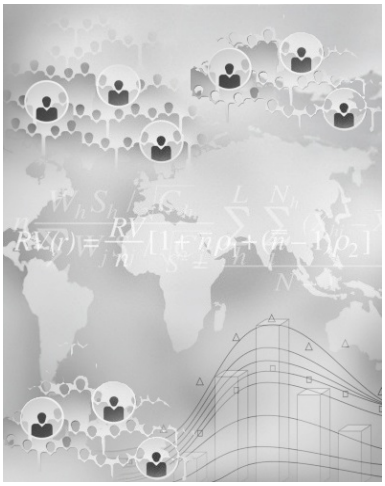
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

No. 94 **July 2026**

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

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



The Survey Statistician No. 94, July 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

6 Report from the Scientific Secretary

8 News and Announcement

16 Debate
- Can machine learning effectively reduce potential nonresponse bias? By David Haziza and Roderick J. Little.

19 Ask the Experts
- On the use of geospatial data in small area estimation by Nikos Tzavidis, Luciano Perfetti Villa, and Vasilis Chasiotis.

32 New and Emerging Methods
- Data Privacy in the AI Era and Implications for Survey Practice by Shurong Lin and Aleksandra Slavković.

47 Early Career Survey Statistician
- High-Dimensional Variance Estimation for the Generalized Regression Estimator by Kalil Bouhadra and Mehdi Dagdoug.

70 Book Review
- “Statistics in Survey Sampling” by Changbao Wu.

72 *Country Reports*
- Argentina, Australia, Brazil, Canada, Croatia, The Netherlands, United States, Uruguay.

80 Upcoming Conferences and Workshops

82 In Other Journals

87 Welcome New Members!

88 IASS Executive Committee Members

89 IASS Institutional Members

Letter from the Editors

Dear colleagues and readers,

This issue of *The Survey Statistician* (TSS) features contributions from Partha Lahiri, the elected president of the International Association of Survey Statisticians (IASS), and Jenny Thompson, the appointed scientific secretary.

We are delighted to announce the winners of the 2026 Hukum Chandra Memorial Prize, which is awarded by the IASS. We also have news about the 33rd Morris Hansen Lecture. We are also pleased to announce that all back issues of TSS from 1978 onwards are now available to read on the IASS website. This would not have been possible without the efforts of Eric Rancourt, former editor of TSS.

This issue contains several sections. In the `*Debate*' section, David Haziza and Roderick J. Little answer `yes' and `no', respectively, to the question `*Can machine learning effectively reduce potential nonresponse bias?*'. In the `*Ask the Experts*' section, Nikos Tzavidis, Luciano Perfetti Villa, and Vasilis Chasiotis focus on the use of geospatial data in small area estimation. In the `*New and Emerging Methods*' section, Shurong Lin and Aleksandra Slavković discuss data privacy in the artificial intelligence era and the implications for survey practice. The `*Early Career Survey Statistician*' section includes the paper `High-Dimensional Variance Estimation for the Generalized Regression Estimator' by Kalil Bouhadra and Mehdi Dagdoug. Changbao Wu reviews the book `Statistics in Survey Sampling' by Jae K. Kim. Finally, the country reports are provided.

This issue ends with the list of upcoming conferences and workshops, a list of articles published in other journals, and the list of new IASS members.

We would like to thank all the authors and contributors to this issue, and to express our gratitude to the TSS editors for their rigour and dedication: Gaia Bertarelli, Mehdi Dagdoug, Francesco Pantalone, Jenny Thompson, Ton de Waal, and Peter Wright.

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 July 2026 issue!

Alina Matei and **Andrea Diniz da Silva**,

TSS Editors

Letter from the President

Dear IASS Members,

In my previous letter, I shared our core long-term vision. Today, I want to dedicate this space to a theme that is vital to the sustained growth of our association: strategic collaboration. To expand our global footprint and deliver maximum value to survey statisticians in a broad sense, the IASS is actively seeking, establishing, and deepening partnerships with associations, national societies, and international working groups.

Our ongoing monthly webinar series serves as a prime testament to this outward-looking strategy. We have achieved remarkable success by organizing featured webinars jointly with the Survey Research Methods Section of the American Statistical Association (ASA-SRMS), the International Society for Bayesian Analysis (ISBA), and the Inter-Secretariat Working Group on Household Surveys, mandated by the United Nations Statistical Commission. These joint initiatives enrich our shared intellectual space, bringing diverse methodologies together under a unified, accessible umbrella.

A recent highlight of this collaborative approach was our partnership with the International Statistical Education Center (ISEC). Together, we conducted an exceptionally successful hybrid workshop titled *Practical SAE: Implementing Solutions Using Multiple Data Sources* at the Indian Statistical Institute (ISI) Kolkata campus. I want to express my sincere gratitude to the instructors, as well as the ISEC and ISI PO staff, for their exceptional coordination. As detailed in the accompanying Scientific Secretary's report, this new initiative was a truly global success, drawing 41 participants from more than 20 countries across six continents.

Looking ahead, we are actively translating this momentum into a vibrant future calendar. We expect to partner with the Washington Statistical Society (WSS) and various regular conferences such as the Small Area Estimation (SAE) conference series and World Statistics Congress (WSC) Satellite meetings to offer specialized short courses of direct interest to our membership. Most excitingly, we are officially collaborating with the global SAE community to co-sponsor a WSC 2027 Satellite Conference in Hue, Vietnam, from June 21–25, 2027.

I am also delighted to reflect on the outstanding success of the recent IASS co-sponsored *Small Area Estimation, Survey and Data Science Conference (SAE 2026)*, held from June 15 to 19, 2026, at the University of Bucharest, Romania (further details can be found in the dedicated article featured in this issue). The event uniquely underscored the growing importance of data integration and model-based methods in contemporary official statistics. It is against the backdrop of this vibrant scientific gathering that I extend my warmest, heartfelt congratulations to Professor Isabel Molina for being honored with this year's prestigious SAE Award. Her achievements inspire us all and perfectly embody the standard of scientific excellence that the IASS strives to support.

On behalf of the entire IASS Executive Committee, it is my absolute pleasure to congratulate the joint winners of this year's prestigious Hukum Chandra Memorial Prize: Professor Sixia Chen (University of Oklahoma Health Sciences and U.S. Census Bureau) and Professor Yan Li (University of Maryland, College Park). Both awardees will showcase their research during our upcoming

webinar series in October and November 2026. I highly encourage you to attend; please click *here* to view the IASS webinar schedule and related information.

The strength of the IASS lies within its members. As we look to expand our collaborative horizons, I warmly invite you to reach out to me directly at **plahiri@umd.edu** with any new ideas for promoting our mission, establishing new institutional ties, or introducing innovative topics to our scientific programs.

Thank you for your continued dedication to our community.

Warmest regards,
Partha Lahiri
IASS President 2025–2027

Report from the Scientific Secretary

The IASS Executive Committee is delighted to announce that Sixia Chen (University of Oklahoma Health Sciences, USA and U.S. Census Bureau) and Yan Li (University of Maryland, College Park, USA) have been jointly awarded the Hukum Chandra Memorial Prize. Dr. Chen is being recognized for contributions to robust methods for handling influential units and missing data in survey sampling, with a webinar presentation entitled *"Robust imputation procedures in the presence of influential survey data based on adaptive tuning constants."* Professor Li is being recognized for contributions to survival analysis and health disparity decomposition using complex survey and observational data, with a webinar presentation entitled *"Correcting Bias from Covariate-Dependent Censoring in Survival Disparity Decomposition."* The presentations will be delivered in the IASS webinar series in October and November 2026. As expected, the pool of nominees was exceptionally strong, and we thank the selection committee - Robert Clark, Alina Matei and Maria Giovanna Ranalli - for their careful consideration.

The IASS is serving the international statistical committee in a variety of ways, partnering with other organizations to extend our reach. At the invitation of Professor SP Mukherjee, Chairman of the Board of Directors of the *International Statistical Education Center (ISEC)*, we conducted a hybrid IASS-ISEC workshop entitled "*Practical SAE: Implementing Solutions Using Multiple Data Sources*" from 18 May 2026 to 21 May 2026 held at the Indian Statistical Institute (ISI) Kolkata (India) campus, co-taught by Partha Lahiri (University of Maryland) and Yuting Chen (Eastern Kentucky University). Forty-one participants represented six continents covering more than 20 countries: 18 onsite ISEC funded participants from 11 developing countries and 23 online participants paying 120 euros in registration fees. Other IASS EC contributions came from Ralf Münnich (inaugural session remarks), Haoyi Chen (capacity building discussion), and Gaia Bertarelli (Q&A sessions). Given the success of this workshop, we are pursuing a similar collaboration with the Washington Statistical Association (WSS), one of the largest chapters of the American Statistical Association. This would likely be an online workshop, and we are brainstorming with WSS officials on a suitable topic.

We have conducted six successful webinars since January:

IASS Webinar 60: "*Survey Quality Frameworks and Quality Assessment*" presented by Julie de Jong in partnership with the Inter-secretariat Working Group on Household Surveys (ISWGHS) of the United Nations, 21 January 2026.

IASS Webinar 61: "*Reducing Measurement and Sampling Biases in Non-probability Surveys*" presented by Aditi Sen in partnership with the American Statistical Association's Section on Survey Research Methods, 27 February 2026.

IASS Webinar 62: "*Seminal Ideas and Controversies in Statistics*" presented by Professor Roderick Little in partnership with the American Statistical Association's Section on Survey Research Methods and International Society for Bayesian Analysis, 23 March 2026.

IASS Webinar 63: "*How We Travel: Insights from Household Travel Surveys and Digital Twin Modeling*" presented by Dr. Cinzia Cirillo, Professor, the University of Maryland and Director, Center for Multimodal Mobility, 21 April 2026.

IASS Webinar 64: "*A Practical Guide for NSOs on Addressing Nonresponse Bias in Social Surveys*" presented by Cilanne Boulet (Statistics Canada), Dr. Eric Lesage

(INSEE), and Kenza Sallier (Statistics Canada) in partnership with the ISWGHS, 19 May 2026.

IASS Webinar 65: "*Data Collection - UN Handbook of Surveys of Households and Individuals Chapter Preview*" presented by Elizabeth J. Zechmeister (Vanderbilt University), Joost Kappelhof (Netherlands Institute for Social Research, and Michael Robbins (Arab Barometer) in partnership with the ISWGHS, 16 June 2026.

All webinars were well attended, with co-sponsorship generally increasing participation: 187 attendees at Webinar 60, 156 attendees at Webinar 61, 242 attendees at Webinar 62, 88 attendees at Webinar 63, 176 attendees at Webinar 64, and 92 attendees at Webinar 65. See our events page for updates in 2026.

The IASS provides no-cost co-sponsorships to conferences of interest to IASS members. Since January, these conferences include the *Survey Cost Workshop* in Washington, D.C. from 9-10 February 2026 and the *IASS–RSS–NLG International Conference 2026* in Nigeria on 10 March 2026. The IASS organized seven Invited Paper Sessions (IPS) and an Special IPS sesion at the *5th ISI Regional Statistics Conference* held in Valletta, Malta from 3-5 June 2026. We thank Gaia Bertarelli for representing the IASS on this conference's organizing committee. The IASS and the International Association of Official Statistics (IAOS) co-sponsored the *Small Area Estimation Conference 2026* (SAE 2026) in Bucharest, Romania, from 15-19 June 2026; see the President's report for more details.

Finally, a few miscellaneous items of interest. First, we hope that you enjoy this issue's New and Emerging Methods paper entitled "*Data Privacy in the AI Era and Implications for Survey Practice*" by Shurong Lin and Aleksandra Slavković (both at the Pennsylvania State University, USA). This thought-provoking paper summarizes material presented at the 2025 Links Lecture conducted by the American Statistical Association. Second, look for an upcoming special issue of the *Revista Colombiana de Estadística* celebrating the 140th anniversary of the International Statistical Institute, developed in collaboration with the ISI Outreach Committee for Latin America and the Caribbean and the ISI Young Statisticians Committee, and formally endorsed by the ISI Executive Committee.

For timely updates and information from the IASS, visit us at LinkedIn and Facebook. 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. 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 Hukum Chandra Memorial Prize

We are proud to share that Professor Yan Li and Associate Professor Sixia Chen are the joint recipients of the 2026 Hukum Chandra Memorial Prize.

Dr. Yan Li

Dr. Sixia Chen

About the Hukum Chandra Memorial Prize

The IASS established a biennial prize in 2022 in honour of Dr Hukum Chandra who passed away in 2021. Dr Chandra was a remarkable statistician and much-missed colleague who authored many important papers in small area estimation, survey and official statistics and the application of demography to statistics, for which he was recognised by the prestigious Cochran-Hansen Award.

The prize is awarded to a mid-career statistician whose work is related to that of Hukum Chandra, namely survey sampling, small area estimation, official statistics, spatial analysis applied to official and survey statistics, and agricultural statistics. The prize is aimed at mid-career researchers whose career stage is close to Dr Chandra's career trajectory.

The prize includes an honorarium of 500 euros, and invited presentations in the IASS webinar series in October and November 2026. Further details will be announced closer to the dates.

About Dr. Sixia Chen and Dr. Yan Li

Sixia Chen (University of Oklahoma Health Sciences Center, USA) and Yan Li (University of Maryland, USA) are leading researchers in survey statistics and methodology. Sixia Chen's work has advanced the analysis of missing data, complex surveys and integrated data sources, with important applications in public health and health disparities research. Yan Li's research has advanced the design and analysis of complex surveys, nonprobability samples and population health studies, particularly in improving population representativeness and addressing selection bias. Both have published extensively in leading statistical journals and have provided distinguished leadership through research, education and service to the international survey statistics community.

Selection Committee

The prize recipient was selected from a strong field by a three-person committee, consisting of Robert Clark (chair), Alina Matei and Maria Giovanna Ranalli.

Robert Clark,

Chair of the 2026 Hukum Chandra Prize Committee

The 33rd Annual Morris Hansen Lecture

Dr. Partha Lahiri

The Washinton Statistical Society's 33rd Annual Morris Hansen Lecture was delivered by **Dr. Partha Lahiri** of the University of Maryland on March 30, 2026 at Summit Consulting LLC, with 105+ attendees.

The lecture, titled *`Combining Information from Multiple Data Sources Using Statistical Modeling and Methods',* featured discussions by Rebecca Steorts of Duke University and Lisa Mirel of the National Science Foundation. The session was chaired by Carolina Franco. A catered reception followed the lecture.

