on REML estimate tends to perform well in prediction accuracy.

The simulation results also show that the effectiveness of parametric bootstrap prediction intervals depends heavily on the variance component estimator. When zero estimates are frequent, particularly when using the PR variance estimator, bootstrap intervals become unreliable due to the induced degeneracy. In contrast, the adjusted REML estimator reduces boundary estimates and supports stable bootstrap inference, leading to improved coverage across all simulation settings considered. This indicates that parametric bootstrap intervals based on adjusted REML estimates could be an effective alternative, when m is small.

There are promising directions for future work. For example, a deeper theoretical investigation of adjusted REML under non-normal random effects, including refined bias corrections and accurate MSPE estimation of EBLUP with adjust REML variance estimate, would strengthen its methodological foundations. Overall, the findings highlight that positive and stable estimation of variance components is essential for reliable small area prediction and inference. Adjusted likelihood methods offer a practical and robust alternative in applications where the normality assumption for random effects may not hold.

References

Bell, W. R. and E. T. Huang (2006). “Using the t-distribution to deal with outliers in small area estimation”. In: *Proceedings of statistics Canada symposium*.

Chatterjee, S., P. Lahiri, and H. Li (2008). “Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models”. In: *The Annals of Statistics* 36.3, pp. 1221 –1245. DOI: `10.1214/07-AOS512`. URL: `https://doi.org/10.1214/07-AOS512`.

Chen, Y., M. Y. Hirose, and P. Lahiri (2024). “Impact of existence and nonexistence of pivot on the coverage of empirical best linear prediction intervals for small areas”. In: *arXiv preprint arXiv:2410.11238*.

Cressie, N. (1990). *REML estimation in empirical Bayes smoothing of census undercount*. Iowa State University. Department of Statistics. Statistical Laboratory.

Datta, G. S. and P. Lahiri (2000). “A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems”. In: *Statistica Sinica*, pp. 613–627.

Fabrizi, E. and C. Trivisano (2010). “Robust linear mixed models for small area estimation”. In: *Journal of Statistical Planning and Inference* 140.2, pp. 433–443.

Fay, R. E. and R. A. Herriot (1979). “Estimates of income for small places: an application of James-Stein procedures to census data”. In: *Journal of the American Statistical Association* 74.366a, pp. 269–277.

Hawala, S. and P. Lahiri (2018). “Variance modeling for domains”. In: *Statistics and Applications* 16.1, pp. 399–409.

Hirose, M. Y. and P. Lahiri (2018). “Estimating variance of random effects to solve multiple problems simultaneously”. In: *The Annals of Statistics* 46.4, pp. 1721–1741.

Jiang, J. (1996). “REML estimation: asymptotic behavior and related topics”. In: *The Annals of Statistics* 24.1, pp. 255–286.

Jiang, J. and M. Torabi (2022). “Goodness-of-fit test with a robustness feature”. In: *TEST* 31.1, pp. 76–100.

Laird, N. M. and J. H. Ware (1982). “Random-effects models for longitudinal data”. In: *Biometrics*, pp. 963–974.

Li, H. (2007). *Small area estimation: An empirical best linear unbiased prediction approach*. University of Maryland, College Park.

Li, H. and P. Lahiri (2010). “An adjusted maximum likelihood method for solving small area estimation problems”. In: *Journal of multivariate analysis* 101.4, pp. 882–892.

Otto, M. C. and W. R. Bell (1995). “Sampling error modelling of poverty and income statistics for states”. In: *American Statistical Association, Proceedings of the Section on Government Statistics*, pp. 160–165.

Patterson, H. D. and R. Thompson (1971). “Recovery of inter-block information when block sizes are unequal”. In: *Biometrika* 58.3, pp. 545–554.

Prasad, N. N. and J. N. Rao (1990). “The estimation of the mean squared error of small-area estimators”. In: *Journal of the American statistical association* 85.409, pp. 163–171.

Rao, J. N. and I. Molina (2015). *Small area estimation*. John Wiley & Sons.

Yoshimori, M. and P. Lahiri (2014). “A second-order efficient empirical Bayes confidence interval”. In: *The Annals of Statistics* 42.4, pp. 1233 –1261. DOI: 10.1214/14-AOS1219. URL: https://doi.org/10.1214/14-AOS1219.

Robust Small Area Estimation: Methods, Theory, Applications, and Open Problems

Sylvia Harmening

Department of Statistics and Econometrics, Otto Friedrich University Bamberg, Germany, sylvia.harmening@uni-bamberg.de

Jiang, J., & Rao, J. S. (2025). *Robust Small Area Estimation: Methods, Theory, Applications, and Open Problems* (1st ed.). Chapman and Hall/CRC, doi: 10.1201/9781003395171, 257 pages.

Small area estimation (SAE) has become increasingly important in both research and practical applications. A small area refers either to a geographic area or to a subpopulation for which direct estimates are unreliable because of limited domain-sample sizes. SAE addresses this issue by borrowing strength - that is, by using auxiliary information through statistical modeling to improve estimation accuracy. However, this improvement involves a trade-off, which forms the central theme of the book. The authors address a crucial gap between classical SAE models and the growing demand for reliable estimation in the presence of model misspecification, outliers, and complex data structures. Intended for researchers and graduate students in statistics, data science, and related fields, the book also offers practical guidance for practitioners, including those working in government and public sector organizations.

The book comprises seven well-structured chapters that progress logically from fundamental concepts to cutting-edge developments. Each chapter reinforces its theoretical discussions with illustrative examples, simulation experiments, or case studies based on real data.

Chapter 1 ‘Small Area Estimation: A Brief Overview’ presents the motivation and fundamental ideas behind SAE. It outlines key estimation strategies, including direct and indirect estimation methods such as the Fay-Herriot model and the nested-error regression (NER) model, and provides an overview of available software packages.

Chapter 2 ‘SAE Methods Built on Weaker Assumptions’ examines how SAE methods and mean squared prediction error (MSPE) estimation can be developed with fewer or less restrictive statistical assumptions, making them more robust to model misspecification. It introduces techniques such as the robust empirical Bayes estimator, the regression average, non-Gaussian mixed models and heteroscedastic NER models. The chapter also includes simulation studies and practical examples, such as grape production and income data.

Chapter 3 ‘Outlier Robustness’ explores methods that make SAE robust to outliers. Robust techniques like the robust EBLUP, M-quantile regression, and density power divergence, which reduce outlier influence while keeping efficiency, are introduced. Detecting and adjusting for outliers to improve prediction accuracy are also covered.

Chapter 4 ‘Observed Best Prediction’ introduces a method designed to make SAE more robust to model misspecification. Unlike traditional EBLUP, observed best prediction (OBP) estimates model parameters by minimizing the observed mean squared prediction error, giving more weight to areas with high sampling variance leading to predictions that remain reliable even when the assumed model is partly wrong. Moreover, the observed best selective predictor (OBSP), which combines variable selection and parameter estimation, and the compromised best predictor (CBP), which can be described as a weighted average of the EBLUP and the OBP, are explored.

Chapter 5 'More Flexible Models' addresses advanced SAE methods that use semi-parametric, non-parametric, and functional models to reduce dependence on parametric assumptions. These models employ tools like splines, kernel functions, and functional mixed-effects models to capture complex, nonlinear relationships between variables. By allowing model flexibility, they improve robustness against model misspecification and perform better even for time-series data.

Chapter 6 'Model Selection and Diagnostics' discusses how to choose and validate models in SAE. It reviews classical and modern selection tools such as information criteria, fence methods, and shrinkage selection. The chapter also explains diagnostic techniques like (robust) goodness-of-fit tests and the tailoring method to detect violations of the model assumptions.

Chapter 7 'Other Topics' discusses several additional topics connected to robust SAE. It covers benchmarking, Bayesian, and machine learning methods (like mixed-effects random forests, neural networks, and gradient boosting), as well as approaches for handling missing data and classified mixed model prediction. The chapter concludes by discussing new challenges - such as Big Data, data quality, and privacy protection (differential privacy) - and calls for future SAE methods that remain robust in modern data environments.

The book is an impressive and timely contribution to the literature on SAE, skillfully combining a deep theoretical framework, modern methodological advances, and practical insights for real-world applications. A notable strength of the book lies in its integration of theory and applications. The authors devote substantial attention to real-world examples, including data on income and poverty, agricultural yields, and health indicators. Although the book is mainly aimed at researchers and graduate students, this hands-on approach is particularly valuable for practitioners, helping them understand and adopt robust SAE methods with greater ease.

R2BEAT: An R Package for Performing Optimal Allocation and Sample Selection

Giulio Barcaroli[1], **Ilaria Bombelli**[2], **Andrea Fasulo**[3], **Alessio Guandalini**[4], **and Marco D. Terribili**[5]

[1]Independent consultant, Italy, giulio.barcaroli@gmail.com
[2]Italian Institute of Statistics (Istat), Italy, ilaria.bombelli@istat.it
[3]Italian Institute of Statistics (Istat), Italy, andrea.fasulo@istat.it
[4]Italian Institute of Statistics (Istat), Italy, alessio.guandalini@istat.it
[5]Italian Institute of Statistics (Istat), Italy, marcodionisio.terribili@istat.it

Abstract

The review presents the main features of the **R2BEAT** `R` package (Fasulo et al., 2023), which is designed for optimal sample allocation. The package integrates the Bethel (1989) algorithm, which extends optimal allocation (Tschprow, 1923; Neyman, 1934) to the multi-domain and multi-purpose case, and it also implements the extension proposed by Falorsi et al. (1998) for handling complex sampling designs. The package streamlines the entire sample design workflow, from sample optimisation to selection of sampling units.

Keywords: sampling, stratification, two-stage, design effect.

1 Introduction

Sample surveys conducted by National Statistical Institutes (NSIs) and other organisations often pursue multi-domain and multi-purpose objectives. Consequently, they are required to produce accurate estimates for multiple parameters and across various domains, both geographical and non-geographical.

Since surveys are subject to budgetary and logistical constraints, their design must be carefully planned to ensure high-quality estimates for the parameters of interest. Within this framework, several crucial decisions must be made, including determining the sample size, defining the stratification scheme, and allocating sampling units across strata and stages.

The proposed package, **R2BEAT** (standing for *R “to” Bethel Extended Allocation for Two-stage*), has been developed within this context (Barcaroli et al., 2023). It integrates the Bethel (1989) algorithm, which extends optimal allocation (Tschprow, 1923; Neyman, 1934) to the multi-domain and multi-purpose case, and it also implements the extension proposed by Falorsi et al. (1998) for handling complex sampling designs. Therefore, it fills an existing gap in the range of statistical software dedicated to sample size allocation, providing an advanced and flexible tool for the R community.

The paper is organised as follows. Section 2 describes the structure of the package and the case study used to illustrate its functionality. Section 3 explains the workflow for stratified sampling design - very common in economic surveys - while Section 4 focuses on two-stage sampling design with stratification of the primary stage units - widely used in household surveys. Finally, Section 5 provides conclusions.

For additional details and overview of further functions, readers may refer to the companion paper by Barcaroli et al. (2023). The workflow presented is based on the most recent functions available on the GitHub page of the package which are expected to be included in the forthcoming official release of the package.

2 Preliminarities

2.1 Structure of the package

The **R2BEAT** package provides a comprehensive set of functions for designing and selecting samples through optimal allocation, both for stratified and two-stage with stratification of the primary stage units.

The appropriate sampling design to apply in a specific situation depends on the information available in the sampling frame, for example, for implementing stratification or an unequal-inclusion-probability sampling design. In addition, to perform optimal allocation, information on the target variable(s) or at least on a suitable proxy is required. Such information may be obtained from a sampling frame, such as a register, or from a sample survey, either a concurrent survey or a previous wave of the same survey, and can be used to guide the allocation of the sampling units.

R2BEAT is able to manage both the scenarios and the typical workflow for designing and selecting a sample involves three main steps: (1) preparing the input data, (2) defining the sampling design and computing the optimal allocation, and (3) selecting the final sample units.

To install the latest stable version of `R2BEAT` from CRAN, use the command `install.packages("R2BEAT")` within the `R` environment. The most recent development version is available on GitHub and can be installed by executing `devtools::install_github("barcaroli/R2BEAT_2.0")`.

2.2 Case study

In this paper, we develop the workflow under a stratified sampling design and a two-stage sampling design with stratification of the primary sampling units[1]. In both cases, a sampling frame covering the entire population of interest is required, whereas information on the target variable(s) is obtained from a previously conducted sample survey. The case in which such information is available directly on the sampling frame differs slightly from this setting, and readers are referred to Barcaroli et al. (2023).

The sampling frame considered in this paper, `pop.RData`, refers to a population of 2,258,507 individuals and contains the following variables:

```
'data.frame':   2258507 obs. of  13 variables:
$ id_ind       : int  1 2 3 4 5 6 7 8 9 10 ...
$ id_hh        : Factor w/ 963018 levels "H1","H10","H100",..: 1 1 1 2 3 3 3 3 ...
$ municipality : num  1 1 1 1 1 1 1 1 1 1 ...
$ province     : Factor w/ 6 levels "north_1","north_2",..: 1 1 1 1 1 1 1 1 1 1 ...
$ region       : Factor w/ 3 levels "north","center",..: 1 1 1 1 1 1 1 1 1 1 ...
$ sex          : int  1 2 1 2 1 1 2 2 1 1 ...
$ cl_age       : Factor w/ 8 levels "(0,14]","(14,24]",..: 3 7 8 5 4 6 6 4 4 1 ...
$ active       : num  1 1 0 1 1 1 1 1 1 0 ...
$ unemployed   : num  0 0 0 0 0 0 0 0 0 0 ...
$ inactive     : num  0 0 1 0 0 0 0 0 0 1 ...
$ income_hh       : num  30488 30488 30488 21756 29871 ...
```

[1]To reproduce the analyses presented in these examples, all datasets are available for download at https://github.com/barcaroli/R2BEAT_datahttps://github.com/barcaroli/R2BEAT_data.

In particular, it contains

- `id_ind`: individual identifier,
- `id_hh`: household identifier to which the individual belongs,
- `municipality`: municipality identifier in which the individual lives,
- `province`: province (NUTS3) identifier in which the individual lives,
- `region`: region (NUTS2) identifier in which the individual lives;

demographic information:

- `sex`: sex of the individual,
- `cl_age`: age class of the individual in ten-year intervals;

information on target variables:

- `active`: binary indicator for occupational status “active”,
- `inactive`: binary indicator for occupational status “inactive”,
- `unemployed`: binary indicator for occupational status “unemployed”,
- `income_hh`: income.

Furthermore, for the present purpose, sampling data from a previous survey are also considered, `sample.RData`, comprising a two-stage (municipalities and individuals) sample of 9,421 units drawn from `Pop.RData`. This dataset includes the same variables described above and, in addition, the following variables useful for the present purpose:

- `weight`: the sampling weights assigned to each sampling unit,
- `stratum_2`: the strata used for stratifying the municipalities,
- `SR`: binary indicator for the Self-Representative (SR) municipalities. It is equal to 1 for the municipalities that are included certainly in the sample (inclusion probability equal to 1), 0 otherwise.

3 Stratified sampling design

Stratification of the sample is very common and highly effective. When one or more variables correlated with the survey’s target variables are available in the sampling frame, it is possible to partition the sampling units into strata and select an independent sample from each of them in order to obtain more efficient estimates.

Defining the proper number of sampling units to be collected in each stratum is an allocation problem. The optimal allocation (Tschprow, 1923; Neyman, 1934) assigns a larger portion of the sample to strata with greater population size and, in particular, to those characterised by higher variability of the target variable. In such strata, a greater sample size is required to achieve the desired level of efficiency of the estimates.

The allocation problem in the multivariate and multi-domain case can be formulated as an optimisation problem (Bethel, 1989), where the objective is to minimise the cost of the survey, usually expressed in terms of sample size, subject to a set of precision constraints on the estimates.

3.1 Step 1: Input preparation

From this premise, it follows straightforwardly that the inputs required to perform optimal allocation are strata information and a set of precision constraints for the estimates of the target variables.

The strata information can be obtained using the function `prepareInputToAllocation_beat.1st`. The parameters to be specified in the function are:

- `frame`: the sampling frame containing necessarily the identifier of the units, strata and domain variables and optionally the target variable(s).
- `sample` (optional): sample survey data, containing necessarily the strata and domains variables, the target variable(s) and the sampling weights. In this way, statistical summaries of the target variables will be estimated on the sample. Strata and domains variables must be consistent with those defined for `samp_frame` dataframe. Default is NULL, meaning that just sampling frame data are used,
- `ID`: name of the identifier of the units in the sampling frame.
- `stratum`: either name of the variable in `samp_frame`, which is taken as the stratum, or the name of the variables which have to be concatenated to obtain the stratum. In the latter case, the variables used to build the stratum are retained.
- `domain`: name of the variable(s) identifying the domain(s) for which estimates of the target variables must be disseminated. Domain(s) must be aggregation of the strata.
- `target`: names of the variable(s) in the sampling frame identifying the target variable(s) leading the planning of the survey.
- `weight` (optional): the sampling weights, whether the target variable(s) is (are) available only on sample data. The default is NULL, meaning that the target variables are available in the sampling frame and, therefore, the statistical summaries are computed on it.

Suppose that the sample is to be stratified by province, while the estimates of mean income and the incidence of unemployed individuals are to be controlled at the regional level. The parameter in `prepareInputToAllocation_beat.1st` can be set as follows:

```
input1 <- prepareInputToAllocation_beat.1st(frame=pop,
                                            sample=samp,
                                            ID="id_ind",
                                            stratum=c("province"),
                                            domain="region",
                                            target=c("income_hh","unemployed"),
                                            weights="d")
```

The function returns three objects:

1. `file_strata`: a dataframe of strata in which the population size, the mean (`M1`, `M2`, ...) and the standard deviation of the target variables in the population (`S1`, `S2`, ...) is provided. Furthermore, one column is specified for each domain. The global domain is included by default and is named `DOM1`. Two additional columns are filled automatically in: `CENS`, an identifier whether the stratum must be censused or not (the default is equal to 0 for all of them) and `COST` indicating the cost of the each interview in the stratum (the default is equal to 1 for all of them).

```
'data.frame': 6 obs. of  11 variables:
 $ STRATUM : Factor w/ 6 levels "north_1","north_2",..: 3 4 1 2 5 6
 $ province: Factor w/ 6 levels "north_1","north_2",..: 3 4 1 2 5 6
 $ DOM1    : Factor w/ 1 level "Total": 1 1 1 1 1 1
 $ DOM2    : Factor w/ 3 levels "north","center",..: 2 2 1 1 3 3
 $ N       : num  591517 205173 462420 482122 336608 ...
 $ M1      : num  21673 21054 27930 24332 16923 ...
 $ M2      : num  0.1032 0.1547 0.0212 0.0316 0.2922 ...
```

```
 $ S1      : num  19618 16565 26151 19252 14686 ...
 $ S2      : num  0.304 0.362 0.144 0.175 0.455 ...
 $ CENS    : num  0 0 0 0 0 0
 $ COST    : num  1 1 1 1 1 1
```

2. `var_list`: a vector of target variables as they appear in `file_strata` (i.e. `M1`, `M2`, …or `S1`, `S2`, …).
3. `ID_stratum`: a dataframe reporting the stratum to which each unit in the `samp_frame` belongs.