About the Morris Hansen Memorial Lecture Series

The Morris Hansen Lecture series was established by the Washington Statistical Society (WSS) in 1990 with financial support from Westat, Inc. to honor Morris Hansen, whose pioneering contributions to survey sampling and related statistical methods during his long and distinguished career at the Census Bureau and at Westat established many standards and methods, mostly still in use, for the conduct of surveys. Morris Hansen served as the inaugural president of the International Association of Survey Statisticians (IASS). The Morris Hansen Lecture usually is held in the fall of the year, typically October or November. The usual format is to have a primary speaker of outstanding merit cover an important topic of wide interest, and two discussants, one local. The Hansen Lecture series seeks to achieve balance between theory, applications, and policy; and to highlight the diversity of disciplines that inform survey practice. The speaker receives an honorarium of $1000. Travel expenses are paid for the speaker and discussants.

The WSS, Summit Consulting LLC, Westat, and USDA NASS provided support for the 33rd lecture.

About Dr. Partha Lahiri

Dr. Partha Lahiri is a Professor of the Joint Program in Survey Methodology (JPSM) and the Department of Mathematics at the University of Maryland College Park (UMD). He served as the JPSM Director from January 2021 through June 2025. Prior to joining UMD, Dr. Lahiri was the Milton Mohr Professor of Statistics at the University of Nebraska-Lincoln. Dr. Lahiri is serving as the President of the International Association of Survey Statisticians during 2025–2027. His research interests include survey statistics, Bayesian statistics, statistical data integration, and small-area estimation. He has published over 85 papers in peer-reviewed journals, delivered 20 plenary/keynote presentations and over 90 invited talks in professional meetings worldwide. He has served on several advisory committees, including the U.S. Census Bureau Advisory Committee of Professional Associations (chair in 2006) and a U.S. National Academy of Sciences panel, and served as consultant/advisor for international organizations such as the United Nations Statistics Division and the World Bank. Dr. Lahiri is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He received

the 2021 SAE Award at the 63rd World Statistics Congress Satellite Meeting on Small Area Estimation in recognition of his lifetime contributions to small area estimation research. More recently, Dr. Lahiri was awarded the Neyman Medal at a joint session of the 3rd Congress of Polish Statistics and the 2022 Conference of the International Association of Official Statistics held in Krakow, Poland, for outstanding contributions to the development of statistical sciences.

Summary of the lecture

The demand for statistics on diverse topics - including socio-economic conditions, agriculture, health, and transportation - is on the rise, while governments and survey organizations have strived to address the increasing costs of conducting high-quality surveys. Along with technological advancements, the increasing accessibility of various data sources, including administrative records, geospatial data, social media data, and AI-generated data, presents researchers with new opportunities to produce improved estimates. In addition, this allows for the investigation of complex problems that would be challenging using only a single data source. Recently there has been a significant surge in statistical methodological research for diverse applications that is focused on combining information from multiple data sources.

Dr. Lahiri began by discussing the scope of statistical modeling for harnessing information from multiple data sources to produce precise estimates at a granular level, conduct multivariate analysis when a single data source lacks all relevant variables, reduce nonsampling errors in probability samples, mitigate self-selection biases in nonprobability samples, and address other emerging challenges. He then focused on recent statistical methodological developments for combining information from multiple data sources in the context of small area estimation for poverty mapping—a topic of significant interest to various national statistical offices and international agencies.

For more information on the Morris Hansen Lecture, please visit: https://washstat.org/hansen/#PROGRAMS

Dr. Benmei Lu,

WSS President, 2025-2026

Other News

All back issues of *The Survey Statistician* on the IASS website

Until May 2026, *The Survey Statistician* issues from 2000 onwards were available in PDF format on the IASS website.We have now been able to post all back issues of *The Survey Statistician* onto the IASS website. They are available from the very first issue in 1978. Prior to *The Survey Statistician*, there were 8 issues of an IASS smaller news bulletin. Issue #8 of that bulletin constituted a "dress rehearsal" and thus was called issue #0 of *The Survey Statistician*. This issue is also included. We would like to thank **Eric Rancourt** (former editor of *The Survey Statistician*) from Statistics Canada who made this possible and who has graciously used his collection to provide scanned issues to the IASS thanks to Statistics Canada. His few missing issues were completed by the help of Gordon Brackstone's collection as well as a couple of issues from Statistics Canada's library. We would also like to thank Benoit Rehayem and Zumreta Demirovic from Statistics Canada for scanning the documents and making them available to the IASS. Final thanks go to Ujjayini Das for mounting all these documents onto the IASS website. If you are interested in some statistics about and history of *The Survey Statistician*, please see Eric Rancourt's article in the July 2023 issue #88.

Modernizing Establishment Statistics: Insights from EESW25

The *European Network for Better Establishment Statistics (ENBES)* is dedicated to improving cooperation on methodology, theory and applications within European establishment statistics. Establishment statistics is statistics about - and for - businesses and other forms of corporate units.

The ninth *European Establishment Statistics Workshop (EESW25)*, organized by ENBES and sponsored by IASS, was held in Rome from 5–7 November 2025, hosted by the Italian National Institute of Statistics (Istat). Under the theme *Innovations in Establishment Statistics with Special Attention to Data Collection*, the workshop brought together experts from National Statistical Institutes (NSI), academia, and international organizations to discuss innovations in establishment statistics, with a focus on data collection, integration, and modernization.

A first set of contributions addressed nonresponse, response burden, and data integration. Evidence confirmed that longer questionnaires increase perceived burden without substantial gains in output quality. Several papers demonstrated the value of integrating survey and administrative data, including large-scale data linkages and the use of fiscal sources. A key message was the need to move from NSI-centered to business-centered data collection, with processes and instruments aligned to business practices.

Administrative data and new collection methods were widely discussed. Tax data and e-invoicing systems were shown to improve timeliness and reduce burden, although issues related to definitions and availability persist. Innovative collection modes, such as live video interviews, reported positive effects on data quality and respondent engagement. The use of paradata provided insights into respondent behaviour and supported the development of more adaptive survey designs.

Innovation and automation were central themes throughout the workshop. Several contributions presented applications of machine learning and large language models for classification tasks, highlighting both their potential and the need for transparency and robustness. Adaptive approaches, such as Active Collection Management, showed improvements in response rates and cost-efficiency. Other works focused on editing and estimation, including automated systems for outlier treatment and improved robustness of estimates.

The final sessions focused on digital transformation and redesign. Contributions explored the integration of survey and web data, semantic interoperability, and machine-to-machine data transmission. Respondent-centered redesigns, such as modular questionnaires and improved web interfaces, demonstrated clear gains in data quality and usability.

Overall, three main directions emerged: a shift toward business-centered data collection; stronger integration of administrative, fiscal, and survey data; and increased use of technological innovation, including paradata, machine learning, and automated data transmission.

ENBES gratefully acknowledges Istat for hosting the workshop, as well as the support of ESRA, IASS, ONS, the Statistical Society of Slovenia, and Statistics Netherlands.

For more information about ENBES's activities, visit https://sites.google.com/enbes.org/home/home

Pasquale Papa,

Member of the ENBES Steering Committee

News from SAE 2026

The *Small Area Estimation, Survey and Data Science Conference 2026* (SAE 2026) was held at the University of Bucharest, Romania, from 15 to 19 June 2026. The conference brought together researchers, practitioners and experts from official statistics, academia and international organisations, providing an important forum for discussing recent methodological developments and applied challenges in small area estimation, survey statistics and data science.

SAE 2026 achieved truly global representation, bringing together 82 participants from 23 countries across 6 continents. The conference featured a comprehensive scientific program, comprised of 3 keynote lectures, 16 invited sessions, 6 contributed sessions, and 2 short courses, culminating in a total of 66 scientific presentations.

SAE 2026 participants, Bucharest, Romania

The programme was highly relevant for the survey statistics community, as it addressed key issues related to survey design, estimation, integration of data sources, model-based methods, uncertainty assessment and the production of reliable indicators for small domains and specific population groups. The contributions presented at the conference were collected in the Book of Abstracts, which documented the diversity of topics, methods, and applications discussed during the event. Additionally, the volume included a special write-up detailing the history of the SAE conference series.

The conference highlighted the growing importance of small area estimation within modern survey statistics. In a context where national statistical institutes and other data producers are increasingly required to provide timely, precise and geographically detailed information, survey statisticians face the challenge of making the best possible use of survey data, administrative records, geospatial information and other auxiliary sources. The sessions and discussions at SAE 2026 showed how small area estimation methods can contribute directly to improving the quality, relevance and policy usefulness of official statistics.

In particular, the keynote presentations addressed several important directions in contemporary SAE research and practice for the professionals interested in combining information to make reliable inference at subnational levels. Isabel Molina (talk `*Conciliation: the Key to Success in Small Area Estimation*') emphasized the growing importance of reconciling design-based and model-based approaches, arguing that their conciliation enables statisticians to effectively borrow strength across areas, integrate multiple data sources, and improve estimation for area means as well as more complex indicators such as poverty and inequality measures. Gauri Datta (talk `*A Bayesian Framework for Multi-Goals Small Area Inference: Estimation, Ranking and Benchmarking*') extended the discussion through a Bayesian framework that addressed multiple inferential goals simultaneously, including estimation, ranking, and benchmarking of small areas, while providing meaningful uncertainty quantification and incorporating benchmark constraints directly into the posterior distribution. His work demonstrates how Bayesian methods can support more accurate and stable decision-making for policy and resource allocation. Complementing these methodological advances, Cristina-Rodica Boboc (talk `*Measuring Skill Mismatch in the Romanian Labour Market: Are Small Area Estimation Methods a Solution*?') presented a practical application focused on skill mismatch in the Romanian labor market, where national statistics reveal persistent disparities between workers' qualifications and job requirements but fail to provide sufficiently detailed local insights. Her keynote explored whether SAE methods can overcome data limitations and generate reliable county-level estimates to better inform regional labor market policies. Together, these presentations illustrated how innovative SAE methodologies can both advance statistical theory and address pressing socioeconomic challenges through improved local-level evidence.

In addition to the scientific sessions, SAE 2026 included two short courses delivered by internationally recognised experts. The first one, entitled `*Entity Resolution*' was given on 15th of June by Ted Enamorado from Washington University in St. Louis, USA. The second one, entitled `*Bayesian Small Area Estimation*', with an emphasis on low- and middle-income countries was given on 19th of June by Jon Wakefield from the University of Washington, Seattle, USA. These courses offered participants the opportunity to deepen their knowledge on entity resolution and on small area estimation methods in demographic and health survey contexts. By combining theoretical foundations with practical applications, the training activities contributed to capacity building among survey statisticians and strengthened the link between research, official statistics and applied survey methodology.

The conference also featured the SAE Award ceremony, which celebrated outstanding contributions to the field and recognised Professor **Isabel Molina** (Complutense University of Madrid, Spain) as

the recipient of the **SAE 2026 Award**, further highlighting the strength and impact of the small area estimation community.

Professor Isabel Molina received the SAE 2026 Award from Partha Lahiri, the president of the IASS. Jon Rao is standing nearby. Congratulations, Professor Molina!

Beyond the scientific programme, SAE 2026 also offered several opportunities for informal exchange, networking, and cultural discovery. Participants had the opportunity to meet during informal dinners, the official conference dinner, the visit to the Palace of Parliament, and the walk through the Bucharest Botanical Garden. These social events created a warm and collegial atmosphere, allowing researchers, practitioners, and representatives of national statistical offices and international organisations to continue discussions beyond the conference sessions, strengthen professional connections, and experience Bucharest as a meeting place for the international SAE community.

SAE 2026 participants during dinner, Bucharest, Romania

As SAE 2026 comes to a close, we extend our sincere gratitude to the organizing committee, scientific committee volunteers, and all contributors whose dedication and hard work made this conference a tremendous success. We also gratefully acknowledge the support of all sponsors, cosponsors, institutional partners, and local organisers who contributed to the organisation and visibility of SAE 2026: the Faculty of Administration and Business of the University of Bucharest, which served as the local organiser and host institution of the conference, the International Association of Survey Statisticians (IASS), the International Association for Official Statistics (IAOS), the Federation of European National Statistical Societies (FENStatS), the Institute of National Economy of the Romanian Academy, and the Romanian National Institute of Statistics. Most importantly, we thank the international statistical community for its continued commitment to advancing methodology and its applications to real-world challenges. The vibrant exchange of knowledge at SAE 2026 demonstrates the strength and relevance of our field. We look forward to building on these connections and achievements in the year ahead and warmly welcome researchers, practitioners, students, and policymakers to join us at **SAE 2027** during 21-25 June 2027 in Hue, Vietnam for another exciting opportunity to learn, collaborate, and shape the future of small area estimation together. Safe travels, and we look forward to seeing you next year.

Ana Maria Ciuhu, co-chair of Local Organizing Committee

Andreea Erciulescu and **Domingo Morales Gonzalez,** co-chairs of Program Committee

Marius Stefan, member of Program Committee

Can machine learning effectively reduce potential nonresponse bias?

YES

David Haziza
University of Ottawa, Canada, dhaziza@uottawa.ca

Short answer: yes, it can — but with important caveats.

To start with, reducing non-response bias fundamentally relies on having auxiliary variables that are observed for all sampled units and that are related both to the outcome of interest and to the response mechanism. Under a missing-at-random type assumption, conditioning on such variables allows us, at least in principle, to remove non-response bias. Additional variables that are related to the outcome can also help improve efficiency; see, e.g., Little and Vartivarian (2005). From that perspective, machine learning methods can be very useful. Whether we are dealing with unit non-response or item non-response, the idea is similar. In the unit non-response case, we estimate response probabilities and construct adjusted weights. In the item non-response case, we impute missing values using predicted outcomes. In both settings, machine learning can help by providing flexible estimators of either the response model or the outcome model.

The main advantage of machine learning is its ability to capture complex relationships without requiring a fully specified parametric model. In practice, parametric approaches require us to get both the functional form and the set of covariates right, including interactions and nonlinearities. This is rarely realistic. Machine learning methods are much more robust to these kinds of misspecifications, which means they can, in principle, lead to better bias reduction — and sometimes improved efficiency as well.

However, there is no free lunch.

NO

Roderick J. Little
University of Michigan, Ann Arbor, USA, rlittle@umich.edu

Machine learning methods use training data to develop algorithms for prediction. Statistical methods also apply algorithms fitted to data, and can provide predictions, as when applied to imputation for nonresponse. I view machine learning as a form of statistical modeling, whereas computer scientists might consider statistical modeling as a form of machine learning. These semantic squabbles do little to advance science – although machine learners might pay more attention to, and reference, the extensive statistical literature on missing data.