Then, the precision constraints in terms of coefficient of variation (CV) for each target variable in each domain have to be planned. Assume that the maximum acceptable coefficients of variation are 2% at the national level and 5% at the regional level for the mean of income, while 5% at the national level and 7% at the regional level for the incidence of unemployment. Then:

```
cv1 <- data.frame(DOM=c("DOM1", "DOM2"),
                  CV1=c(0.02, 0.05),
                  CV2=c(0.05, 0.07))
cv1
  DOM  CV1  CV2
1 DOM1 0.02 0.05
2 DOM2 0.05 0.07
```

3.2 Step 2: Optimal allocation

The optimal allocation is then computed using the `beat.1st` function and the inputs `file_strata` and `cv1` previously described:

```
alloc1 <- beat.1st(file_strata=input1$file_strata,
                   errors=cv1)
```

The final sample size of the optimal resulting allocation satisfying the precision constraints is 9,688. The object `alloc1` is a list of seven output objects:

1. `n`: a vector containing the optimal allocation for each stratum. Its total, `sum(alloc1$n)`, is equal to 9,688.
2. `file_strata`: the input dataset `file_strata` with an additional column, `n`, indicating the optimal allocation.
3. `alloc`: a dataframe specifying for each stratum the optimal allocation (`OPT`), the proportional allocation (`PROP`), and the uniform allocation (`UNIF`).
4. `sensitivity`: a dataframe with the precision constraints (`PlannedCV`), the expected CV (`ExpectedCV`), i.e. the CV that are expected to be obtained the optimal allocation), and the sensitivity (`Sensitivity 10%`) for each variable and each domain category. Sensitivity provides a suggestion about the expected variation in sample size if the planned errors change by 10%.
5. `ExpectedCV`: a dataframe with the maximum of the expected coefficients of variation (`Actual CV`), for each variable in each domain.
6. `PlannedCV`: a dataframe with the maximum coefficients of variation admissible for each domain and for each variable. It is the input errors dataframe, provided by the user.
7. `param_alloc`: a vector summarising all the parameters used for performing the optimal allocation.

In general, a reduction in the CV corresponds to a higher required level of precision, which in turn requires a larger sample size, and vice versa.

3.3 Step 3: Selection of sampling units

Given the allocation, the sample can be selected using the function `strata` from the `R` package **sampling** (Tillé and Matei, 2023). For proper implementation, prior to sample selection, it is recommended that both the sampling frame and the allocation dataframe are ordered by stratum:

```
library(sampling)
alloc1$file_strata <- alloc1$file_strata[order(alloc1$file_strata$STRATUM),]
pop <- merge(pop,input1$ID_stratum,by="id_ind")
pop <- pop[order(pop$STRATUM),]
s <- strata(data=pop,
            stratanames="STRATUM",
            size=alloc1$file_strata$n,
            method="srswor")
sample_str <- getdata(data=pop, m=s)
```

Finally, `sample_str` is the sample of size 9,688, stratified by province, which will yield estimates of mean income and the incidence of unemployment consistent with the planned precision constraints defined in `cv1`.

4 Two-stage sample design with stratification of the primary stage units

Sampling units may be organised in clusters; for example, individuals within households, workers within enterprises, or households within enumeration areas or municipalities.

For logistical and economic reasons, it may be useful to exploit this clustering. The typical case is household surveys. In these surveys, municipalities (Primary Stage Units, PSUs) are usually stratified. Then, within each stratum, a sample of municipalities is selected, typically with probability proportional to size, and within the selected municipalities a sample of households (Secondary Stage Units, SSUs) is drawn.

This sampling design is more convenient because it reduces the management complexity and therefore costs. However, this advantage comes at the expense of a reduction in the efficiency of the sample design, which must be taken into account when planning the sample.

In this context, the allocation problem is more complex, since the both PSUs and SSUs must be allocated. A solution can be obtained by following Falorsi et al. (1998). They propose iterating the Bethel algorithm, adjusting the design effect[2] at each iteration. Convergence is usually achieved within 5–6 iterations.

4.1 Step 1: Input preparation

The function `prepareInputToAllocation2` behaves similarly to the function described in Section 3.1 and likewise generates all the input objects required for the optimal allocation.

However, since this sample design is more complex, it needs more parameters:

[2] It denotes how much the sampling variance under the adopted sampling design is inflated with respect to SRS, on equal sample size. The design effect for SRS is equal to 1, whereas for clustered sampling designs it is greater than 1.

- `frame`: the sampling frame containing necessarily the identifier of the units, strata and domain variables and optionally the target variable(s).
- `RGdes`: the sampling data containing necessarily the strata and domains variables, the target variable(s) and the sampling weights. It must be a design object created with the `R` package **ReGenesees**[3].
- `RGcal` (optional): the sampling data containing necessarily the strata and domains variables, the target variable(s) and the sampling calibrated weights. It must be a calibration object created with the `R` package **ReGenesees**[4]. If NULL (default), it is set equal to the design object, `RGdes`.
- `id_PSU`: name of the identifier of the PSU.
- `id_SSU`: name of the identifier of the SSU.
- `stratum`: name of the variable in the sampling frame which is taken as the stratum. In contrast to the function used for the stratified sampling design, in the current version, this function does not perform variable concatenation. Therefore, it is recommended to prepare the concatenated variables beforehand.
- `target`: names of the variable(s) in the sampling frame identifying the target variable(s) leading the planning of the survey.
- `deff_level`: name of the variable(s) identifying the domain level at which compute the design effect. Although this information is applied in the algorithm at the stratum level, it is advisable to aggregate it to a higher hierarchical level to obtain more stable design effect estimates. The resulting design effect value is then applied to all strata belonging to the corresponding higher-level domains.
- `domain`: name of the variable(s) identifying the domain(s) for which estimates of the target variables must be disseminated. Domain(s) must be aggregation of the strata.
- `delta`: the average size of SSUs in terms of elementary units in each stratum. If SSUs match the survey units, `delta` must be equal to 1 in all the strata. Otherwise, it should be set equal to the average size of SSUs in terms of elementary units in the stratum.
- `minimum`: minimum number of SSUs to be interviewed in each selected PSU.

Suppose that the sample is to be a two-stage, municipalities and individuals, in which the municipalities are stratified by province, while the estimates of mean income and the incidence of unemployed individuals are to be controlled at the regional level and have been previously investigated in another sample survey. Furthermore, the design effect will be computed at the regional level.

A propedeutic step, before preparing the inputs for the optimal allocation, is to create the design object useful for computing the design effects and the estimator effects. The package **ReGenesees** is used for the present purpose also in the `prepareInputToAllocation2` function.

Since `samp` is a two-stage (municipalities, `municipality`, and individuals, `id_ind`) sample with stratification of the municipalities (`stratum_2`), the design object, `RGdes`, is defined as follows:

```
library(ReGenesees)
samp$stratum_2 <- as.factor(samp$stratum_2)
RGdes <- e.svydesign(data=samp,
                     ids=~municipality+id_ind,
                     strata=~stratum_2,
                     weights=~weight,
                     self.rep.str=~SR,
```

[3]For all the details see Zardetto (2015) and Zardetto (2023).
[4]See above.

```
                         check.data=TRUE)
```

Then, the parameter in `prepareInputToAllocation2` can be set as follows:

```
input2 <- prepareInputToAllocation2(frame=pop,
                                    RGdes=RGdes,
                                    id_PSU="municipality",
                                    id_SSU="id_ind",
                                    stratum="province",
                                    target=c("income_hh", "unemployed"),
                                    deff_level="region",
                                    domain="region",
                                    delta=1,
                                    minimum=120)
```

The function returns six dataframes:

1. `strata`: a dataframe with the same structure as the output provided by the function `prepareInputToAllocation_beat.1st` described in Section 3.1.
2. `deff`: a dataframe of strata with the design effect for each variable (`DEFF1`, `DEFF2`, ...) and the average size of the SSUs in the PSUs (`b_nar`).
3. `effst`: a dataframe with the estimator effect for each variable in each stratum[5]. When `RGcal` is NULL the estimator effect is equal to 1 for all the variables in each stratum.
4. `rho`: a dataframe with the intraclass correlation coefficient[6] for each variable in each stratum and for municipalities included for sure in the sample (Self-Representative). The correlation coefficient for larger municipalities (i.e. included certainly in the sample, since their selection probability is equal to 1) is equal to 1 by default.
5. `psu_file`: a dataframe of PSUs with the related stratum and their size (`PSU_MOS`) .
6. `des_file`: a dataframe of strata with the size, delta and minimum[7].

All of these objects, except `deff` (included for documentation purposes only), serve as inputs for the optimal allocation step.

Then, as before, the precision constraints in terms of coefficient of variation (CV) for each target variable in each domain have to be planned. Assume, in this case, that the maximum acceptable coefficients of variation is 2% at the national level and 5% at the regional level for the mean of income, while 5% at the national level and 7% at the regional level for the incidence of unemployment. Then:

```
cv2 <- data.frame(DOM=c("DOM1", "DOM2"),
                  CV1=c(0.02, 0.05),
                  CV2=c(0.05, 0.07))
```

[5]The estimator effect measures how much the sampling variance under the chosen estimator is inflated or deflated relative to the Horvitz–Thompson estimator (Horvitz and Thompson, 1952), under the same sample design. By definition, the Horvitz–Thompson estimator has an estimator effect equal to 1, while for instance a calibrated estimator (Deville and Särndal, 1992) typically yields values lower than 1.

[6]The correlation coefficient captures the degree of similarity among units within clusters. Positive values indicate strong within-cluster similarity, leading to higher design effects and poorer CVs. In contrast, negative values reflect greater within-cluster heterogeneity.

[7]By modifying this dataframe, it is possible to set different minimum values according to the strata.

```
cv2
  DOM  CV1  CV2
1 DOM1 0.02 0.05
2 DOM2 0.05 0.07
```

4.2 Step 2: Optimal allocation

The optimal allocation is then computed using the `beat.2st` function and the inputs previously described:

```
alloc2 <- beat.2st(file_strata=input2$file_strata,
                   errors=cv2,
                   des_file=input2$des_file,
                   psu_file=input2$psu_file,
                   rho=input2$rho,
                   effst=input2$effst)
```

```
  iterations PSU_SR PSU NSR PSU Total   SSU
1          0      0       0         0  9688
2          1     17      42        59 12677
3          2     19      76        95 11962
4          3     20      68        88 11944
```

The final sample size of the optimal resulting allocation satisfying the precision constraints comprises 88 PSUs, 20 Self-Representative (`PSU_SR`) and 68 Non-Self-Representative (`PSU_NSR`), and 11,944 SSUs.

The object `alloc2` is a list of eight output objects:

1. `iteractions`: a dataframe that, for each iteration, provides a summary of the number of PSUs (`PSU_Total`), distinguishing between Self-Representative (`PSU_SR`) and Non-Self-Representative (`PSU_NSR`) units, as well as the number of SSUs (`SSU`). This output is also printed to the screen.
2. `file_strata`: a dataframe equal to the input dataframe `file_strata` with additional columns: `DEFT1`, `DEFT2`, …reporting the square root of the design effect for each variable within each stratum, and `n`, specifying the optimal allocation.
3. `alloc`: a dataframe with optimal (ALLOC), proportional (PROP), equal (EQUAL) sample size allocation.
4. `planned`: a dataframe with the precision constraints (`Planned CV`) for each variable in each domain.
5. `expected`: a dataframe with the expected CVs with the given optimal allocation (`Expected CV`) for each variable in each domain.
6. `sensitivity`: a dataframe with a summary of the sensitivity at 10% for each domain and each variable.
7. `deft_c`: a dataframe with the design effect for each variable in each domain in each iteration. Note that `DEFT1_0`, `DEFT2_0`, …is always equal to 1 if `deft_start` is `NULL`. Otherwise is equal to `deft_start`. While `DEFT1`, `DEFT2`, …are the square root of the final design effect related to the given allocation reported also in `file_strata`.
8. `param_alloc`: a vector with a resume of all the parameter given for the allocation.

As before, a reduction in the CV corresponds to a higher required level of precision, which in turn requires a larger sample size, and vice versa. Moreover, for a fixed sample size, reducing the minimum number of units per PSU decreases the CV, since the sample is spread across more PSUs and the design effect decreases. Conversely, increasing the minimum leads to higher CVs.

4.3 Step 3: Selection of sampling units

The PSUs are then selected using:

```
sample_1st <- select_PSU(alloc=alloc2, type="OPT", pps=TRUE)
```

The selected PSUs are stored in the `sample_PSU` element of the output list. Using these, the final sample of secondary units can be selected:

```
PSU_sampled <- sample_1st$sample_PSU
sample_2st <- select_SSU(df = pop,
                         PSU_code ="municipality",
                         SSU_code ="id_ind",
                         PSU_sampled=PSU_sampled)
```

Finally, `sample_2st` is the two-stage sample of size 13,090, with the municipalities stratified by province, which will yield estimates of mean income and the incidence of unemployment consistent with the planned precision constraints defined in `cv2`.

A slight discrepancy may arise between the number of SSUs determined during allocation and those obtained after PSU selection. This occurs because the PSU selection process enforces the minimum number of SSUs (here, 120) per selected PSU, which may result in an increase in the total number of SSUs.

The two samples, `sample_str` and `sample_2st`, achieve the same level of precision, in terms of CVs, for the estimates of average income and unemployment incidence at both national and regional level. However, in the two-stage design the sampling units are clustered, which reduces the efficiency of the sample. As a result, a larger sample size, 13,090 instead of 9,688, is required to satisfy the same precision constraints.

5 Concluding remarks

R2BEAT stands out for its comprehensive approach to statistical data production, covering all stages from design to sample selection. It is especially flexible and adaptable, offering optimal allocation for both stratified and two-stage with stratification of the primary stage units sampling designs. This makes it valuable for various organisations, including national statistical institutes (NSIs), private research firms, research institutes and universities.

R2BEAT leverages auxiliary variables, improving sample design and allocation by making use of additional data from registers or previous surveys. Its user-friendly output allows for easy analysis and validation of the allocations and sample used in the survey. As the package stems from an ongoing development effort, it is continuously updated and maintained to guarantee maximum consistency and efficiency in the implementation of the methodology for sample design and selection.

References

Barcaroli, G., A. Fasulo, A. Guandalini, and M. Terribili (2023). "Two Stage Sampling Design and Sample Selection with the R Package R2BEAT". In: *The R Journal* 15.3, pp. 191–213.

Bethel, J. W. (1989). "Sample allocation in multivariate surveys". In: *Survey methodology* 15.1, pp. 47–57.

Deville, J.-C. and C.-E. Särndal (1992). "Calibration Estimators in Survey Sampling". In: *Journal of the American Statistical Association* 87.418, pp. 376–382. DOI: `10.1080/01621459.1992.10475217`.

Falorsi, P. D., M. Ballin, C. De Vitiis, and G. Scepi (1998). "Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'ISTAT". In: *Statistica Applicata* 10.2, pp. 235–257.

Fasulo, A., G. Barcaroli, S. Falorsi, A. Guandalini, D. Pagliuca, and M. D. Terribili (2023). *R2BEAT: Multistage Sampling Allocation and Sample Selection*. R package version 1.0.5. URL: `https://CRAN.R-project.org/package=R2BEAT`.

Horvitz, D. G. and D. J. Thompson (1952). "A Generalization of Sampling Without Replacement from a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685.

Neyman, J (1934). "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". In: *Journal of the Royal Statistical Society* 97.4, pp. 558–625.

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

Tschprow, A. (1923). "On the mathematical expectation of the moments of frequency distributions in the case of correlated observations". In: *Metron* 2, pp. 646–683.

Zardetto, D. (2015). "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys". In: *Journal of Official Statistics* 31.2, pp. 177–203. DOI: `10.1515/jos-2015-0013`. URL: `https://doi.org/10.1515/jos-2015-0013`.

— (2023). *ReGenesees: R Evolved Generalised Software for Sampling Estimates and Errors in Surveys (R package)*. Version 2.3. R package. URL: `https://github.com/DiegoZardetto/ReGenesees`.

Argentina

Reporting: **Verónica Beritich**

INDEC hosted officials from the United Nations System in Argentina

On September 11th, the authorities of INDEC, together with the authorities of the Ministry of Foreign Affairs, International Trade and Worship of Argentina, took part in a meeting with representatives of 14 organizations of the United Nations System in Argentina. The purpose of this meeting was to strengthen data production in the country. The institutional work plan was presented, highlighting the use of official statistics to promote economic, social, and environmental development, and inviting international agencies to work in coordination with INDEC on projects involving the National Statistical System.

The use of administrative records is one of the structural pillars of the projects shared in this meeting, and INDEC's obligation is to harness all that potential to produce better data. Other key pillars highlighted by INDEC's authorities included the crucial support of the United Nations agencies in Argentina in breaking down the natural barriers that exist within the State, and the need to work in a coordinated manner to provide information that meets the growing demand and thus deliver timely, high-quality statistics for decision-making.

The working session was divided into two parts. At the beginning, representatives of each multilateral organization briefly described the projects they are promoting, which are primarily based on information produced by INDEC. Afterwards, the main current lines of work of the Institute were presented:

- Strengthening the statistical structure through projects that include the continuous updating of population projections and the implementation of the new Master Urban Household Sample of the Argentine Republic (MMUVRA).
- Expanding the conceptual frameworks for the production of new economic, social, and environmental statistics and incorporating the integrated governance system.
- Developing an integrated system of administrative records for statistical use, based on statistical records of population, housing, and economic units.
- Applying technological innovation in statistical operations and in the harmonization of administrative records.
- Strengthening institutions by coordinating activities between national State agencies that provide information and produce statistics and international organizations in the field.

General information can be found at https://www.indec.gob.ar.

For further information, please contact https://www.indec.gob.ar/indec/web/Institucional-Indec-Contacto.

Australia

Reporting: **Paul Schubert**

Enhancing household survey frames when you don't have a population register

National Statistical Offices (NSOs) across the globe have an urgent need to address increasing collection costs, caused by the increased difficulty in making contact with households and securing their cooperation, while maintaining data quality. The biggest threat to data quality that all NSOs are working to safeguard against is nonresponse bias given the increasing trends of survey nonresponse.

Unlike Scandinavian countries (for example), Australia does not have a population register; frame information to assess and help adjust for nonresponse biases in household surveys has been very limited.