Breiman (2001) advocated "black box" algorithmic approaches over more classical statistical models like linear or logistic regression (see also Little 2026, chapter 12). Since that paper, machine learning has come to be associated with algorithmic methods such as regression trees and forests, neural networks and gradient boosting. A more specific question is whether these methods can be used to reduce nonresponse bias. In their favor, these methods tend to be more flexible than classical models, and have the ability to allow for interactions that might otherwise be missed. There is some evidence that they produce better predictions than standard statistical models. I believe that mixing a variety of methods can be better than relying on one single approach, as in random forests rather than regression trees; I favor Bayesian forms of mixing as in BART (Chipman et al. 2010), because they give more weight to better-fitting models. On the other hand, machine learning algorithms are not easy to interpret, and underlying assumptions tend to be buried. Because of this, the methods tend to be viewed as "assumption-free", and hence always superior to classical statistical models. This is not true – the only universal approach is to avoid missing data! I think that

First, machine learning methods typically converge more slowly than parametric estimators, and therefore may require larger sample sizes to perform well. Second, and more importantly in the context of official statistics, point estimation is not enough. We also need valid variance estimators and, sometimes, confidence intervals. This is where things become much more challenging.

While recent developments such as double/debiased machine learning (Chernozhukov et al., 2018) provide promising solutions in some settings (for instance in the context of imputation for the treatment of item nonresponse), they often require rethinking the estimator itself in order to recover valid statistical inference.

The situation is even more complex in the case of unit non-response. Here, we construct adjusted weights that are meant to be used across all variables in the survey. Ensuring valid inference in that context, including variance estimation and asymptotic normality, remains a difficult problem, and in many cases an open research question; see Dagdoug and Haziza (2026) for a discussion.

Another practical issue is that machine learning methods require the user to choose an architecture and tuning parameters. In practice, default settings are often used, but for highly flexible methods this can be risky and may lead to unstable or biased estimators.

Finally, it is important to emphasize that machine learning does not solve the fundamental identification problem. If key variables related to both response and the outcome are not observed, no method — machine learning or otherwise — can fully eliminate non-response bias.

In summary, machine learning offers powerful tools to reduce non-response bias by better approximating complex relationships in the data. But it also introduces new challenges, particularly for statistical inference and practical implementation. Addressing these challenges remains an active area of research.

attention to the setting and underlying causal mechanisms of nonresponse can be important. To give two examples: (a) one application is to predict probabilities of response, which can be used in design to target units where nonresponse is high, or in analysis for nonresponse weighting. The "probability of response" depends on the set of predictors (Little 2021), and including auxiliary variables that are highly predictive of response but unrelated to survey outcomes is counterproductive, increasing variance without reducing bias (Little & Vartivarian 2005). Identifying such variables is context-dependent, and varies across the survey variable being imputed. Throwing the kitchen sink into a prediction algorithm ignores this issue. (b) More directly, machine-learning methods can be used to predict the missing values of survey variables. Here the fact that the algorithms are trained on existing data implies some form of missing at random (MAR, Rubin 1976) or very specific missing not at random (MNAR) mechanisms if response indicators are included as predictors (Little 2020; Fischer et al. 2026). MNAR models are chronically unidentified, and often best treated using a sensitivity analysis (Little & Rubin 2019, chapter 15).

Machine-learning algorithms focus on best predictions of the missing values; surveys concern statistical inferences for model parameters or finite population quantities. When missing values are replaced by best estimates, standard errors based on the filled-in data are underestimated. Also estimates of nonlinear functions of the data are biased. Multiple imputation (MI, Rubin 1987) imputes draws of the missing values from a predictive distribution. Estimates are obtained for each imputed data set and then averaged, rather than averaging the draws, the simulation analog of imputing best predictions. The MI approach is ingenious, less subject to bias for nonlinear statistics, and yields valid standard errors as well as point estimates. However, MI requires a statistical model to generate a predictive distribution for each missing value, which machine learning algorithms generally avoid. Machine learning approaches could be incorpo-

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261-265.

Dagdoug, M., & Haziza, D. (2026). Machine learning methods for finite population parameter estimation in survey sampling. *arXiv preprint arXiv:2604.01160*

Little, R. J., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means?. *Survey Methodology*, 31(2), 161.

rated into fully-specified statistical models, and MI applied to the predictive distributions of the missing values.

Finally, a comment about the question: it focuses on bias, assuming that a sizeable sample size make variance irrelevant; but the number of ultimate clusters in multistage sampling is often small, and for questions like subgroup analysis and small area estimation variance really matters. Methods that trade some bias for lower mean squared error, such hierarchical models, ridge regression, or the machine-learning methods discussed here, have become increasingly important. For me, these "biased" approaches constitute one of the seminal ideas in modern statistics (Little 2026, Chapter 7).

References

Breiman, L. (2001). Statistical modeling: two cultures. *Statist. Sci.* 16, 3, 199-231.

Fischer, M., Little, R.J. & West, B. T. (2026). Multiple imputation under missing not at random: incorporating response indicators into sequential imputation. In press, *Journal of Statistical Computation and Simulation*.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Statist.*, 4, 1, 266-298.

Little, R.J. (2021). A Note About the Definition of Propensity Weights. *Journal of Survey Statistics and Methodology.* 10 (4), 1098-1106.

Little, R.J. (2026). *Seminal Ideas and Controversies in Statistics*, Chapter 7. Baseball averages, foreign cars, and shrinkage estimation, and Chapter 12: Exploratory data analysis and data science. Chapman & Hall/CRC Press. Little, R.J. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition, Wiley Press.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Meth.*, 31, 161-168. Rubin, D.B. (1976). Inference and missing data (with discussion), *Biometrika*, 63, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley Press.

On the use of geospatial data in small area estimation

Nikos Tzavidis[1,a], **Luciano Perfetti Villa**[1,b], **Vasilis Chasiotis**[2]

[1]Department of Social Statistics and Demography & Southampton Statistical Sciences Research Institute, University of Southampton, UK
[2]Department of Statistics, Athens University of Economics and Business, Greece
[1,a]n.tzavidis@soton.ac.uk, [1,b]L.Perfetti-Villa@soton.ac.uk, [2]chasiotisv@aueb.gr

Abstract

In an era where emphasis is placed on reducing operating costs whilst increasing the timeliness and granularity of survey estimates, it is natural to think how to make best use of alternative data sources beyond traditional sources of population data, e.g., from censuses. The use of administrative data has been the focus of extensive theoretical and applied research, mainly in data rich contexts where administrative data are available. Census and administrative data are, however, not available or regularly updated in many parts of the world. In this short paper, we discuss the use of geospatial data as a source of auxiliary information in model-based small area estimation. Using geospatial data in small area estimation is not new. For example, the seminal paper by Battese, Harter, and Fuller, 1988 used geospatial covariates in the nested error regression model. The growing availability, easy access, global coverage, frequent updates, and quality of remote sensing data has led to renewed interest in their use in model-based estimation. Their use however is not without challenges. How to process geospatial data ready for estimation, how to best integrate survey, census, and geospatial data, how to specify and build a small area model, and how to use geospatial data for intercensal updating are areas of current research interest. We study these topics by reviewing relevant literature and using evidence from three applications with access to data of varying detail. Specifically, the first application presents an ideal case with access to the households' geolocations and to census microdata from both a recent and an older census. In the second application, census microdata are not available, while in the third application, we have access only to area-level information. In all applications, geospatial data are used as supplementary or the only source of auxiliary data. We present a summary of current research findings with an emphasis on providing guidance to practitioners. Although the focus of the applications is on the estimation of general parameters of income and poverty indicators, the conclusions have broader applicability.

Keywords: Alternative data; data integration; intercensal updating; poverty mapping; spatial scales.

1 Introduction

Model-based and model-assisted small area estimation is commonly implemented with the aid of statistical models and auxiliary information from population data, for example censuses. Increasing demand for estimates of general parameters means that analysts must increasingly rely on population-level, e.g., census micro-data. Relying on census data can be restrictive. Censuses are less frequent in many parts of the world and are usually only conducted every ten years. Using outdated population data in the intercensal period relies on strong assumptions about the distribution of the census variables over this period.

The increasing availability, improved quality, frequency, and coverage of remote sensing data raises the question of the role that such data sources can play in small area estimation. Chi et al. (2022) have

already developed a methodology that relies on data from alternative sources. Advances in the processing, global coverage, frequency, and free access of remote sensing data have created renewed interest in their use as a source of auxiliary information in small area models, e.g., Van der Weide et al. (2022), Edochie, Newhouse, Würz, et al. (2024), Newhouse et al. (2025). This is in contrast to data, for example, from mobile networks that are not easily accessible. Using geospatial data in small area estimation is challenging. By definition, zonal statistics, i.e., summary statistics of remote sensing data, are computed at an aggregate spatial scale, e.g., a grid cell or an administrative unit. Therefore, zonal statistics are contextual predictors and act only as proxies of household characteristics. The predictive ability of geospatial data depends on the type of outcome we are interested in predicting and is application specific. The spatial scale to be used for the integration of geospatial and survey data, model specification, and model-building are all topics of current research interest.

Although geospatial data have been successfully used in several applications in several countries, we cannot assume that the same geospatial variables and model specifications will be equally predictive in other countries. The types of geospatial data and their utility in countries that are data-rich are also the focus of current research. In an era where emphasis is placed on reducing operating costs, making best use of alternative data sources becomes of paramount importance. Experiences in data-scarce settings offer valuable lessons in data rich settings. Data-rich settings also offer the opportunity to compare geospatial-based estimates against what are considered to be ”gold standard” estimates that use census data. Last but not least, a compelling reason for using geospatial data is that they offer a natural approach to updating the estimates in off-census years.

In this short paper, we present current research evidence on the use of geospatial data in small area estimation. The evidence we present comes from reviewing the literature, and focusing on three applications, part of our research, with access to data of varying detail. Specifically, the first application presents an ideal case with access to the geolocations of sampled households and to census microdata from both a recent and an older census. In the second application, census microdata are not available, while in the third application, we had access only to area-level data. The three applications allow us to explore research questions around the following themes: (a) integrating survey and census with geospatial data, (b) model specifications and model building, and (c) intercensal updating. Although we do not present detailed results, we summarise the key findings with emphasis on providing guidance for practitioners. The remainder of the paper is structured as follows. In Section 2 we introduce geospatial data and describe how to go from raster data to constructing covariates (predictors) and how to integrate geospatial data with survey and census data depending on the information that is available about the geolocations of the units. Section 3 focuses on three model specifications namely, unit-level, unit-context and area-level models, and on variable selection and data quality issues when using geospatial data. In Section 3 we also summarise the evidence from the three applications. Finally, Section 4 summarises the main findings and offers guidance for practitioners.

2 Geospatial data and small area estimation

The term geospatial data is collectively used to indicate information derived from satellite imagery, remote sensing instruments, and other spatial data collection systems that provide measurements at high spatial resolution with near-global coverage. The initial input for geospatial products is satellite imagery, which captures the Earth’s surface across multiple spectral bands, from visible light to infrared and microwave frequencies. These multispectral images undergo a series of processing steps to produce geospatial indicators of interest. First, the spectral bands are decomposed and calibrated

to correct for atmospheric interference and sensor characteristics. Classification/machine learning algorithms are then used to detect specific features of interest, such as building footprints, land cover types, or crop extent. The spectral information can also be transformed into smoothed surfaces, for example, by measuring the intensity of nighttime light emissions or by computing vegetation indices from the ratio of near-infrared to visible light reflectance. The resulting measurements are typically distributed as raster layers, where each pixel stores the value of the derived indicator at a fixed spatial resolution. The resolution varies considerably across geospatial products and determines the finest spatial scale at which meaningful variation can be observed. For example, Landsat imagery is captured at 30 metres per pixel, VIIRS nighttime lights at approximately 500 metres, and climate reanalysis products such as ERA5 at roughly 30 km. Repositories such as Google Earth Engine, Microsoft Planetary Computer, and WorldPop provide access to pre-processed geospatial layers with global coverage. Aggregated values of these layers using different summary functions and spatial scales (e.g., administrative areas) form the so-called geospatial zonal statistics which are used as auxiliary information (predictors) in statistical/machine learning models.

Geospatial data sources have become increasingly accessible over the past two decades, both through publicly funded programmes such as the Copernicus and Landsat missions and through the efforts of private organisations such as Google and Meta (Merfeld et al., 2023). Using geospatial data in small area estimation is not a new idea. The application in the seminal paper by Battese, Harter, and Fuller, 1988 used remote sensing variables as predictors in a unit-level small area model. However, the increased availability of and easy access to geospatial data has opened new opportunities for statisticians working, among other subfields, in model-based small area estimation (SAE), particularly in settings where traditional sources of auxiliary data, such as census and administrative microdata, may be outdated, incomplete, or not accessible. The lack of recent censuses in several parts of the world is a key reason for using geospatial auxiliary data and machine learning algorithms (random forests) to produce population estimates for high-resolution grids (e.g., 100 $\times$ 100 m or 1 $\times$ 1 km grids) (Stevens et al., 2017), (`https://www.worldpop.org/wp-content/uploads/2022/10/top-down-tutorial.html`). Geospatial data are also a key source of auxiliary information for Meta's global estimates of average wealth at 2.4 km^2 resolution (Chi et al., 2022). Although such products have been largely developed outside the core of the small area literature, their popularity shows that the use of geospatial data in applications of small area estimation has utility for practitioners. Commonly used geospatial variables in SAE applications include building counts and density (e.g., from Google Open Buildings or Microsoft Building Footprints), nighttime light intensity from the VIIRS instrument, land cover classifications from MODIS, elevation and slope from digital elevation models, climate variables such as temperature and precipitation, and distance-based features derived from OpenStreetMap and other geographic databases. The range of potentially useful geospatial products continues to expand as new satellite missions, data providers, and processing methods become available.

Unlike other alternative data sources, e.g., mobile phone data, several sources of geospatial data are freely available, frequently updated and have global coverage. However, careful use of geospatial data as predictors in models is needed. This is because, depending on the application, geospatial data do not always provide direct measurements about the unit of measurement (e.g., the household) but only about the context within which the unit of measurement is located. Hence, geospatial data act as proxies for household characteristics. We return to this, in our view, important point later in this paper.

2.1 From rasters to covariates: Integrating survey and geospatial data

A key step in using geospatial data for SAE is creating zonal statistics by aggregating the raster values at pixel-level to spatial units of interest. The aggregation involves computing summary statistics, most commonly the weighted average, for all pixels that fall within the spatial unit. The weights correspond to the fraction of each pixel covered by the corresponding polygon, which ensures that the partial coverage of the boundary pixels is properly accounted for (Baston, 2023). Because survey and census data are typically georeferenced at administrative area boundaries, it is common for raster values to be aggregated at that level. However, this does not necessarily have to be the case. How we decide to aggregate the raster values also depends on the information available about the geolocation of the units in the survey data. Figure 1 illustrates this point by showing how to integrate household data with geospatial data with survey data depending on the information available about geolocations in the survey data. The plot on the left (original raster) shows the case where the households' geolocations are available (red points). In this case, an alternative to administrative-level aggregation is to compute zonal statistics within a buffer zone around each household's coordinates (blue circles). For example, one may calculate the average intensity of nighttime lights or the average density of buildings within a radius of 1 km for each household. This approach attempts to produce zonal statistics that are close to, albeit still contextual, unit-level geospatial data. The choice of buffer radius introduces a similar trade-off to the choice of grid resolution: smaller buffers capture more localized conditions but may be noisier, while larger buffers smooth the spatial variation. The plot on the right (aggregated to Enumeration Area (EA)) shows a more common case where the exact geolocation of the households is not available (red points are illustrative of household locations in EAs). A possible solution in this case is obtaining zonal statistics at the lowest possible spatial level available (e.g., EAs). All households within the same EA will be assigned the same covariate values. This results in a smoother version of the nightlights variable on the right hand side plot. We return to the point regarding the information on geolocations that needs to be available in survey and population data later in this paper when we discuss different model specifications.

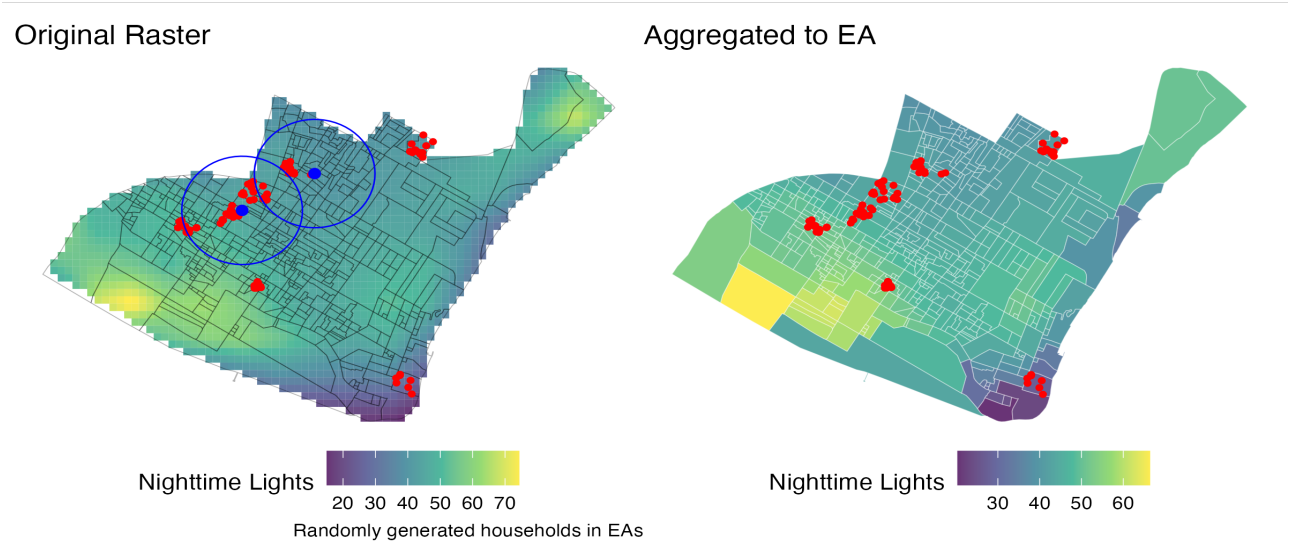


Figure 1: Nighttime Lights: Raster and aggregation to EA.

The choice of aggregation level for geospatial variables is a practical decision with methodological consequences because this will impact their subsequent use in models/algorithms for estimation/prediction purposes. We argue that if possible, analysts must be given full flexibility for producing zonal statistics. For example, analysts may decide to aggregate geospatial data to a variety of spatial supports, including regular grids of varying resolution (e.g., hexagonal or square grids at 1, 5, or 10 km^2) or administrative boundaries at different hierarchical levels (enumeration areas, sub-districts, or districts). This choice has consequences due to the Modifiable Areal Unit Problem (MAUP), which implies that the relationships between the geospatial predictors and the outcome may change depending on the spatial scale at which the predictors are measured (A. E. Gelfand, Zhu, and Carlin, 2001; Trevisani and A. Gelfand, 2013). Different geospatial variables may also exhibit different sensitivity to aggregation. For example, climate variables such as precipitation and temperature can vary considerably because the underlying spatial processes operate at different scales.

3 Model-based SAE with geospatial auxiliary information

The source of evidence presented in this section is current research on the use of geospatial data for small area estimation of income and poverty indicators. The estimation targets are general small area parameters including averages and poverty indicators (e.g., the mean consumption and the head count ratio). The model specifications below are presented in their simplest form, but, of course, several extensions to these specifications exist in the small area literature. We draw our evidence from three applications with access to data of varying detail. One application is from collaboration with the World Bank in Mozambique, the second application from collaboration with the Office for National Statistics in the UK, and the third application uses data from Greece, part of a wider collaboration between the World Bank and Eurostat on poverty mapping in Europe.

In the Mozambique application, we used data from the IOF 2022 household budget survey (15,382 households across 149 in-sample districts, with household geolocations available), the IOF 2019/20 survey (13,656 households, 154 districts), census microdata from the 2017 census (6,159,244 households with WorldPop 2022 population projections, 161 districts), 2007 census data (4.6 million households, 149 districts), and over 50 geospatial raster layers from WorldPop, Google Earth Engine, and Microsoft Planetary Computer. In the case of the UK application we used data from the Family Resources Survey (FRS), geospatial covariates aggregated at 100 $\times$ 100 m grids alongside administrative data such as benefits claimant counts, median house prices, and council tax bands, among others. In the application to data from Greece, we had access to area-level data from the EU-SILC survey for several years starting in 2010, geolocated information at the postcode level and access to corresponding geospatial data. The Mozambique application provided an ideal data scenario with census micro-data and household geolocations for processing raster data to various spatial scales. The availability of census data from 2007 allowed us to explore methods for intercensal updating. The UK data set-up was less favourable due to the lack of access to census microdata. However, unlike the case in Mozambique, administrative data were also available in this case. Finally, the application to data from Greece was the least favourable in terms of data access and exploring different model specifications.

3.1 The unit-level model

The unit-level nested error regression model, originally introduced by Battese, Harter, and Fuller (1988) and extended for estimating general parameters through an Empirical Best Predictor (EBP) by Molina and Rao (2010), is a widely used model specification for SAE. The ideal data scenario for

estimating general small area parameters assumes the availability of unit-level survey and population (e.g., census) microdata and the survey being conducted close to the census date. The predictors in the survey and census data are assumed to be measured using the same definitions. Under this data scenario, and assuming that the model has been built carefully, the unit-level model is a preferred method that has been shown to perform well. In practice, however, the ideal data scenario is rarely met because secondary analysts do not usually have access to census microdata and in the best case, censuses are only conducted every five to ten years.

The unit-level model specifies the relationship between a welfare variable y_{ij} (e.g., household income or consumption expenditure) for household j in area i and a vector of covariates $\boldsymbol{x}_{ij}$ as

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \tag{1}$$

where β is the vector of regression coefficients, and in the most common case, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ is an area-level random effect capturing unobserved heterogeneity across areas, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$ is the household-level error. A transformation of the response variable, $h(y_{ij})$, such as the logarithmic-shift or data-driven Box-Cox transformation (Rojas-Perilla et al., 2019), is used when the distributions of the model residuals do not follow those assumed under the model, as is typically the case when modelling the raw income or consumption variable. Model (1) is estimated using the survey data. The target quantities in poverty mapping applications include the head count ratio (HCR), the poverty gap and inequality measures. Molina and Rao (2010) proposed the EBP which works by generating simulated/synthetic populations under the model, computing the target parameter in each simulated population, and averaging over the simulation replications. As mentioned above, the method assumes that the predictors $\boldsymbol{x}_{ij}$ are also available for each household in the population. Geospatial predictors can also be brought in the model as contextual variables alongside census microdata. Contextual variables can help explain part of the unobserved heterogeneity between areas and provide a valuable data source that we believe analysts should explore also under the unit-level model framework.

In the application to data from Mozambique, the EBP with the 2017 unit-level census data and 2019/20 unit-level survey data served as the gold standard for model-based estimation. At the district level, the EBP estimates of mean consumption and HCR under this model showed high Spearman correlations with direct estimates at both the district and province levels. Comparison at the province level is used because direct estimates at this spatial scale have acceptable statistical reliability. When the 2007 census was used instead, assuming that there was no census in 2017, the correlations between model-based and direct estimates declined. In this case, model-based estimates of average consumption in certain areas were underestimated, which also led to higher model-based estimates of HCRs than the corresponding reliable direct estimates. This pattern was particularly evident in urban areas. We suspect that this is caused because the support of the covariates available from the 2007 census is different from that from the 2017 census. Therefore, current evidence suggests that the use of outdated census data must be used with caution because they can affect the quality of small area estimates. In the applications with the data from the UK and Greece, we had no access to census microdata so producing EBP under the unit-level model estimates was not possible.

3.2 The unit-context model

The use of remote sensing data as auxiliary information in SAE dates back to the foundational work of Battese, Harter, and Fuller (1988), who used crop classifications at pixel-level from land observatory satellites, aggregated to segment level, as covariates to predict the average hectares by crop type

for segments in Iowa counties. The unit-context model (see also Nguyen (2012)) is a nested error regression model used in settings where the predictors are not measured at the unit level but are available instead at an aggregate spatial unit g (such as a grid cell, enumeration area, sub-district, etc.) within area i. In the simplest case, the unit-context model is specified as

$$y_{igj} = \boldsymbol{x}_{ig}^T\boldsymbol{\beta} + u_i + e_{igj}, \tag{2}$$

where $\boldsymbol{x}_{ig}$ denotes the vector of predictors aggregated to spatial unit g, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ is a random effect at the area-level and $e_{igj} \overset{iid}{\sim} N(0, \sigma_e^2)$ is the error term at the household-level. Note that under model (2) all units within the same spatial unit g share the same covariate values $\boldsymbol{x}_{ig}$, while the outcome variable y_{igj} remains measured at the unit level. This model specification is not unique to the use of geospatial covariates. Similar model specifications have been used when the source of auxiliary information is administrative data but no linkage between the survey records and the administrative records exists. This unit-context specification has been adopted in a number of applications using geospatial data for poverty mapping, including Masaki et al. (2022), Newhouse et al. (2025), and Edochie, Newhouse, Tzavidis, et al. (2025). Small area estimation of general parameters under this model also requires access to population sizes for each spatial unit g for the purposes of constructing the appropriate number of predictions to be used for deriving the small area estimates. If a recent census is available, and information about the geolocations of the households in the census is available too, these population sizes are readily available. However, exact geolocated information for all population units is rarely available. If access to such data is not possible also because there is no recent census, the population sizes can be approximated by using population projections from gridded population data such as the data produced by the WorldPop team at the University of Southampton. This situation is not unique to the use of geospatial data under the unit-context model. Deriving EBP estimates of general parameters under the unit level model also requires access to the population sizes for the target areas. However, getting access to the population sizes for higher level administrative units may be more straightforward than gaining access to population sizes for smaller spatial units.

Compared to the unit-level model, the use of the unit-context model has been criticized see e.g. Corral, Henderson, and Segovia (2025). Because all households within the same spatial unit share the same covariate values and in poverty mapping applications geospatial variables act only as proxies for household-level characteristics, it is reasonable to expect that the predictive ability of the unit-context model will not be as good as that of the unit-level model, possibly leading to higher unexplained variance and to less precise small area estimates. Despite this, the unit-context model can be a valuable tool for practitioners working in data limited settings. Its appeal lies in its applicability when census microdata are outdated and/or not accessible. As we outline below, our current view about the performance of the unit-context model is that it is application/context specific, and as a result, this model specification must be used with great caution.

In the case of the application in Mozambique, estimates under the unit-context model, with LASSO-selected geospatial covariates, achieved high Spearman correlation coefficients with direct estimates at the district and province levels for both small area means and HCRs. As expected, the predictive power of the unit-context model was lower than that of the unit level model and the corresponding MSEs generally larger than the MSEs of EBP estimates under the unit-level model. However, this was the case under the ideal data scenario for the unit-level EBP, i.e., when recent census data are available. When the estimates under the unit-context model were compared with the unit-level EBP estimates produced with the 2007 census data, the unit-context estimates were preferable. These

results indicate that in the case of Mozambique, carefully-selected geospatial covariates can provide reliable estimates at the district and province levels, and represent a clear improvement over the use of outdated census data. The aggregation level at which geospatial predictors are computed impacts on model performance. In Mozambique, geospatial predictors aggregated at coarser administrative levels (district, Posto) yielded estimates with lower RMSEs compared to the use of geospatial predictors at high-resolution hexagonal grids. This suggests that aggregation to finer grids introduces noise without improving predictive power. Using variable selection, e.g., via LASSO, is recommended when many geospatial layers are available, as it identifies which variables are important at each spatial resolution and addresses multicollinearity among correlated raster-derived covariates.

The above conclusions change when assessing the results from using the unit-context model in the application with the UK data. In this case, the unit-context model using geospatial covariates aggregated at 100 $\times$ 100 m grids performed poorly when used to estimate income-type measures for UK Local Authority Districts. In the UK, the rural/urban income difference was not explained well by geospatial variables. Our conjecture is that in developed countries the spatial variation in income is driven by socioeconomic factors (e.g., employment, benefits, housing costs) that are not captured well by remote sensing data alone. Including administrative data at low spatial scale improved the fit of the unit-context model, but the small area estimates remained unsatisfactory when compared to reliable direct estimates. In the application to data from Greece, we did not use the unit-context model because we worked only with area-level data.

3.3 The area-level model

The Fay-Herriot model (Fay and Herriot, 1979) is the standard area-level model in SAE. Unlike the unit-level and unit-context models which operate at the household level, the Fay-Herriot model is specified directly with area-level aggregates. The Fay-Herriot model is well suited for use with geospatial data because predictors, $\boldsymbol{x}_i$, are naturally linked to direct estimates at the area level. It comprises two parts. The first part is the sampling model which describes the relationship between the direct estimates $\hat{\theta}_i^{Dir}$ of the target area parameter with the true values θ_i,

$$\hat{\theta}_i^{Dir} = \theta_i + e_i, \quad e_i \overset{ind}{\sim} N(0, \sigma_{e_i}^2), \tag{3}$$

where e_i denotes the sampling error with variance $\sigma_{e_i}^2$, estimated under the sampling design. The second part is the linking model that relates the true area parameter to area-level covariates $\boldsymbol{x}_i$,

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2), \tag{4}$$

where u_i is the area-level random effect. Estimation of the Fay-Herriot model assumes that the variances of the direct estimates are known. The estimation of the sampling variances under the design is important because this determines the weight placed on the direct and the regression synthetic estimators used to construct the EBLUP.