The goal of the Australian Bureau of Statistics' (ABS) Data Improved Frames From the Address Register (DIFFAR) project is to address these challenges head-on by delivering improved survey frames for household surveys. The improvement comes from augmenting existing survey frames - a full population list of addresses eligible for selection - with categorical auxiliary information derived from the Person Level Integrated Data Asset (PLIDA) that can support more efficient and targeted survey designs, operations and processes (for more about PLIDA, see: Person Level Integrated Data Asset (PLIDA) | Australian Bureau of Statistics.)

The key methodological innovation of the DIFFAR project is the use of a random forest model, a machine learning approach, to predict the probability that an address has certain characteristics based on its reported PLIDA data – for example, whether a low-income household resides at the address or a child is present at the address. These predicted probabilities are then used to group addresses into decile categories which can be used to:

- ensure representation of key subpopulations in the sample design and selection
- track response during data collection operations to prioritise groups for follow-up
- improve estimation results through this use of auxiliary information.

It is important to note that this method protects the privacy of personal information from PLIDA by not using it directly to categorise addresses.

While with DIFFAR the ABS is still in the early days of introducing more and better use of auxiliary information in our survey processes, it has already enabled us to reduce biases in survey estimates, and improve efficiency of sample designs by up to 20%. It has also identified subpopulations that are under-represented in response so far during enumeration, allowing data collection staff to prioritise these subpopulations in follow-up, and delivering a more representative final sample. The use of DIFFAR information in estimation has provided new insights into subpopulations that are currently under-represented in weighted estimates, and is providing the information required to remove the resulting nonresponse bias from estimates.

For further information, please contact Bruce Fraser.

Brazil

Reporting: **Andrea Diniz da Silva**

Learning Paths: Professional Training and Statistical Literacy

The Brazilian Institute of Geography and Statistics has created a set of Learning Paths for various audiences: professionals from the public and private sectors, academics, policy makers, and other professionals interested in the topics offered. The courses will be offered free of charge, remotely, with synchronous and asynchronous activities. To ensure regional reach, spots are reserved for each region of Brazil. Spots are also distributed equally between men and women. This aims to ensure nationwide reach and a diverse range of trainees. Participants are expected to disseminate the knowledge, references, and materials provided in each course.

There are six Learning Paths: Current Situation Analysis and Public Communication; Improving Municipal Planning: Regulations and Indicators; Data Science, Big Data, and Artificial Intelligence; Demographic Census and Municipal Planning; Statistics, Territory, and Public Policies; and the National System of Statistics and Geography. The courses in these Paths are mostly taught by IBGE staff, but also include guest professors.

All Paths focus on statistics, their production, use, and understanding, thus fulfilling a dual role: professional training and statistical literacy.

Canada

Reporting: **Darren Gray**

Statistics Canada's generalized systems are migrating to R and Python

Statistics Canada's generalized systems have supported our statistical production infrastructure for many years, providing efficient, vetted, and reusable solutions that can be deployed across multiple programs. These systems are used throughout the Integrated Business Statistics Program (IBSP), numerous economic and social surveys, and in our Census program. They are also shared internationally at no cost, supporting methodological consistency and cooperation among National Statistical Organizations (NSOs).

Statistics Canada is modernizing its broader statistical systems through a progressive shift toward open-source technologies such as R and Python. These languages offer extensive statistical, analytical, and machine-learning capabilities for better scalability and performance, and they align with the technical skill sets of new statisticians, analysts, data scientists, and programmers joining the agency. This modernization strengthens our ability to work within global statistical and data-science communities.

As part of this broader initiative, we are migrating our SAS-based generalized systems to R and Python. For some of the systems, this includes not just a change in programming language but new functions, features, and overall enhancements. The modernization is already well underway: Banff and G-Series were released publicly in early 2025, and additional systems are on the way. In addition to using R and Python, we also plan on making the source code publicly available on our GitHub account to improve transparency, enhance code quality, and encourage collaboration. The table below summarizes the migration plan and targeted release dates for our SAS-based generalized systems:

Current system (SAS-based)	Functionality	New system (Language)	Release date
Banff	Editing and imputation	Banff & Banff Processor (Python)	January 2025
G-Series	Time series adjustment	G-Series (R)	January 2025
G-Sam	Probabilistic sampling	Jasper (R)	October 2025 (internal); preparing for public dissemination
G-Confid	Disclosure control	G-Confid (Python)	March 2026 (planned)
G-Link	Record linkage	G-Link (Python)	March 2026 (planned)
G-Est	Weighting and estimation	Yoho (R)	June 2026 (planned)

Releasing these systems publicly will strengthen collaboration, reduce duplication of effort across agencies, and promote international methodological consistency. For questions or further information, please feel free to reach out to Darren Gray at darren.gray@statcan.gc.ca.

Released systems:

Statistics Canada (2025). *Banff.* Python package version 3.1.3,

https://github.com/StatCan/gensol-banff.

Statistics Canada (2025). *Banff Processor.* Python package version 2.0.3, https://github.com/StatCan/gensol-banff-processor.

Statistics Canada (2025). *G-Series.* R package version 3.0.2,

https://github.com/StatCan/gensol-gseries.

Croatia

Reporting: **Ksenija Dumičić**

New Developments in the 2024 EU SILC Survey Methodology in Croatia

In 2025, Državni zavod za statistiku (DZS) released the 2024 wave of EU-SILC, based on an updated sampling design and updated data-collection practices. The 2024 sample was drawn as a random sample of private-household dwellings, using the 2021 Population, Households and Dwellings Census as the sampling frame. From 13,049 randomly selected households, 9,410 completed interviews were obtained, yielding a household-level response rate of 77.53 % (DZS, 2025a; DZS, 2025b). The survey retains its panel-design structure: selected households remain in the sample for a four-year rotation, enabling both cross-sectional and longitudinal analyses of income, poverty, material deprivation, and other living conditions indicators. Data collection was carried out primarily via computer-assisted personal interviewing (CAPI), supplemented by telephone interviewing (CATI) when necessary, by authorised and certified DZS interviewers (DZS, 2025b). Importantly, the 2024 wave introduced a **methodological innovation** by integrating administrative data sources for income variables, such as wages, pensions, and social transfers, alongside traditional household interviews. This hybrid approach strengthens data accuracy, reduces respondent burden, and marks a break in the time series compared with previous years when income was collected solely via interviewing (DZS, 2025a). The updated sampling frame based on the 2021 Census ensures up-to-date population coverage; weighting procedures account for unequal selection probabilities, unit non-response, and post-stratification to known population totals at national and regional (HR-NUTS 2) levels. These enhancements considerably improve the representativeness, reliability, and European comparability of Croatian EU-SILC data, aligning the survey with best practices in longitudinal income and living conditions measurement across EU member states.

References

DZS (2025a) *Indicators of Poverty and Social Exclusion, 2024 (EU-SILC Survey)*. First Release, ZUDP-2025-1-1. Zagreb: DZS. Available at: https://podaci.dzs.hr/2025/en/97252

DZS (2025b) *Income and Living Conditions Survey – survey documentation and methodological notes*. Zagreb: DZS. Available at: https://dzs.gov.hr/in-focus/survey-researches/income-and-living-conditions-survey/1787

France

Reporting: **Philippe Brion**

Audience Measurement in France

Audience measurement aims to understand and quantify the number of people exposed to media content, whether on a television channel, radio station, website, or any other medium.

Behind this generic term "measurement" lie, beyond technical aspects, concepts and approaches that can be quite different. There are two main types of audience measurement. On one hand, there are declarative measurements, based on sample surveys, generally involving large samples. Press and radio audience measurements mainly rely on this type of method. On the other hand, there are automatic measurements, which require the installation of a technical measurement system and are based on panel tracking. Television and internet measurements have been built on this type of method.

Many factors contribute to making audience measurement increasingly complex. First, media are evolving, and their usage is becoming more and more fragmented. Accurately measuring the audience of the multitude of available content would require significantly increasing sample sizes. However, the digitalisation of media now offers other data sources: return path data from operator boxes or from publishers' websites and applications. This has led to the emergence of hybrid measurements, resulting from the combined use of these return path data for a precise estimate of the number of visitors/viewers, together with panel data to provide socio-demographic profiles.

Moreover, historically, audience measurements have been constructed in silos: one measurement per medium. However, media are no longer managed in silos, and their content circulates from one channel to another, creating the need for a global cross-media measurement, particularly for advertising campaign measurement. This is why statistical fusion approaches have multiplied, allowing different studies to be brought together by associating respondents based on their similarity across a number of criteria, prioritised according to their explanatory power on the variables of interest, here the audiences. For example, a respondent in the television audience measurement is associated with the radio audience of a “twin” to estimate duplication between the two media. These approaches are becoming more sophisticated today to more broadly take into account media and available data, whether from sample surveys or exhaustive sources. The aim is to create virtual populations on which advertising contacts are distributed according to probabilistic models.

In France, television, radio, and internet audience measurements are carried out by Médiamétrie. More recently, this institute has launched audience measurement for SVOD platforms.

More information :

https://static-webmail.mediametrie.com/Livre+blanc+Hybride+et+IA/EN/Mediametrie+White+Paper+Hybrid+and+AI.pdf

Contact : Aurélie Vanheuverzwyn (avanheuverzwyn@mediametrie.fr)

Hong Kong Special Administrative Region, China

Reporting: **Ronald Chan**

Applications of Large Language Models in Statistical Work

The Census and Statistics Department (C&SD) of Hong Kong, China is actively exploring the applications of Large Language Models (LLMs) to streamline and enhance statistical operations. This initiative supports C&SD's digital transformation strategy to modernise workflows, improve service quality and optimise resource allocation.

Key LLM applications include:

- **Program development**: Automating code generation for statistical programming (Python and SAS), assisting with debugging, and improving documentation efficiency.
- **Data processing**: Extracting and analysing information from unstructured text data, including open-ended survey responses, sales receipts and corporate announcements.
- **Insights generation**: Supporting data validation and review by identifying potential trends, anomalies and patterns in statistical datasets.
- **User services**: Developing intelligent chatbots to handle public inquiries and enhance internal knowledge management.

C&SD is implementing these LLM applications through a carefully planned, phased approach, focusing on enhancing cost-effectiveness of internal processes in the initial phase. To ensure responsible implementation, C&SD will establish robust safeguards, including expert validation of LLM outputs to prevent potential errors and biases. Additionally, computing capabilities are being enhanced to securely handle internal data while providing sufficient power to run LLMs effectively.

For more information, please contact Ronald Chan (rchchan@censtatd.gov.hk).

Lithuania

Reporting: **Danutė Krapavickaitė**

Baltic-Nordic-Ukrainian (BNU) network workshop on survey statistics

The 28th event of the BNU network was organized in Vilnius on August 25-29, 2025. 51 statisticians participated onsite and online from the network countries and from France, Great Britain, the Netherlands and Switzerland. The main topic was “Addressing nonresponse in survey statistics”, other topics were also included.

The keynote speakers gave the talks relevant to the practitioners and theoreticians in survey statistics. The lecture of Alina Matei was entitled “Spatially balanced sampling and its applications in official statistics”. It stimulates to introduce sample coordination system in the case of enterprise surveys and to create the sampling designs for agriculture and social surveys. The lecture of

Guillaume Chauvet "Bootstrap methods in survey sampling with focus on the rescaling bootstrap" will find applications in the future activities of the network. It is useful in social surveys when the survey population is large. The lecture of Jacek Wesolowski "Rotation sampling schemes and Chebyshev polynomials" showed a view of pure mathematics to survey sampling. We were honoured by participation of Professor Carl-Erik Särndal with the talk "The nonresponse dilemma: some thoughts on its origin, impact and future role in survey statistics". It was a pleasure to have among us the founders of the network Imbi Traat and Jānis Lapinš.

Representatives from the network countries gave invited lectures. Other participants gave the contributed presentations, which were discussed by the discussants, appointed in advance. Topics relevant to official statistics, application of machine learning in survey statistics, adjustment for nonresponse were popular. The round table discussions were going on the following urgent topics: "Teaching of survey sampling"; "Modern methods in survey sampling: machine learning, AI, SAE, nonprobability samples"; "Dealing with non-sampling errors and accuracy estimation".

Many young statisticians took part in the workshop. It was their first workshop, in which they presented results of their research. We expect that this event will stimulate their positive attitude to the further development of the statistical science and participation in it.

The workshop was sponsored by Vilnius Gediminas Technical University, State Data Agency, Nordplus, Lithuanian Statistical Society, Lithuanian Mathematical Society and International Association of Survey Statisticians. On behalf of the participants of the workshop, we express sincere thanks for support which has made the event interesting and pleasant.

You may visit the workshop home page:

https://wiki.helsinki.fi/xwiki/bin/view/BNU/Events/Workshop%20on%20Survey%20Statistics%202025/

The Netherlands

Reporting: **Lianne Tessers-Ippel**

Smartphone-first questionnaire design at Statistics Netherlands (CBS)

As smartphones have become the primary device for online survey participation in household surveys, Statistics Netherlands (CBS) developed a *smartphone-first* redesign of its web questionnaires and evaluated this in a large-scale field experiment. The experiment examined six factors: a revised login process, a smartphone-first designed questionnaire layout, alternative grid formats (including carousel and accordion designs), the inclusion of smileys and icons, a speech-to-text encouragement for open questions, and questionnaire length. The questionnaire redesign also aligned with the CBS corporate design system and accessibility requirements.

The results revealed few significant differences across indicators of response behaviour, data quality, and respondent satisfaction. This indicates that the smartphone-first design can be introduced without jeopardising comparability over time or introducing measurement bias. The new design will be gradually implemented in production surveys to ensure a user-friendly experience that is consistent across devices.

A detailed report on the experiment by Deirdre Giesen, Maaike Kompier and Jan van den Brakel is available at "Experiment smartphone-first questionnaire layout" *(*https://www.cbs.nl/en-gb/background/2025/42/experiment-smartphone-first-questionnaire-layout*).*

For more information please contact Deirdre Giesen at d.giesen@cbs.nl.

Poland

Reporting: **Tomasz Żądło**

The 5th Polish Statistics Congress took place from July 1 to 3, 2025, in Warsaw. It was organized by Statistics Poland and the Polish Statistical Association. The event covered topics such as mathematical statistics, survey sampling and small area estimation, population, social, economic, regional, and spatial statistics, data analysis and classification, AI methods, big data and data science, Polish statistics in the international context, history of Polish statistics, communication and education, public statistics and data management systems, as well as data integration and harmonization in official statistics.

The program included two keynote speeches by:

- Professor Partha Lahiri (University of Maryland, College Park) on "Poverty Mapping”
- Professor Ronald Lee (University of California, Berkeley) on “How low fertility and shrinking populations will impact our economies”.

A panel discussion titled “From data to decisions - for social and economic development" was also held, along with 28 sessions and a poster session.

The survey sampling and small-area estimation session organized by Janusz Wywiał, Mirosław Szreder, and Tomasz Żądło contained four presentations:

- invited presentation: “Item Count Techniques under Some Assumption Violations” by Barbara Kowalczyk and Robert Wieczorkowski,
- invited presentation: “Comparing Institutional Performance” by Nicholas Longford,
- “On the Maximum Likelihood Estimation of Population and Domain Means” by Janusz Wywiał,
- “On Complex Estimators Under the Pathak Sampling Scheme” by Krzysztof Szymoniak-Książek.

Program of the conference: https://kongres2025.stat.gov.pl/en/Program

Abstracts and presentations: https://kongres2025.stat.gov.pl/en/Ksiega_abstraktow

Video recordings: https://kongres2025.stat.gov.pl/en/Transmisja

10th Edition of the Statistical Olympiad

The Statistical Olympiad is organized by Statistics Poland and the Polish Statistical Association and is addressed to high school students. Its goal is to promote statistical knowledge and develop skills in socio-economic data analysis through a three-stage competition that selects finalists and laureates. The 10th edition is co-financed by the Ministry of Education and Science, and its first stage took place in November 2025.

The Statistical Olympiad has three consecutive stages designed to gradually evaluate and enhance students' statistical reasoning and data analysis skills. The first stage, the school level, is organized by each participating school and involves an online test that assesses basic knowledge and problem-solving skills in statistics. Schools register their contestants in the Olympiad IT system and appoint a School Committee to oversee the round. The Central Committee then provides the School Committees with the results achieved by the participants.

The regional stage is conducted simultaneously across the country’s 16 regions (voivodeships) under supervised conditions and features a longer, more challenging online exam that includes applied and interpretive problems. Successful participants from the school stage compete at this level for spots that qualify them for the central stage, and rankings are based on test scores with clearly defined tie-breaking rules.

The central stage is held in two parts: an initial central-round online test that filters the top participants, followed by a final written examination for a limited number of contestants who work on more complex, open-ended problems requiring deeper analysis, interpretation of socio-economic data, and clear written justification of methods and conclusions. The final ranking and the list of laureates are based on the combined results from both parts of the central stage.

Winners and finalists of the Statistics Olympiad can gain admission to many prestigious universities in Poland without needing to go through the regular admissions process. In the ninth edition of the competition, the three finalists received gift cards valued at PLN 6,000, PLN 5,000, and PLN 4,000.

United States

Reporting: **Andreea L. Erciulescu**

Principles and Practices for a Federal Statistical Agency

The Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine, identified 5 principles and 10 practices for federal agencies and units to adopt when conducting their activities involving collection, compilation, processing, or analysis of information for statistical purposes. These principles and practices, along with the 16 U.S. federally recognized statistical agencies and units, are provided in the eighth edition of the report titled Principles and Practices for a Federal Statistical Agency and available at the following link: Principles and Practices for a Federal Statistical Agency - 8th Edition | The National Academies Press.

The 5 principles are as follows:

1) relevance to policy issues and society
2) credibility among data users and stakeholders
3) trust among the public and data subjects
4) independence from political and other undue external influence
5) continual improvement and innovation

The 10 practices are as follows:

1) a clearly defined and well-accepted mission
2) necessary authority and procedures to protect independence
3) commitment to quality and professional standards of practice
4) professional advancement of staff
5) an active research program
6) strong internal and external evaluation processes for an agency’s statistical programs
7) coordination and collaboration with other agencies
8) respect for data subjects and data holders and protections of their data
9) dissemination of statistical products that meet users’ needs
10) openness about sources and limitations of the data provided

The 16 federally recognized statistical agencies and units are as follows:

- Bureau of Economic Analysis (U.S. Department of Commerce)
- Bureau of Justice Statistics (U.S. Department of Justice)
- Bureau of Labor Statistics (U.S. Department of Labor)
- Bureau of Transportation Statistics (U.S. Department of Transportation)
- Census Bureau (U.S. Department of Commerce)
- Economic Research Service (U.S. Department of Agriculture)
- Energy Information Administration (U.S. Department of Energy)
- National Agricultural Statistics Service (U.S. Department of Agriculture)
- National Center for Education Statistics (U.S. Department of Education)
- National Center for Health Statistics (U.S. Department of Health and Human Services)
- National Center for Science and Engineering Statistics (National Science Foundation)
- Office of Research, Evaluation, and Statistics (Social Security Administration)
- Statistics of Income Division (U.S. Department of the Treasury)
- Center for Behavioral Health Statistics and Quality (Substance Abuse and Mental Health Services Administration; U.S. Department of Health and Human Services)
- Microeconomic Surveys Unit (U.S. Federal Reserve Board)
- National Animal Health Monitoring System (Animal and Plant Health Inspection Service, U.S. Department of Agriculture)

Uruguay

Reporting: **Marcelo Bisogno, Juan Pablo Ferreira and Juan José Goyeneche**

Weighting the 2023 Uruguay Census: Correcting Omission Through Doubly Robust Estimators

The 2023 Population, Household, and Housing Census of Uruguay applied a mixed methodology that combined the traditional census operation (CAWI + CAPI) with the inclusion of individuals identified through administrative registers. This integration improved population coverage and reduced net undercount by incorporating individuals who were not effectively enumerated but showed “signals of life” in systems such as education, health, or social security. This approach contributed to adjusting aggregate population totals and the age–sex structure.