Transformed versions of the Fay-Herriot model have been proposed. The use of a logarithmic transformation has been studied for example by Slud and Maiti (2006). When the target parameter is a proportion, as is the case for example with the HCR, the use of transformations can be crucial especially when direct estimates are close to the boundary of the support of θ_i. For proportions, the arcsin transformation provides a transformation that can stabilise the variance (Casas-Cordero Valencia, Encina, and Lahiri, 2016). Under this transformation, direct estimates are transformed

$\hat{\theta}_i^{Dir,arcsin} = \sin^{-1}\left(\sqrt{\hat{\theta}_i^{Dir}}\right)$, and the sampling variance, approximated using a Taylor series expansion, is given by $1/(4\tilde{n}_i)$, where $\tilde{n}_i$ is the effective sample size. The Fay-Herriot model is estimated on the transformed scale, and a back-transformation based on the conditional expectation under the model can be employed to account for the non-linearity of the inverse transformation (Sugasawa and Kubokawa, 2017). Other transformations for proportions, including the logit, probit, and complementary log-log, have also been considered in the literature. Jacob et al. (2025) provided a detailed comparison of these alternatives in an application to producing quarterly estimates of extreme poverty rates in Brazil. An alternative framework for area-level modelling of proportions is offered by the extended Beta model proposed by De Nicolò, Fabrizi, and Gardini (2024). Under this approach, the area-level proportions are modelled directly by using a beta distribution, which naturally accounts for the bounded support of the response. The Beta model can be extended to include zero and one-inflated components, which is important in poverty mapping applications, and spatial random effects. This approach has been recently used in a poverty mapping application with remote sensing data, see De Nicolò, Fabrizi, and Gardini (2024). The `tipsae` R package (De Nicolò and Gardini, 2024) provides a Bayesian implementation of the extended Beta models.

In the case of the application in Mozambique, a Fay-Herriot model that uses the arcsin transformation and geospatial predictors aggregated at the district level achieved high correlations with direct estimates at the district and province levels for both small area means and HCRs. These model-based estimates were also comparable to the model-based estimates produced with the Fay-Herriot model that used aggregate predictors from the 2017 census. What was interesting in this case is that using a Fay-Herriot model with district-level predictors coming from the 2007 census produced estimates that were also well correlated with direct estimates. This indicates that area-level aggregation in this case mitigated the effect of possible structural changes in census covariates that, as we described above, can affect the performance of the EBP estimates produced with the unit-level model.

The use of a Fay-Herriot model in the application to data from the UK produced estimates that were substantially better than those under the unit-context model. Combining geospatial covariates with administrative data sources (benefits claimant counts, median house prices, council tax bands, among others) proved to be an effective strategy. This suggests that in some contexts the Fay-Herriot framework can be more effective than unit-context models especially when geospatial data are supplemented with data from other data sources, e.g., administrative data. Initial results from using a Fay-Herriot model with geospatial predictors to data from Greece showed that the resulting estimates are of acceptable quality. Overall, our research evidence so far suggests that area-level models combined with geospatial predictors perform well across applications.

3.4 Variable selection and data quality

When many geospatial layers are available, as is increasingly common with repositories providing several raster products, variable (feature) selection becomes essential. High correlation between geospatial variables can lead to multicollinearity. For example, nighttime lights, road density, and building counts capture aspects of human settlement and economic activity. In addition, as mentioned before, rasters can be aggregated at different spatial scales and by using different summary functions. Regularization methods such as LASSO (Tibshirani, 1996) or elastic net (Zou and Hastie, 2005) provide principled approaches for selecting a parsimonious set of geospatial predictors while addressing multicollinearity. Although moderate multicollinearity may not severely affect SAE point estimates (Merfeld et al., 2023), it can cause numerical instability in model fitting.

Data quality issues also deserve the attention of practitioners. Cloud coverage can introduce noise or missing values in optical satellite imagery. The South Atlantic Anomaly affects the operation of certain satellite instruments, leading to gaps in data coverage over parts of southern Africa and South America. In some cases, specific geospatial products may have incomplete spatial coverage. For example, in the application in Mozambique, we found that Google Open Buildings V3 has missing building footprints in certain regions of the country. This problem was addressed by combining multiple data sources, for example, by combining Google and Microsoft building footprint counts where both are available. Awareness of these limitations is important for practitioners, as the quality of geospatial covariates directly affects the quality of the resulting SAE estimates.

4 Concluding remarks and guidance for practitioners

How to best use geospatial data in model-based small area estimation is application specific. The research evidence from the applications in Mozambique, the UK, and Greece offers valuable insights. In this concluding section, we summarize the main findings.

Unit-level models: Using a well-specified unit-level model and micro-data from a recent census as the source of auxiliary information is the preferred approach for model-based small area estimation. Geospatial data can complement census predictors to assist with reducing the unexplained between-area variability. The main limitation in this case is that the method depends on the availability of population micro-data from a recent census. As illustrated in the case of Mozambique, the use of outdated census data can be risky. Supplementing outdated census data with geospatial data does not guarantee a solution to this problem because census predictors can remain important to achieve a good model fit.

Unit-context models: This model specification requires survey data, the geolocations of the sample units at the chosen spatial scale, population projections at the same spatial scale (if counts from a recent census at the same spatial scale are not available), and corresponding geospatial predictors, making it applicable in settings where recent census microdata are unavailable. How the geospatial predictors are aggregated depends on what information about the geolocations of the sampling units is available, and the type of geospatial variable. As is the case with every model, model-building in the case of the unit-context model requires carefully considering how to include geospatial predictors at different scales. In Mozambique, the use of a unit-context model produced small area estimates of acceptable quality. A finding of practical relevance in this case is that using coarser aggregation levels for geospatial predictors was preferred to using finer-resolution grids. However, the use of the unit-context model in the UK application failed to produce estimates of acceptable quality. This shows that using the unit-context model must be done with great caution.

Area-level models: Area-level models offer a natural framework for integrating survey and geospatial data. Area-level models are deceptively simple. Model specification requires careful consideration of the survey design for point and variance estimation, possible use of variance smoothing methods, and advanced models. Using area-level models with geospatial predictors performed consistently well across the applications we have considered in this paper.

Intercensal updating with geospatial data: Appealing features of geospatial data include global coverage, ease of access, and continuous updating. The last feature is a key reason that has been used to promote the use of geospatial data in SAE applications. Economic development, urbanisation, migration, and policy changes can change the distribution of household characteristics during the intercensal period. A common practical challenge is that, depending on the context, census data

can quickly become outdated. Using outdated census data as predictors relies on the assumption that the relationship between the outcome and the predictors remains stable over time. The data setup in Mozambique allowed us to study the impact of using outdated census data in small area estimation. Comparisons of data from the 2007 and 2017 Mozambique censuses revealed changes in key housing characteristics, including the type of energy source used by households and variables describing the housing quality. These changes reflect a decade of economic development. Using outdated 2007 census data in a unit-level model affected the quality of model-based small area estimates. Instead, the use of a unit-context or an area-level model with geospatial predictors resulted in estimates of acceptable quality. For practitioners working in intercensal periods and without access to population micro-data, these results indicate that geospatial predictors as the source of auxiliary information can be effective. However, a cautious approach to how the model is specified is needed. If analysts are keen to continue using census data, SPREE-type methods (Isidro, Haslett, and Jones, 2016; Luna Hernandez, 2016; Zhang and Chambers, 2004) offer an alternative approach to intercensal updating. Current research focuses on different estimation strategies in the off-census years and shows that SPREE-type methods can be also effective for intercensal updating. Whichever model specification is selected, carefully validating model-based estimates against reliable model-free estimates remains of critical importance.

Software: Processing geospatial data is perhaps the most challenging part for statisticians without access to GIS expertise. Outside specialised GIS software, preprocessing steps to prepare geospatial data for estimation are performed either by using R or Python, both of which now offer mature toolchains for raster and vector operations, as well as access to remote-sensing data. In R, `exactextractr` (Baston, 2023) provides fast zonal statistics for arbitrary polygons, `blackmarbler` (Marty and Stefanini Vicente, 2024) retrieves and processes NASA Black Marble nighttime lights products, and `rgee` (Aybar et al., 2020) exposes the Google Earth Engine catalogue (Gorelick et al., 2017) for Landsat, Sentinel and VIIRS imagery. The recently released `GeoLink` package (Llyod et al., 2025), jointly developed by the World Bank Group and the University of Southampton, provides a higher-level interface that automates the linking of standard geospatial layers, e.g., nighttime lights, building footprints, climate, land cover, among others, to georeferenced survey microdata or shapefiles, considerably reducing the boilerplate involved in assembling SAE-ready covariate sets. Equivalent functionality is available in Python through `rasterio`, `rasterstats` and `geopandas` for raster and vector handling, and through the `earthengine-api` client for Earth Engine access. For small area estimation, the most established implementations of the models discussed here are in R, although Python alternatives are emerging. Among other packages, the `emdi` package (Kreutzmann et al., 2019) provides a unified interface for the area-level and unit-level models, including MSE estimation, as is the `sae` package (Molina and Marhuenda, 2015). The `povmap` package (Edochie, Newhouse, Würz, et al., 2024) extends `emdi` with features specifically tailored to poverty mapping, including a survey-weighted version of EBP, ex-post benchmarking, and a wrapper for use via Stata. For Beta-type models the `tipsae` package (De Nicolò and Gardini, 2024) is the natural choice.

References

Aybar, C., Q. Wu, L. Bautista, R. Yali, and A. Barja (2020). rgee: An R Package for Interacting with Google Earth Engine. In: *Journal of Open Source Software* 5.51, p. 2272. DOI: 10.21105/joss.02272.

Baston, D. (2023). *exactextractr: Fast Extraction of Raster Values Within Polygons*. R package version 0.9.1. URL: https://CRAN.R-project.org/package=exactextractr.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. In: *Journal of the American Statistical Association* 83.401, pp. 28–36. DOI: 10.1080/01621459.1988.10478561.

Casas-Cordero Valencia, C., J. Encina, and P. Lahiri (2016). Poverty Mapping for the Chilean Comunas. In: *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons, Ltd. Chap. 20, pp. 379–404. ISBN: 9781118814963. DOI: https://doi.org/10.1002/9781118814963.ch20. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118814963.ch20. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118814963.ch20.

Chi, G., H. Fang, S. Chatterjee, and E. Blumenstock J. (2022). Microestimates of wealth for all low and middle income countries. In: *PNAS* 119.3, pp. 381–399. DOI: 10.1073/pnas.2113658119.

Corral, P., H. Henderson, and S. Segovia (2025). Poverty Mapping in the Age of Machine Learning. In: *Journal of Development Economics* 172, p. 103377. DOI: 10.1016/j.jdeveco.2024.103377.

De Nicolò, S., E. Fabrizi, and A. Gardini (2024). Extended Beta Models for Poverty Mapping. An Application Integrating Survey and Remote Sensing Data in Bangladesh. In: *The Annals of Applied Statistics* 18.4. DOI: 10.1214/24-AOAS1934.

De Nicolò, S. and A. Gardini (2024). The R Package tipsae: Tools for Mapping Proportions and Indicators on the Unit Interval. In: *Journal of Statistical Software* 108.1, pp. 1–36. DOI: 10.18637/jss.v108.i01.

Edochie, I., D. Newhouse, N. Tzavidis, T. Schmid, E. Foster, A. L. Hernandez, A. Ouedraogo, A. Sanoh, and A. Savadogo (2025). Small Area Estimation of Poverty in Four West African Countries by Integrating Survey and Geospatial Data. In: *Journal of Official Statistics* 41.1, pp. 96–124. DOI: 10.1177/0282423X241284890.

Edochie, I., D. Newhouse, N. Würz, and T. Schmid (2024). *povmap: Extension to the emdi Package*. R package version 1.0.1. DOI: 10.32614/CRAN.package.povmap. URL: https://github.com/SSA-Statistical-Team-Projects/povmap.

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* 74.366, pp. 269–277. DOI: 10.2307/2286322.

Gelfand, A. E., L. Zhu, and B. P. Carlin (2001). On the Change of Support Problem for Spatio-Temporal Data. In: *Biostatistics* 2.1, pp. 31–45. DOI: 10.1093/biostatistics/2.1.31.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. In: *Remote Sensing of Environment* 202, pp. 18–27. DOI: 10.1016/j.rse.2017.06.031.

Isidro, M., S. Haslett, and G. Jones (2016). Extended Structure Preserving Estimation (ESPREE) for Updating Small Area Estimates of Poverty. In: *Annals of Applied Statistics* 10.1, pp. 451–76. DOI: 10.1214/15-AOAS900.

Jacob, G. A. P., N. Tzavidis, A. Luna Hernandez, and P. L. D. N. Silva (2025). Quarterly Small Area Estimates of Extreme Poverty in Brazil Using Transformed Fay-Herriot Models. In: *Journal of Survey Statistics and Methodology* 13.5, pp. 552–586. DOI: 10.1093/jssam/smaf025.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators. In: *Journal of Statistical Software* 91.7, pp. 1–33. DOI: 10.18637/jss.v091.i07.

Llyod, C. T., I. Edochie, D. Jaganjac, J. D. Merfeld, D. Newhouse, L. Perfetti-Villa, D. A. Gomez, and N. Tzavidis (2025). *GeoLink R Package*. DOI: 10.5281/zenodo.15231144. URL: https://doi.org/10.5281/zenodo.15231144.

Luna Hernandez, A. (2016). "Multivariate Structure Preserving Estimation for Population Compositions". PhD thesis. University of Southampton. 127 pp.

Marty, R. and G. Stefanini Vicente (2024). *blackmarbler: Black Marble Data and Statistics*. R package. URL: https://CRAN.R-project.org/package=blackmarbler.

Masaki, T., D. Newhouse, A. R. Silwal, A. Bedada, and R. Engstrom (2022). Small Area Estimation of Non-Monetary Poverty with Geospatial Data. In: *Statistical Journal of the IAOS* 38.3, pp. 1035–1051. DOI: 10.3233/SJI-210902.

Merfeld, J. D., H. Chen, P. Lahiri, and D. Newhouse (2023). Small Area Estimation with Geospatial Data: A Primer. Working Paper.

Molina, I. and Y. Marhuenda (2015). sae: An R Package for Small Area Estimation. In: *The R Journal* 7.1, pp. 81–98. DOI: 10.32614/RJ-2015-007.

Molina, I. and J. N. K. Rao (2010). Small Area Estimation of Poverty Indicators. In: *Canadian Journal of Statistics* 38.3, pp. 369–385. DOI: 10.1002/cjs.10051.

Newhouse, D., A. Ramakrishnan, T. Swartz, J. Merfeld, and P. Lahiri (2025). Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. In: *Oxford Bulletin of Economics and Statistics*. DOI: 10.1111/obes.12678.

Nguyen, V. C. (Dec. 2012). A Method to Update Poverty Maps. In: *Journal of Development Studies* 48.12, pp. 1844–1863. DOI: 10.1080/00220388.2012.682983.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2019). Data-Driven Transformations in Small Area Estimation. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 183.1, pp. 121–148. DOI: 10.1111/rssa.12488.