The inclusion of individuals from administrative registers particularly improved coverage at higher geographic levels, such as departments. However, this strategy presents limitations when more granular territorial information is required. Geographic allocation based on administrative registers tends to concentrate individuals in urban areas, where administrative activity is more complete, potentially leading to underestimation of the population in rural areas or smaller communities. Moreover, administrative registers cannot replace the census operation, as they do not contain key census variables such as housing conditions, household equipment, household structure and composition, or socioeconomic characteristics. For this reason, individuals “incorporated through registers" improve coverage but do not constitute complete observations for producing many census indicators.

Despite the strategy implemented, overall omission reached 10.3%, meaning that one in ten individuals was not enumerated. This omission was differential, with higher incidence in lower socioeconomic areas and regions with more difficult access. Consequently, the enumerated population does not represent a random selection of the total population but rather a subset with patterns of self-selection associated with social and territorial characteristics.

To produce valid estimates for census variables not available in administrative registers, it was necessary to construct weights that corrected differential omission. Households that were effectively enumerated were treated as a non-probability sample, and a doubly robust estimation approach was applied, combining a model for the propensity to be enumerated with a superpopulation model for each variable of interest. Under this framework, estimates can be unbiased if at least one of the models is correctly specified.

The propensity to be enumerated was estimated under the assumption that the response mechanism is Missing At Random (MAR), along with the construction of nonresponse classes at the level of groups or enumeration areas, assuming that the propensity (i.e., voluntarism) is homogeneous within each nonresponse class. Finally, a post-stratification estimator (a specific case of the linear regression estimator) was applied so that the estimates match the simple age–sex structure of the population. The resulting weights correct differential omission, reduce bias, and yield reliable indicators at the global level and for different domains of estimation.

For more information: mbisogno@ine.gub.uy , jferreir@ine.gub.uy , jgoyeneche@ine.gub.uy

Events on survey statistics and related areas

2026 Survey Costs Workshop

Date: 9-10 February 2026

Location: Maryland, USA

Webpage: https://aapor.org/aapor-2026-survey-costs-workshop/

20th IAOS Conference

Date: 12-14 May 2026

Location: Vilnius, Lithuania

Webpage: https://www.isi-next.org/conferences/iaos-2026/

14th International Francophone Conference on Surveys

Date: 19-22 May 2026

Location: Vannes, France

Webpage: https://sondages2026.sciencesconf.org/

5th ISI Regional Statistics Conference

Date: 3-5 June 2026

Location: Valleta, Malta

Webpage: https://www.isi-next.org/conferences/isi-rsc-malta-2026/

The Seventh International Workshop on Business Data Collection Methodology

Date: 8-10 June 2026

Location: Heerlen, the Netherlands

Webpage: https://2026bdcmw.wordpress.com/

The Small Area Estimation Conference 2026

Date: 15-19 June 2026

Location: Bucharest, Romania

Webpage: https://sae2026.faa.ro/

BNU Workshop on survey statistics 2026

Date: 24-28 August 2026

Location: Riga, Latvia

Webpage:

https://wiki.helsinki.fi/xwiki/bin/view/BNU/Events/Workshop%20on%20Survey%20Statistics%202026/

Journal of Survey Statistics and Methodology

Volume 13, Issue 4, September 2025
https://academic.oup.com/jssam/issue

Survey Methodology

An Experimental Comparison of Modular and Non-Modular Approaches for Administering Surveys via Smartphone Apps
Christopher Antoun, Brady T. West, Xin (Rosalynn) Yang, Syed Junaid M. A. Zaidi and Jennifer Sinibaldi

Question form Matters: Examining Trust in Government Through Open and Closed Survey Items
Jana Bernhard-Harrer and Katharina Pfaff

Is Consent to Further Panel Participation Selective? The Case of a Self-Administered Family Panel Survey Announcing Organizational Change
Almut Schumann and Claudia Schmiedeberg

Survey Statistics

Synthesizing Surveys with Multiple Units of Observation: An Application to the Longitudinal Aging Study in India
Joshua Snoke, Erik Meijer, Drystan Phillips, Jenny Wilkens and Jinkook Lee

Bayesian Tree Models for Survey Sample Data
Daniell Toth, Scott H. Holan and Diya Bhaduri

Journal of Official Statistics

Volume 41, Issue 4, December 2025
https://journals.sagepub.com/toc/jofa/41/4

Articles

The Poverty Free Life Expectancy in Europe
Gianni Betti, Federico Crescenzi and Andrea Nigri

Official Statistics and Government Decision Making: A Bibliometric and Thematic Analysis of Policy-Related Academic Research
Ana Božič Verbič

Why are Measures of Aggregate Hours Worked by the Unincorporated Self-Employed So Volatile?
Cindy Michelle Cunningham and Sabrina Wulff Pabilonia

Decomposing Residential Resale House Prices into Structure and Land Components
Erwin Diewert and Ning Huang

Connected and Uncooperative: The Effects of Homogenous and Exclusive Social Networks on Survey Response Rates and Nonresponse Bias
Jonathan Eggleston and Chase Sawyer

STAHL: Seasonal, Trend, and Holiday Decomposition with Loess
Vincent Haller, Sebastien Daniel and Benoit Bellone

Age-Period Modeling of Mortality Gaps: The Cases of Cancer and Circulatory Diseases
Giacomo Lanfiuti Baldi, Andrea Nigri and Han Lin Shang

Higher-Level Aggregation Using Long-Term Links in the Swedish CPI
Olivia Ståhl

When Cleaning Data Introduces Bias: A Critical Examination of Post-Hoc Methods in Detecting Insufficient Effort Responding
Melissa Dan Wang, Leifeng Xiao and Xuan Zang

Research Note
Which Nordic Countries Are the Most and Least Urban? The Muddling World of Urbanity Statistics
Marianne Tønnessen

Book Review
Review of "Register-Based Statistics – Registers and the National Statistical System"
Paul A. Smith

Volume 41, Issue 3, September 2025
https://journals.sagepub.com/toc/jofa/41/3

Foreword

Celebrating JOS Forty Years: Future Research Needs in the New Era of Official Statistics
Lilli Japec, Henri Luomaranta-Helmivuo, Li-Chun Zhang, Suad Elezovic and Yingfu Xie

JOS40

First Forty Years of Journal of Official Statistics
Risto Lehtonen

Challenges in a Federal Statistical Agency Ecosystem: The U.S. Census Bureau Robert L. Santos
GDP and Beyond: Dilemmas and Heresies
Steve MacFeely

Future Pathways Embracing Multisource Statistics and Novel Data Sources at National Statistical Offices
Anders Holmberg

Some Dimensions of Statistical Ethics and Scientific Integrity That Warrant Exploration Through Empirical Studies of Stakeholder Information Needs
John L. Eltinge

In Other Journals

Input Privacy Enhancing Technologies for Statistical Production: Motivations and Challenges
Fabio Ricciato

Statistical Disclosure Control: Moving Forward
Josep Domingo-Ferrer, David Sánchez and Krishnamurty Muralidhar

Competence, Training, and Collaboration of Universities with National Statistical Offices
Danny Pfeffermann

Where Have the Respondents Gone? Did We Lose Them or Failed to Win Them? And Is It Too Late?
Barry Schouten

The Future of Interviewer-Administered Surveys
Kristen Olson

Future Research Considerations for Mixed-Mode Surveys
Leah Melani Christian

Modernizing Data Collection
Frauke Kreuter

Census Transformation and the Future of Population Statistics
James J. Brown and James Chipperfield

Calibration Weighting for Analyzing Non-Probability Samples
Jae Kwang Kim

Blending Administrative and Nonprobability Survey Data to Enhance National and Subnational Estimates
Dan Liao and Paul P. Biemer

Future Challenges in Sampling and Estimation
Guillaume Chauvet

The Unknown Future of Statistical Data Editing: Some Imputations
Sander Scholtus

Machine Learning Methods for Estimation in Official Statistics
M. Giovanna Ranalli

Small Area Estimation in the Era of Machine Learning and Alternative Data Sources: Opportunities, Challenges, and Outlook
Nikos Tzavidis

Some Challenges and Research Needs for the Analysis of Integrated Data
Raymond L. Chambers
Challenges and Opportunities for Analytic and Causal Inference with Official Statistics
F. Jay Breidt, Robert Ashmead and Susan M. Paddock

Enhancing Microsimulation by Open Data
Ralf Thomas Münnich

Challenges and Future Directions for International and Cross-Cultural Comparability
Julie de Jong