Slud, E. V. and T. Maiti (2006). Mean-Squared Error Estimation in Transformed Fay–Herriot Models. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.2, pp. 239–257. DOI: 10.1111/j.1467-9868.2006.00542.x.

Stevens, F., A. Gaughan, C. Linard, and A. Tatem (2017). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. In: *PLoS ONE* 10(2): e0107042. DOI: 10.1371/journal.pone.0107042.

Sugasawa, S. and T. Kubokawa (2017). Transforming Response Values in Small Area Prediction. In: *Computational Statistics & Data Analysis* 114, pp. 47–60. DOI: 10.1016/j.csda.2017.03.017.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Trevisani, M. and A. Gelfand (2013). "Spatial Misalignment Models for Small Area Estimation: A Simulation Study". In: *Advances in Theoretical and Applied Statistics*. Studies in Theoretical and Applied Statistics. Berlin, Heidelberg: Springer, pp. 269–279. DOI: 10.1007/978-3-642-35588-2_25.

Van der Weide, R., B. Blankespoor, C. Elbers, and P. Lanjouw (2022). How Accurate Is a Poverty Map Based on Remote Sensing Data? An Application to Malawi. World Bank Policy Research Working Paper. URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/b6d07e8a-7a03-58fe-b8f1-2ac86fa12011/content.

Zhang, L.-C. and R. Chambers (2004). Small area estimates for cross□classifications. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.2, pp. 479–496. DOI: 10.1111/j.1369-7412.2004.05266.x.

Zou, H. and T. Hastie (2005). Regularization and Variable Selection Via the Elastic Net. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320.

Data Privacy in the AI Era and Implications for Survey Practice[1]

Shurong Lin[1] **and Aleksandra Slavković**[2]

Department of Statistics, The Pennsylvania State University, USA
[1]shurong@psu.edu, [2]sesa@psu.edu

Abstract

Balancing data confidentiality and data utility is a long-standing challenge in survey and official statistics. In modern data ecosystems, this challenge has intensified due to increased data availability, advances in computational methods, and now the widespread use of artificial intelligence (AI). The rise of AI does not create the familiar privacy–utility tradeoff anew, but fundamentally reshapes the information environment in which it must be understood. Protected survey data can now be reused, linked, and modeled more intensively than before, while AI-assisted tools are transforming survey data collection, processing, and analysis. In this paper, we present a perspective grounded in statistical data privacy (SDP), an integrative framework that connects traditional statistical disclosure control with formal privacy approaches such as differential privacy. We emphasize that privacy-preserving mechanisms alter released data or summaries by introducing additional sources of uncertainty that must be explicitly modeled to support valid downstream analysis and statistical inference. This perspective is important in the AI era, especially for surveys where complex design and processing features already shape inference. AI further reshapes the privacy–utility tradeoff by enabling adaptive reuse and accelerating the accumulation of privacy risk. This reinforces the need for methodological transparency and alignment of privacy choices with the policy goals and intended uses of survey data.

Keywords: Artificial intelligence (AI), statistical data privacy, differential privacy, statistical disclosure control, survey statistics, official statistics.

1 Introduction

Statistical agencies, survey organizations, and data stewards have long operated under a dual mandate: to protect the confidentiality of respondents while ensuring that released data remain sufficiently informative to support statistical inference, scientific discovery, and policy. This balance is foundational to the mission of official statistics and has guided decades of methodological development (e.g., see Hundepool et al., 2012; Skinner, 2009; Matthews and Harel, 2011). The tension between minimizing disclosure risk and maximizing data utility has traditionally been framed as a *risk–utility tradeoff* (Duncan, Keller-McNulty, and Stokes, 2001). In this article, we adopt the broader term *privacy–utility tradeoff* and use the terms *privacy* and *confidentiality* interchangeably to refer to protection against the disclosure of sensitive information by a third party.[2]

The privacy-utility tradeoff is especially difficult to attain in survey settings. Survey products are often intended to support finite-population inference under complex designs, unequal weighting, nonresponse adjustment, imputation, and other processing steps (Groves, F. J. Fowler, et al., 2011; Särndal, Swensson, and Wretman, 1992; Lohr, 2021). For decades, the survey community has worked

[1]This paper draws on the 2025 Links Lecture by Slavković: *PrAIvacy (noun) Pronounced: /'prā-və-sē/ Balancing data confidentiality and utility* (https://www.amstatleads.org/events).

[2]Informally, privacy is the right of individuals to control the dissemination of, or access to, information about themselves. Confidentiality is the dissemination of data without public identification (i.e., no disclosure) by a third party.

with a broad collection of methods from statistical disclosure control (SDC), also referred to as statistical disclosure limitation (Hundepool et al., 2012; Matthews and Harel, 2011). More recently, formal privacy frameworks such as differential privacy (DP) (Dwork, McSherry, et al., 2006; Dwork and Roth, 2014) have received attention. A major example in official statistics is the adoption of DP for the 2020 U.S. Decennial Census (J. M. Abowd, 2018; J. Abowd et al., 2022). We do not review the formal definition of DP here; for a recent overview in this venue, see Charest and Drechsler, 2026. We use the term *statistical data privacy* (SDP) to refer to methods and frameworks for protecting confidential information in data release and analysis, including both SDC and DP (Slavković and Seeman, 2023).

Traditionally, sample surveys were designed to produce tabular estimates from a single dataset. Early total survey error frameworks focused primarily on measurement and representation errors and did not explicitly address non-sampling errors induced by SDC (Groves, J. Fowler F. J., et al., 2004). Later survey quality frameworks acknowledged that disclosure-control procedures can affect data quality, but did not fully address their implications for valid statistical inference (Groves and Lyberg, 2010). Over the past two decades, however, survey practice has shifted toward statistically valid estimation based on the integration of multiple data sources, with such combined data products becoming increasingly common. At the same time, the heightened disclosure risks inherent in these integrated settings have not received comparable attention, nor has the role of privacy-induced errors in shaping valid statistical inference.

In today's data ecosystem, maintaining the privacy–utility balance has become increasingly difficult. Data are more abundant, interconnected, and reusable than ever before. Advances in computational power and the availability of auxiliary data have made it possible to combine information across sources, often in ways not anticipated at the time of collection. Early work showed that even aggregate data releases could enable the reconstruction of sensitive information (e.g., Dinur and Nissim, 2003), and more recent work has demonstrated that machine learning models can leak training data or enable inference attacks (e.g., Shokri et al., 2017; Carlini et al., 2021).

Artificial intelligence (AI) has amplified both the opportunities and the risks associated with data. Here, AI refers broadly to machine-learning systems that automate, augment, or support analytical and decision-making tasks, including generative models, large language models, and tools for prediction, classification, linkage, and data generation. On one hand, AI enables richer analysis, improved prediction, and new forms of data integration. On the other hand, it challenges traditional assumptions about how data are used, reused, and shared. AI changes the information environment in which the privacy-utility tradeoff must be understood. Privacy can no longer be viewed as a static property of a dataset; instead, it must be understood as a dynamic feature of a broader data ecosystem.

In survey settings, AI-driven systems can affect multiple stages of the survey pipeline, from design and respondent interaction to processing, analysis, reporting, and synthetic data generation (Rothschild et al., 2025). They can also search across heterogeneous sources and exploit auxiliary information at scale. Released data products, including synthetic data, can therefore be reused, linked, and modeled far more intensively than before (Kapania et al., 2025). As a result, privacy depends not only on whether a released product satisfies a disclosure rule at the time of release, but also on how it is used in downstream analyses and combined with external data. Privacy risk may therefore accumulate and evolve over time. In the AI era, overlooking amplified disclosure risks in combined-data settings becomes especially consequential.

This paper approaches these challenges through the lens of statistical data privacy (SDP), an integrative framework that connects traditional SDC with formal privacy approaches such as DP. SDP is particularly well suited to survey statistics because it emphasizes uncertainty quantification, inferen-

tial validity, and transparency. From a statistical perspective, privacy protection methodology should consider both the disclosure-control aspect and the quality of inference (J. M. Abowd and Schmutte, 2015; Slavković and Seeman, 2023; Awan and Gong, 2024). Privacy-preserving methods and mechanisms, whether based on SDC or DP frameworks, typically alter released data or summaries by introducing additional sources of uncertainty that must be explicitly modeled to support valid downstream analyses. Once confidentiality protection alters released data, summaries, or estimators, standard inferential procedures based on the protected data no longer automatically retain their validity. Privacy protection thus becomes an integral component of the broader error and uncertainty structure that shapes survey inference.

The SDP perspective is valuable for several reasons. First, it provides a common statistical language for understanding both traditional disclosure limitation and modern formal privacy methods. Second, it aligns naturally with the survey community's longstanding focus on multiple sources of error, including sampling variation, nonresponse, weighting, imputation, and linkage. Third, it highlights how AI increases the scale of data reuse and the adaptivity of downstream analysis, making the inferential consequences of privacy protection more important, not less.

In the remainder of the paper, we outline how the challenges we face today are rooted in a long history of statistical practice. We propose that the survey community should treat privacy protection as an essential component of the statistical pipeline through which data are produced, released, and analyzed. In the AI era, this perspective allows us not only to revisit the classic privacy–utility tension, but also to ask a more informative question: under what forms of protection can released data support trustworthy analysis, for which inferential goals, and with what accounting of uncertainty?

2 Privacy in the Era of AI

2.1 A Bit of a Historical Reflection

In thinking about how best to quantify the data privacy–utility tradeoff, we can view AI as an amplifier that stress-tests prior paradigms. The challenges we face today are rooted in a long history of statistical practice. The timeline in Table 1 highlights more than 70 years of intellectual and practical development, particularly within the U.S. federal statistical system, as it has responded to growing data demands. It also illustrates how AI exposes and challenges the assumptions and protections associated with each stage.

Historically, privacy benefited from structural constraints. Data demand was relatively limited, data systems were slow and fragmented, and computation was expensive. Early statistical confidentiality practices often relied implicitly on these constraints for protection. Prior to the 1960s, agencies developed tabulation rules and heuristics to prevent disclosure. These approaches were effective in an environment with limited access and restricted analytical capabilities.

The 1960s marked a turning point with advances in computing, which enabled the development of more systematic statistical disclosure control (SDC) methods. Although data access remained restricted and large-scale linkage across sources was still difficult in practice, increasing computational capacity began to expose vulnerabilities in earlier approaches. Methods such as data swapping and complementary cell suppression addressed the needs of the time, but their outputs became more susceptible to linkage attacks (e.g., Sweeney, 2002) and database reconstruction (Dinur and Nissim, 2003) as computational capabilities improved. Holan et al., 2010 and Webb et al., 2026, among others, also demonstrate strong deficiencies in privacy protection with cell suppression.

As data sharing expanded and computational resources grew, SDC methods evolved to address a setting in which more granular data were requested, auxiliary information became more widely available, and record linkage and inferential reconstruction became increasingly feasible (e.g., Homer et al., 2008; A. Slavkovic and Lee, 2010; Hundepool et al., 2012; Gymrek et al., 2013). The emergence of model-based approaches marked an important shift. Synthetic data and multiple imputation introduced the idea that confidentiality protection could be integrated with statistical modeling (Rubin, 1993; Reiter, 2005), aiming to preserve key statistical properties while reducing disclosure risk. These approaches were later extended to large-scale applications (Kinney et al., 2011).

With the digital age, the growth of data and the increasing availability of diverse forms of public data beyond agency releases increased re-identification risks (Narayanan and Shmatikov, 2008; Dwork, Smith, et al., 2017). Differential privacy represents a subsequent evolution of statistical data privacy, providing a formal framework for bounding disclosure risk through controlled randomness (Dwork, McSherry, et al., 2006). It builds on earlier ideas such as noise injection, while introducing *transparency, composability, and formal guarantees* (Dwork and Roth, 2014). Over the past two decades, differential privacy has become an influential framework, in part because intuitive or ad hoc disclosure-control rules appeared increasingly fragile in data-rich environments with abundant side information and powerful computation. It has also seen large-scale implementation in official statistics, including at the U.S. Census Bureau (J. M. Abowd, Ashmead, et al., 2022) as well as in industry settings such as Apple and Google (Apple, 2017; Erlingsson, Pihur, and Korolova, 2014). At the same time, researchers have been working on the broader challenge of principled integration of DP and other formal privacy guarantees with decades of SDC experience in data utility and statistical approaches to uncertainty quantification.

Today, the emergence of AI represents another shift. AI enables learning, data generation, and reuse at an unprecedented scale. The structural constraints that once limited data reuse have largely disappeared, and privacy risk must now be understood as evolving over time. In the current AI era, the central question is no longer only whether a single data release is safe at the moment of dissemination. Instead, released data may enter broader workflows in which they are reused, transformed, linked with auxiliary sources, and incorporated into downstream models and generated outputs. As a result, privacy risk becomes increasingly temporal, adaptive, and cumulative. At the same time, while these developments expand what can be learned from data, the errors and biases introduced by both learning algorithms and privacy-preserving mechanisms may also accumulate. This raises further questions about data utility, including accuracy, precision, credibility, and relevance.

Table 1: How AI stress-tests long-standing privacy challenges across successive eras of confidentiality protection.

Era	What changed	How AI stress-tests it
Early confidentiality (pre-1960s)	Low data reuse	AI enables large-scale inference from aggregates
Traditional SDC (1960s–1990s)	Limited linkage	AI automates linkage and reconstruction
Modern SDC and synthetic data (early 2000s)	Valid inference	AI increases model capacity and can raise re-identification risk
Differential privacy (mid-2000s)	Worst-case guarantees	AI makes worst-case attackers more realistic
AI (now)	**Learning everywhere**	**Privacy risk is temporal, adaptive, and cumulative**

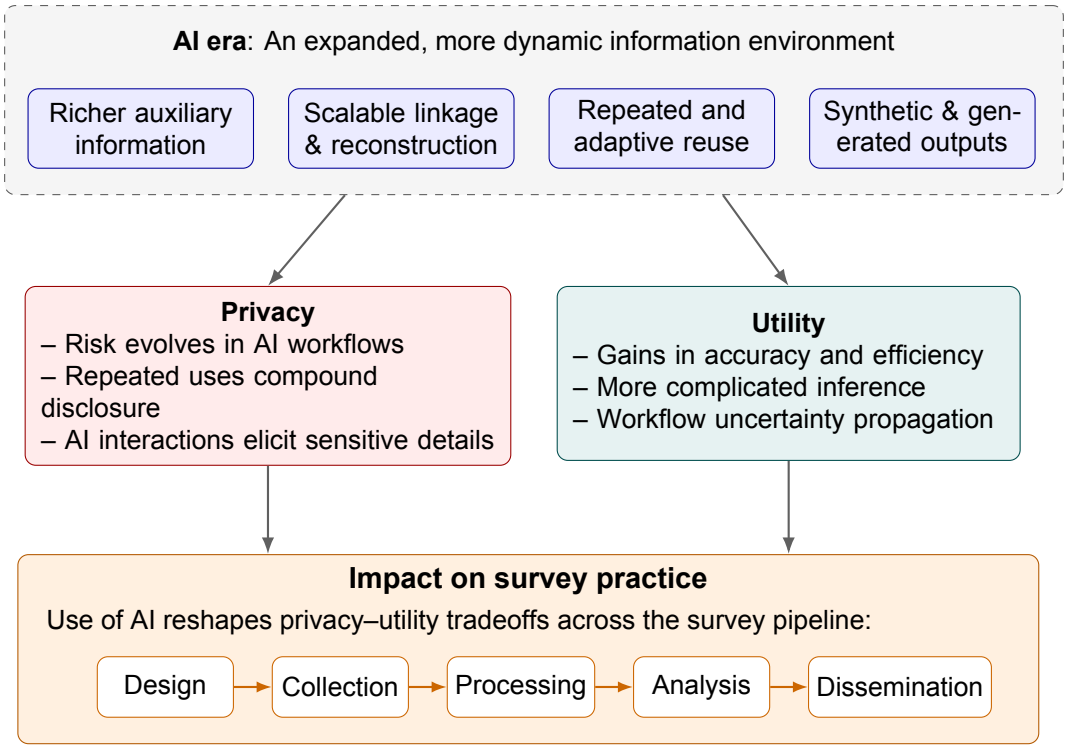


Figure 1: How the AI era reshapes privacy and utility and their implications across the survey pipeline.

2.2 AI as an Amplifier of Both Utility and Risk

AI raises the stakes of data privacy. In earlier settings, privacy protection was often framed as a release-stage problem, in which a table, statistic, or synthetic dataset was produced, evaluated, and then disseminated. In the AI era, this view is too narrow. Released data may instead enter broader information environments, where they can be queried repeatedly, linked to other sources, incorporated into model-based systems, and used to generate new outputs. This expanded setting makes robust and transparent privacy protection more critical, particularly because informal disclosure-control rules may become fragile when data are reused and combined in unforeseen ways. While this shift is not unique to survey data, it is especially consequential in survey settings, where these dynamics affect multiple stages of the survey pipeline, including questionnaire design, data collection, processing, analysis, and dissemination. Figure 1 summarizes, at a high level, these implications for the privacy–utility tradeoff.

On the privacy side, released survey products may now be combined with richer auxiliary information, linked or reconstructed at scale, reused repeatedly and adaptively, or transformed into synthetic and other generated outputs. As a result, privacy risk becomes increasingly dependent on the surrounding information environment rather than solely on the released object itself. It may evolve as data move through AI-enabled workflows, with repeated use compounding disclosure concerns. In conversational or chatbot-based survey administration, AI may also elicit richer and potentially more sensitive respondent information than a fixed questionnaire would have collected (Wuttke et al., 2025; Xiao et al., 2020).

On the utility side, AI-assisted tools can improve efficiency and, in some settings, task-level accuracy by supporting question development, coding open-ended responses, summarizing complex outputs, and making survey results more accessible to broader audiences (Törnberg, 2023; Mellon et al., 2024). These tools may also reduce costs by automating or augmenting labor-intensive steps in the

survey workflow (Jansen, Jung, and Salminen, 2023). In addition, synthetic data and other model-generated outputs are increasingly discussed as ways to expand access while reducing direct exposure of confidential data (Kapania et al., 2025). Such products may support exploration, software development, education, early-stage analysis, or model prototyping.

These potential gains in utility do not automatically translate into inferential validity. In survey settings, utility depends not only on whether specific steps become faster or more accurate, but also on whether the resulting data products support credible estimation and appropriate uncertainty quantification. AI-enabled workflows can propagate errors across stages: inaccuracies in coding may affect downstream estimates, synthetic outputs may preserve certain distributional features while distorting inferential targets, and automated reporting tools may present results without clearly conveying uncertainty arising from sampling, processing, and privacy protection. As a result, AI may expand the practical uses of protected survey data while making the validity of inference more difficult to assess.

These issues are especially important for survey data because such data occupy a distinctive role in the broader data ecosystem. Survey data are carefully curated, closely tied to population-level inference and public policy, and often contain sensitive information. They are also produced through a complex inferential pipeline shaped by sampling, weighting, editing, imputation, calibration, and, in some cases, record linkage. In this context, privacy protection cannot be treated as a simple technical add-on; it interacts directly with data production, downstream use, and statistical interpretation.

Recent discussions in the survey community have highlighted both the opportunities and the risks of AI. While AI may enhance many aspects of survey work, it also raises concerns about transparency, validation, and the reliability of generated outputs (Rothschild et al., 2025). These concerns align naturally with privacy. As it becomes increasingly difficult to distinguish between original data, protected releases, and generated outputs, it becomes more important to understand how privacy protection enters the survey pipeline and how it affects both disclosure risk and inferential utility.

3 Privacy, Uncertainty, and Inference

Statistical inference typically reflects a sequence of data-generation, collection, and processing steps rather than a single-step calculation. Privacy protection fits naturally into this broader framework because it systematically modifies the released object on which downstream analysis is based; see, for example, Figure 1 in Slavković and Seeman (2023). In this section we outline a perspective that the additional uncertainty introduced by privacy protection is both unavoidable and essential to account for in order to support valid statistical inference. This perspective extends beyond survey data alone, but it is especially important in survey settings, where inference is already shaped by multiple sources of uncertainty, including sampling variation, nonresponse, weighting, imputation, and linkage (Groves, F. J. Fowler, et al., 2011).

3.1 Privacy Risk and Statistical Utility

A useful way to formalize privacy and inference is to distinguish among the confidential data object, the released (sanitized) object, and the information available to different data users. Here, *privacy (disclosure) risk* refers to what can be learned about individuals or confidential records from a release, whereas *statistical utility* refers to what the release still supports in terms of estimation, uncertainty quantification, and downstream analysis.[3]

[3] Our notation aligns with Seeman (2023).

Let $\theta \in \Theta$ denote the population quantity or model parameter of interest, $f_\theta(\cdot)$ the data-generating model, $X \in \mathcal{X}$ the confidential data, and $Y \in \mathcal{Y}$ the released output after privacy protection. We write $M_X(\cdot)$ for the privacy-preserving release mechanism or method. In DP, such mechanisms are typically randomized, for example through noise addition; in SDC, they may be randomized, as in data swapping or noise addition, or deterministic, as in recoding. Then we write

$$X \sim f_\theta, \qquad Y = M_X(X). \tag{1}$$

Let $\pi_A(\cdot)$ and $\pi_D(\cdot)$ denote, respectively, the information available to an adversary and to a data analyst. Here, an adversary refers to a party attempting to infer sensitive information about individuals or confidential records from released data and possible auxiliary information. Privacy risk can then be viewed through the adversary's updated information,

$$\pi_A(\cdot \mid M_X, Y), \tag{2}$$

whereas statistical utility may be viewed through the analyst's resulting estimator,

$$\hat{\theta} = \hat{\theta}(M_X, Y, \pi_D). \tag{3}$$

This framework naturally accommodates additional information sources. Let Z denote external data, such as linked datasets, auxiliary information, prior or related releases, or AI-generated inputs. Such information may improve downstream analysis, but in both SDC and DP settings it can also increase privacy risk by facilitating record linkage, attribute inference, or reconstruction across releases.

The cumulative perspective is particularly challenging for SDC, where protection is often assessed for a specific release under context-dependent assumptions about what an intruder may know. As the surrounding information environment evolves, these assumptions may become unstable. As discussed in Slavković and Seeman (2023), SDC typically operates with an *absolute* notion of disclosure risk, whereas DP defines a *relative* notion of risk over a broader set of adversarial contexts, quantifying differences in disclosure risk between similar datasets. In the AI era, privacy risk becomes increasingly *relative* to the broader information environment, as it depends on prior releases, auxiliary data, and patterns of downstream reuse.

Differential privacy provides a clear illustration of this point. Relaxed notions such as (ϵ, δ)-DP permit a small failure probability, and repeated releases affect not only the cumulative privacy-loss parameter ϵ but also the failure-probability term δ. Thus, while DP provides a principled framework for accounting for repeated analyses, the cumulative risk may still grow over time. More broadly, privacy concerns may increase in mixed release systems that combine protected and unprotected outputs. For example, an earlier release may provide unprotected state-level summaries, while a later release provides protected county-level outputs derived from the same underlying data. Although the latter release may satisfy its formal privacy guarantee in isolation, disclosure risk may increase when the two are interpreted jointly. This motivates formal approaches to partially private data (Seeman, Reimherr, and Slavković, 2022). Taken together, these considerations suggest that privacy should be evaluated not only at the level of individual outputs, but also at the level of the broader system of linked data products, related releases, and external information.

3.2 Two Inferential Regimes

There are two inferential regimes for analysis under privacy protection. In Slavković and Seeman (2023), these are formulated as statistical risk minimization problems, where $L(\cdot, \cdot)$ denotes a loss

function. In what they call the *design problem*, the privacy mechanism and the estimator are chosen jointly for a particular inferential goal. A generic formulation is

$$\hat{\theta}_{\text{Design}} = \arg \min_{\tilde{\theta},\, M \in \mathcal{M}} \sup_{\theta \in \Theta} E_{\theta,M} \left[L\{\tilde{\theta}(Y), \theta\} \right], \tag{4}$$

where Θ denotes the parameter space and $\mathcal{M}$ denotes a class of admissible privacy-preserving mechanisms. For a fixed data-generating parameter θ and mechanism $M \in \mathcal{M}$, $E_{\theta,M}$ denotes expectation with respect to both the data-generating process indexed by θ and any randomness introduced by the mechanism M.

By contrast, in the *adjustment problem*, the privacy mechanism is fixed in advance and the task is to construct valid inference from the released object. In that case,

$$\hat{\theta}_{\text{Adjust}} = \arg \min_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta,M} \left[L\big(\tilde{\theta}(Y), \theta\big) \right]. \tag{5}$$

These formulations clarify the distinction at a formal decision-theoretic level and serve as useful conceptual targets. In practice, however, privacy-preserving methods are rarely obtained by explicitly solving these minimization problems, which are often intractable in realistic settings. Instead, methods are typically developed through problem-specific constructions. For example, a wide range of methods has been proposed for differentially private linear regression, and their performance varies across data settings and assumptions rather than yielding a single uniformly optimal solution (Wang, 2018; Amin et al., 2023; Lin, Slavković, and Bhoomireddy, 2026). For this reason, it is often more useful to describe the corresponding settings in terms of workflow. We refer to these as *primary analysis under privacy* and *secondary analysis under privacy*. This terminology emphasizes the practical distinction most relevant for analysts: whether privacy protection is incorporated into the statistical procedure from the outset or instead enters through an already protected release that must be analyzed afterward. A closely related distinction appears in record linkage, where one differentiates between settings in which linkage and analysis are handled jointly and settings in which analysts are provided with linked data for downstream analysis (Kamat and Gutman, 2026). The same lesson carries over to privacy: uncertainty introduced during data construction may either be incorporated directly into the analysis or addressed only after the fact.

Under privacy protection, *primary analysis* refers to settings in which the release mechanism, inferential target, and uncertainty quantification are developed jointly, whereas *secondary analysis* refers to settings in which the analyst is given a protected release and must determine what valid inference remains possible. The distinction is therefore not only chronological but also concerns what is under the analyst's control. This makes transparency and documentation especially important, since reliable inference depends on understanding how the protected object was constructed and what perturbations it reflects.

3.3 Inferential Error and Uncertainty

Privacy protection matters for inference because it can affect both variability and bias. These quantities are especially important in survey settings, where they directly shape the quality of point estimates, interval estimates, and other inferential summaries used in policy and official statistics. Once Y is generated from the confidential data through a randomized release mechanism, its variability reflects both the underlying data-generating and sampling process and the additional randomness

introduced by the release mechanism. For randomized mechanisms, a useful decomposition is

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\mathbb{E}[Y \mid X]\right) + \mathbb{E}\left[\mathrm{Var}(Y \mid X)\right], \tag{6}$$

where the outer expectation and variance are taken with respect to the data-generating and sampling process for X, while the conditional expectation and variance are taken with respect to the randomness of the privacy mechanism given X. The first term captures variability inherited from the data-generating and sampling process, while the second reflects variability introduced by privacy protection.

To isolate the effect of privacy protection on bias, it is useful to compare the release Y to the corresponding non-private quantity $T(X)$ for the same estimation goal. The privacy-induced bias conditional on X may then be written as

$$b_{\mathrm{priv}}(X) = \mathbb{E}[Y \mid X] - T(X). \tag{7}$$

Averaging over the data-generating process gives

$$\mathbb{E}[Y] - \theta = \mathbb{E}\left[\mathbb{E}[Y \mid X] - T(X)\right] + \left(\mathbb{E}[T(X)] - \theta\right), \tag{8}$$

where the first term represents bias introduced by the privacy mechanism and the second corresponds to the bias of the non-private estimator. Valid inference therefore requires both sources of error to be accounted for.

These considerations are particularly evident in confidence interval (CI) estimation. If privacy-induced noise is ignored and the released statistic is analyzed as though it were the corresponding non-private statistic, the resulting inference will likely be adversely affected in ways that may be difficult for the analyst or data user to detect. Confidence intervals may be too narrow, empirical coverage may fall below the nominal level, and point estimates may exhibit bias arising from naive downstream analysis. These issues extend beyond confidence intervals to more general inferential procedures.

Figure 2 illustrates this phenomenon in a linear regression setting using synthetic data generated by BinAgg (Lin, Slavković, and Bhoomireddy, 2026), a differentially private synthetic data method paired with a customized regression procedure. The first panel shows the sampling distribution of the non-private OLS estimator for the regression coefficient and its corresponding CI. The second and third panels show the distributions for estimators based on differentially private synthetic data. The second panel corresponds to a naive approach that fits linear regression directly to the synthetic data, treating them as if they were the original data. The third panel corresponds to the BinAgg approach, which adjusts regression to account for the additional variability introduced by differential privacy.

Both private approaches exhibit greater dispersion than the non-private OLS benchmark, since the synthetic data reflect not only sampling variability but also additional variability induced by the privacy mechanism. The naive analysis of the synthetic data exhibits both noticeable bias and undercoverage of its CIs, whereas the BinAgg estimator is much closer to unbiasedness and its adjusted CI attains coverage close to the nominal level. In fact, the associated large sample theory based on a central limit theorem is designed to justify asymptotic unbiasedness and valid interval estimation under the BinAgg procedure.