Survey Research Methods

Volume 19, No.4, 2025
https://ojs.ub.uni-konstanz.de/srm/

Articles

Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation
Leah von der Heyde, Anna-Carolina Haensch, Bernd Weiß, Jessica Daikeler

A Matter of Perspective? Differences Between Adolescent–Parent and Parent–Teacher Pairs in Responses to the Strengths and Difficulties Questionnaire Using a Scottish National Cohort Study
Madison Bunker, Valeria Skafida, Emma Davidson

The Effects of Study Duration on Nonresponse and Measurement Quality in a Smartphone App-Based Travel Diary
Danielle Remmerswaal, Peter Lugtig, Barry Schouten, Bella Struminskaya

Invitation Messages for Business Surveys: A Multi-Armed Bandit Experiment
Johannes J. Gaul, Florian Keusch, Davud Rostam-Afschar, Thomas Simon

Effects of Replacing Telephone with Web, Mail, and Mixed-Mode Data Collection in an Establishment Follow-Up Survey
Benjamin Küfner, Joseph W. Sakshaug, Stefan Zins, Claudia Globisch

The Effect of Targeted Incentives on Response Rates and Representativeness: Evidence From the Next Steps Age 32 Survey
Alessandra Gaia, Matt Brown, Tugba Adali, Stella Fleetwood, Christy Lai

Retrieving True Preference under Authoritarianism
Jongyoon Baik, Xiaoxiao Shen

Measuring Gender and Sex in Surveys: Lessons Learned from 50 Years of Cross-National Survey Data and Nonresponse Patterns
Ilona Wysmulek

Volume 19, No.3, 2025
Special Issue: Survey Climate and Trust in Scientific Surveys
https://ojs.ub.uni-konstanz.de/srm/issue/view/243

Editorial

Survey Climate and Trust in Scientific Surveys: Introduction to the Special Issue
Henning Silber, Bettina Langfeldt, Bella Struminskaya, Michael Traugott

Articles

Predicting Survey Nonresponse with Registry Data in Sweden between 1993 and 2023: Cohort Replacement or a Deteriorating Survey Climate?
Sebastian Lundmark, Kim Backström

Effects of Survey Design Features on Response Rates: A Meta-Analytical Approach Using the Example of Crime Surveys
Jonas Klingwort, Vera Toepoel

Survey Attitude as Indicator for Survey Climate and as Predictor of Nonresponse and Attrition in a Probability-Based Online Panel
Benjamin Rosche, Hugo Bons, Joop Hox, Edith De Leeuw

Trust, Concerns and Attitudes: Examples for Respondent (Non-)Cooperation in SHARE
Imke Herold, Michael Bergmann, Arne Bethmann

Trust in Survey Results: A Cross-Country Replication Experiment
Adam Stefkovics, Zoltán Kmetty

Public Confidence in Official Statistics in the UK: Characteristics of Respondents and Changes Over Time
Olga Maslovskaya, Annamaria Bianchi

Using Experimental Vignettes to Study How Survey Methods and Findings Affect the Public's Evaluation of Public Opinion Polls: Considering a Dual-Process Approach
Allyson L. Holbrook, Paul J. Lavrakas, Timothy P. Johnson, Andrew Crosby, Polina Polskaia, Xiaoheng Wang, Xiaoyan Hu, Evgenia Kapousouz, Young Ik Cho, Henning Silber

Volume 19, No.2, 2025
https://ojs.ub.uni-konstanz.de/srm/issue/view/244

Articles

Quality of Expenditure Data Collected With a Mobile Receipt Scanning App in a Probability Household Panel
Alexander Wenz, Annette Jäckle, Jonathan Burton, Mick P. Couper, Brendan Read

Pre-Trained Nonresponse Prediction in Panel Surveys with Machine Learning
John Collins, Christoph Kern

Surely Shorter Is Better? A Questionnaire Length Experiment in a Self-Completion Survey
Tim Hanson, Eva Aizpurua, Rory Fitzgerald, Marta Vukovic

Internet Coverage Bias in Web Surveys in Europe
Alessandra Gaia, Emanuela Sala, Chiara Respi

The Impact of Scale Direction on Data Quality
Ting Yan, Alexandru Cernat, Florian Keusch

Response Burden and Response Quality in Web Probing: An Experiment on the Effects of Probe Placement and Format
Patricia Hadler

Effects of Mode and Transitioning to a Mixed-Mode (Web/Phone) Design on Categorical Survey Estimates: Do Question Characteristics Matter?
Mengyao Hu, Vicki Freedman, Justin Kamens

Effects of Changing the Incentive Strategy on Panel Performance: Experimental Evidence From a Probability-Based Online Panel of Refugees
Jean-Philippe Décieux, Sabine Zinn, Andreas Ette

Other Journals

- **Statistical Journal of the IAOS**
  https://content.iospress.com/journals/statistical-journal-of-the-iaos/
- **International Statistical Review**
  https://onlinelibrary.wiley.com/journal/17515823
- **Transactions on Data Privacy**
  http://www.tdp.cat/
- **Journal of the Royal Statistical Society, Series A (Statistics in Society)**
  https://rss.onlinelibrary.wiley.com/journal/1467985x
- **Journal of the American Statistical Association**
  https://amstat.tandfonline.com/uasa20
- **Statistics in Transition – New Series**
  https://sit.stat.gov.pl

Welcome New Members!

We are very pleased to welcome the following new IASS members:

Title	First name	Surname	Country
Mrs.	Clyde E.	Charre de Trabuchi	Argentina
Mr.	Tony	Labillois	Canada
Ms.	Anita	Harmina	Croatia
Dr.	Yannick	Lemel	France
Dr.	Christophe	Lefranc	France
Dr.	Jan Pablo	Burgard	Germany
Dr.	Faisal	Awartani	Israel
Mr.	Leonard Warren	Cook	New Zealand
Mr.	Bjorn K. Getz	Wold	Norway
Mrs.	Awa	Thiongane	Senegal
Dr.	Willie	Lahari	Solomon Islands
Dr.	Hyoung Il	Lee	South Korea
Professor	Beat	Hulliger	Switzerland
Dr.	Philippe	Eichenberger	Switzerland
Professor	H. Öztas	Ayhan	Turkey
Professor	Peter J.	Lynn	United Kingdom
Dr.	Tarek	Al Baghal	United Kingdom
Mr.	Gary	Bennett	United Kingdom
Professor	Martin R.	Frankel	United States
Professor	Roderick J	Little	United States
Dr.	Daniel	Kasprzyk	United States
Mr.	Edward J.	Spar	United States
Dr.	Keith	Rust	United States
Dr.	J. Michael	Brick	United States
Dr.	David Alan	Marker	United States
Ms.	Francesca	Perucci	United States
Professor	Juan Pablo	Ferreira Neira	Uruguay
Mr.	Oliver	J. M Chinganya	Zambia

IASS Executive Committee Members

Executive officers (2025 – 2027)

President:	Partha Lahiri (USA)	plahiri@umd.edu
President-elect:	Ralf Münnich (Germany)	muennich@uni-trier.de
Vice-Presidents:		
Scientific Secretary	Katherine Jenny Thompson (USA)	jennythompson731967@gmail.com
VP Finance and IASS conferences support	Partha Lahiri (USA) Ralf Münnich (Germany)	plahiri@umd.edu muennich@uni-trier.de
Liaising with ISI EC and ISI PO plus administrative matters	Ralf Münnich (Germany)	muennich@uni-trier.de
Chair of the 2025 Cochran-Hansen Prize Committee Chair of the 2024 Hukum Chandra Prize Committee IASS representative on the ISI Awards Committee	Robert Clark (Australia)	robert.clark@anu.edu.au
IASS representatives on the World Statistics Congress Scientific Programme Committee IASS representative on the World Statistics Congress short course committee	Ralf Münnich (Germany)	muennich@uni-trier.de
IASS representative on the ISI publications committee	Partha Lahiri (USA)	plahiri@umd.edu
IASS Webinars 2025-2027 Volunteer for supporting training and Webinar activities within ISI Statistical Capacity Development Committee	Haoyi Chen (China)	
IASS representative on the Regional Statistics Conference 2026 IASS Social Media	Gaia Bertarelli (Italy)	gaia.bertarelli@unive.it
Ex Officio Member:	Conchita Kleijweg (The Netherlands)	c.kleijweg@isi-web.org

IASS LinkedIn Account:
https://nl.linkedin.com/company/international-association-of-survey-statisticians-iass

IASS Facebook Account: https://www.facebook.com/iass.isi/

IASS X Account: https://x.com/iass_isi/

IASS Webmasters: Ujjayini Das (ujstat@umd.edu) and Sabrina Zhan (SabrinaZhang@westat.com)

IASS Institutional Members

International organisations:

- Eurostat (European Statistical Office) – Unit 01: External & Inter., Luxembourg

National statistical offices:

- Australian Bureau of Statistics, Australia
- Instituto Brasileno de Geografia y Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- International Rel. & Statistical Coordination, Israel
- Istituto nazionale di statistica (ISTAT), Italy
- Statistics Korea (KOSTAT), Republic of Korea
- Direcção dos Serviços de Estatística e Census (DSEC), Macao, SAR China
- Statistics Mauritius, Mauritius
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Instituto Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- Office for National Statistics Service (ONS), United Kingdom
- National Agricultural Statistics Service (NASS), United States
- National Center of Health Statistics, United States

Universities:

- Department of Mathematics and Statistics, University of Ottawa, Canada
- Univ. of Tuscany, Dept. Economics & Management, Italy

Other statistical organizations:

- Institut Public de Sondage d'Opinion Secteur (Ipsos), Italy
- WESTAT Inc., United States