The AI-era setting described in Section 2 makes careful inferential accounting even more important. Protected releases are increasingly reused in automated workflows, combined with external information, and incorporated into model-based systems. As a result, the uncertainty introduced by privacy

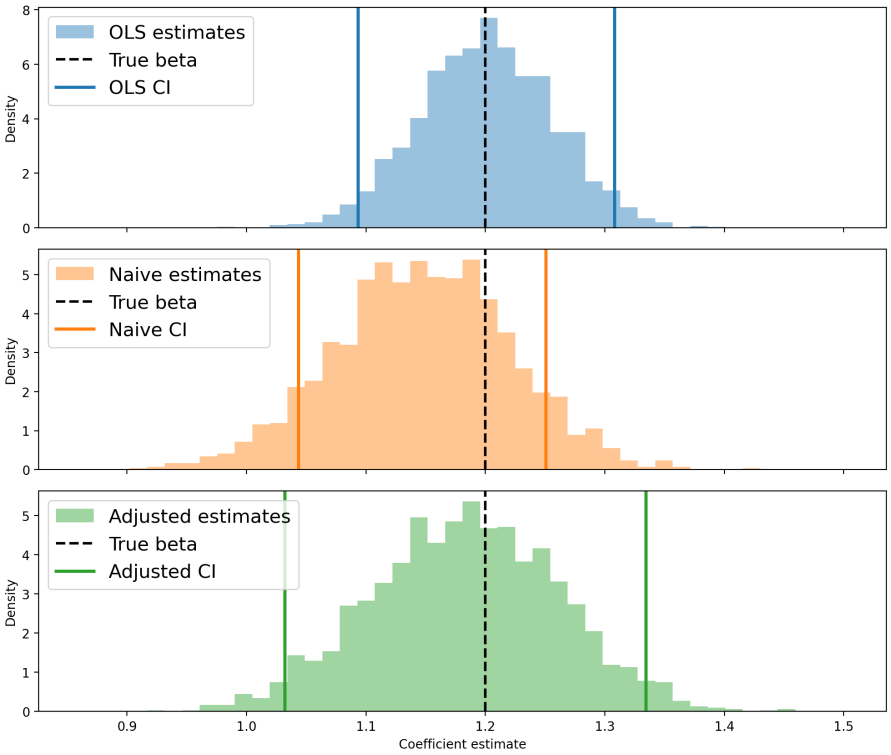


Figure 2: Confidence intervals for the regression coefficient under non-private and differentially private synthetic-data-based simple linear regression for a sample of size $n = 1000$. Synthetic data generated by BinAgg under μ-Gaussian DP (Dong, Roth, and Su, 2022) with $\mu = 1$.

protection may propagate through subsequent stages of processing and analysis rather than remaining confined to the initial release. The inferential consequences of privacy protection therefore extend beyond the point of release and carry into downstream analysis.

4 Statistical Data Privacy Implications for Survey Practice

Viewing privacy protection as part of the data pipeline has important implications for survey practice. Survey data are produced through a sequence of design and processing decisions, including questionnaire design, sample design, field procedures, weighting, calibration, editing, imputation, and, in some cases, record linkage. Each of these steps helps determine what can ultimately be learned from the data.

These design and processing choices are typically aligned with the intended end uses of the survey data. For example, a survey designed primarily to publish tabulations or change estimates may place greater emphasis on preserving valid totals and comparisons, whereas a survey designed to release research microdata may place greater emphasis on preserving broader distributional features. These priorities, in turn, affect both disclosure risk and the choice of privacy protection methods. In this sense, privacy protection is not external to the survey process, but part of the same chain of design and processing decisions. Because privacy protection modifies the released data product, it affects not only disclosure risk, but also the inferential validity, interpretability, and usability of the resulting data products.

It is therefore important to understand how privacy mechanisms interact with survey-specific structures. Complex sample designs, including stratified, clustered, multistage, and unequal-probability

designs, can affect both the sensitivity and the variability of survey estimators. Survey weights may amplify the influence of certain units and change how privacy-induced perturbations propagate through estimation. Nonresponse adjustment, calibration, editing, and imputation introduce additional dependencies and processing layers. Together, these features complicate the design of privacy mechanisms and make it more difficult to determine how privacy-induced perturbations interact with downstream estimation. Existing work has begun to address how specific survey features interact with differential privacy, including imputation, stratified sampling, survey weighting, and record linkage (Das et al., 2022; Lin, Bun, et al., 2024; Seeman, Si, and Reiter, 2026; Lin, Paquette, and Kolaczyk, 2024). For a broader discussion of the challenges of applying differential privacy in survey settings, see Drechsler and Bailie (2026). More generally, privacy must be understood in conjunction with the survey design and processing steps that shape the data object used for analysis.

Another implication is the need for transparency (**Gong2022Transparent**) and documentation, especially for secondary analysis under privacy. Survey methodologists have long emphasized documenting sampling, weighting, nonresponse adjustment, editing, and imputation because these features directly affect inference. Privacy protection should be documented with the same level of care. This does not imply that every operational detail must be disclosed. Full disclosure of SDC procedures is often impractical, whereas DP emphasizes algorithmic transparency. In either case, analysts need sufficient information to understand how privacy protection affects inference. Without such information, protected releases may be treated as if they were unmodified data, leading to misscalibrated downstream analyses and inference; e.g., for simple examples, see A. Slavkovic and Lee, 2010 for the case of two-way contingency table analysis or Woo and A. B. Slavkovic, 2012 for a logistic regression estimation after PRAM as a disclosure avoidance method was applied.

This issue is especially important for synthetic data products. In the generative AI era, such products are increasingly presented as a response to data scarcity, resource constraints, and limits on access to sensitive data (Kapania et al., 2025). Because they are often positioned as accessible alternatives to restricted data, transparency about how they are generated, which features they preserve, and what inferential tasks they are intended to support becomes essential.

A further implication is that privacy choices should be aligned with the policy goals and intended uses of the survey. Survey data are often produced to support population estimates, subgroup comparisons, trend monitoring, and policy decisions. These objectives are not interchangeable, and they are not equally sensitive to privacy protection. A release mechanism that preserves broad national patterns may still distort local heterogeneity, weaken estimation for small areas and subpopulations, or reduce reliability for rare outcomes. In practice, this means that privacy protection cannot be evaluated solely in terms of overall utility or average error. It must be assessed relative to the specific estimands, disaggregations, and inferential tasks the survey is intended to support. In survey settings, privacy design is therefore partly a matter of prioritization: deciding which uses must remain reliable, for whom, and with what accounting of the uncertainty introduced by protection.

5 Conclusion

To unlock the transformative potential of AI, data privacy must remain central to digital strategy. It is fundamental to ethical and socially responsible AI and machine learning. Statistical data privacy methods provide a coherent framework for this setting by combining key properties of formal privacy, such as methodological transparency, robustness to post-processing, and principled accounting of cumulative privacy loss, with careful attention to statistical utility. Privacy-preserving algorithms introduce structured randomness that, if not properly accounted for, can distort downstream analysis, amplify

the effects of data scarcity, and introduce bias. Uncertainty quantification must therefore remain a core principle when working with protected data and when generating curated synthetic datasets. At the same time, AI makes uncertainty quantification more complex and privacy risk more continuous, adaptive, and cumulative.

The AI era amplifies familiar privacy concerns in survey and official statistics by making them more dynamic, more cumulative, and more tightly connected to downstream analysis. Privacy protection must therefore be evaluated not only in terms of disclosure risk, but also in relation to the inferential goals that released data are intended to support. This is especially important in survey settings, where data are produced through multiple stages of design and processing and are often intended to support inference by secondary users. This perspective highlights the importance of survey-specific design features, transparent documentation, and evaluation relative to intended uses.

The survey community brings important conceptual and technical tools to these challenges, including design-based thinking, careful uncertainty quantification, and a sustained focus on intended use. In the AI era, the central question is not simply whether privacy protection reduces risk or utility, but under what forms of protection, for which inferential goals, and with what accounting of uncertainty released data can still support trustworthy, transparent, and policy-relevant analyses in an increasingly complex data ecosystem.

Acknowledgment

This work was supported at Penn State by the Huck Institutes of the Life Sciences through the Dorothy Foehr Huck and J. Lloyd Huck Chair in Data Privacy and Confidentiality, and by a 2025–2026 Rising Researcher Grant from the Institute for Computational and Data Sciences (RRID: $SCR025154$).

References

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

Abowd, J. M. (2018). The U.S. Census Bureau Adopts Differential Privacy. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867.

Abowd, J. M., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, M. Spence, and P. Zhuravlev (June 2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. In: *Harvard Data Science Review*. Special Issue 2. URL: https://hdsr.mitpress.mit.edu/pub/7evz361i.

Abowd, J. M. and I. M. Schmutte (2015). Economic Analysis and Statistical Disclosure Limitation. In: *Brookings Papers on Economic Activity* 50.1 (Spring), pp. 221–267. URL: https://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

Amin, K., M. Joseph, M. Ribero, and S. Vassilvitskii (2023). Easy Differentially Private Linear Regression. In: *International Conference on Learning Representations (ICLR)*.

Apple (2017). *Learning with Privacy at Scale*. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

Awan, J. and R. Gong (2024). "Statistical Inference and Differential Privacy". In: *Handbook of Sharing Confidential Data*. Chapman and Hall/CRC, pp. 115–135.

Carlini, N., F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel (Aug. 2021). "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, pp. 2633–2650. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Charest, A.-S. and J. Drechsler (2026). Differential Privacy and its Application to Survey Data. In: *The Survey Statistician* 93, pp. 13–25.

Das, S., J. Dreschler, K. Merrill, and S. Merrill (2022). Imputation under Differential Privacy. In: *ArXiv* abs/2206.15063. URL: https://api.semanticscholar.org/CorpusID:250144515.

Dinur, I. and K. Nissim (2003). "Revealing information while preserving privacy". In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '03. San Diego, California: Association for Computing Machinery, pp. 202–210. ISBN: 1581136706. DOI: 10.1145/773153.773173. URL: https://doi.org/10.1145/773153.773173.

Dong, J., A. Roth, and W. J. Su (Feb. 2022). Gaussian Differential Privacy. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84.1, pp. 3–37. DOI: 10.1111/rssb.12454.

Drechsler, J. and J. Bailie (2026). "The Complexities of Differential Privacy for Survey Data". In: *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and New Findings*. University of Chicago Press. Chap. 5. URL: https://www.nber.org/books-and-chapters/data-privacy-protection-and-conduct-applied-research-methods-approaches-and-new-findings/complexities-differential-privacy-survey-data.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes (2001). *Disclosure Risk vs. Data Utility: The R-U Confidentiality Map*. Tech. rep. Technical Report Number 121. Research Triangle Park, NC: National Institute of Statistical Sciences.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In: *Theory of Cryptography Conference*, pp. 265–284.

Dwork, C. and A. Roth (Aug. 2014). The Algorithmic Foundations of Differential Privacy. In: *Foundations and Trends in Theoretical Computer Science* 9.3–4, pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

Dwork, C., A. Smith, T. Steinke, and J. Ullman (2017). Exposed! A Survey of Attacks on Private Data. eng. In: *Annual Review of Statistics and Its Application (2017)*.

Erlingsson, Ú., V. Pihur, and A. Korolova (2014). "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. Scottsdale, Arizona, USA: Association for Computing Machinery, pp. 1054–1067. ISBN: 9781450329576. DOI: 10.1145/2660267.2660348. URL: https://doi.org/10.1145/2660267.2660348.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011). *Survey Methodology*. 2nd ed. Wiley.

Groves, R. M., J. Fowler Floyd J., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. New York: Wiley.

Groves, R. M. and L. Lyberg (2010). Total Survey Error: Past, Present, and Future. In: *Public Opinion Quarterly* 74.5, pp. 849–879. DOI: 10.1093/poq/nfq065.

Gymrek, M., A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich (2013). Identifying Personal Genomes by Surname Inference. In: *Science* 339.6117, pp. 321–324. DOI: 10.1126/science.1229566. eprint: https://www.science.org/doi/pdf/10.1126/science.1229566. URL: https://www.science.org/doi/abs/10.1126/science.1229566.

Holan, S. H., D. Toth, M. A. R. Ferreira, and A. F. Karr (2010). Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality. In: *Journal of the American Statistical Association*

105.490, pp. 564–577. DOI: 10.1198/jasa.2009.ap08629. eprint: https://doi.org/10.1198/jasa.2009.ap08629. URL: https://doi.org/10.1198/jasa.2009.ap08629.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In: *PLoS genetics* 4.8, e1000167. DOI: 10.1371/journal.pgen.1000167.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf (2012). *Statistical Disclosure Control*. Wiley.

Jansen, B. J., S.-g. Jung, and J. Salminen (2023). Employing Large Language Models in Survey Research. In: *Natural Language Processing Journal* 4, p. 100020. ISSN: 2949-7191. DOI: 10.1016/j.nlp.2023.100020.

Kamat, G. and R. Gutman (2026). Analysis of Linked Files: A Missing Data Perspective. In: *Statistical Science* 41.1, pp. 28–48. DOI: 10.1214/24-STS939.

Kapania, S., S. Ballard, A. Kessler, and J. W. Vaughan (2025). Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 16 pages. DOI: 10.1145/3715275.3732005.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. In: *International Statistical Review* 79.3, pp. 362–384. DOI: 10.1111/j.1751-5823.2011.00153.x.

Lin, S., M. Bun, M. Gaboardi, E. D. Kolaczyk, and A. Smith (2024). Differentially private confidence intervals for proportions under stratified random sampling. In: *Electronic Journal of Statistics* 18.1, pp. 1455–1494. DOI: 10.1214/24-EJS2234.

Lin, S., E. Paquette, and E. D. Kolaczyk (July 2024). Differentially Private Linear Regression With Linked Data. In: *Harvard Data Science Review* 6.3.

Lin, S., A. Slavković, and D. R. Bhoomireddy (2026). Differentially Private Linear Regression and Synthetic Data Generation with Statistical Guarantees. In: *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research 300.

Lohr, S. L. (2021). *Sampling: Design and Analysis*. 3rd ed. CRC Press.

Matthews, G. J. and O. Harel (2011). Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy. In: *Statistics Surveys* 5, pp. 1–29.

Mellon, J., J. Bailey, R. Scott, J. Breckwoldt, M. Miori, and P. Schmedeman (Jan. 2024). Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale. In: *Research & Politics* 11.1. DOI: 10.1177/20531680241231468.

Narayanan, A. and V. Shmatikov (2008). "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy*, pp. 111–125. DOI: 10.1109/SP.2008.33.

Reiter, J. P. (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1, pp. 185–205. DOI: 10.1111/j.1467-985X.2004.00343.x.

Rothschild, D. M., T. D. Buskirk, S. Eckman, D. S. Hillygus, F. Kreuter, and D. Lazer (2025). Successfully Navigating the Disruption AI Will Bring to Survey Research. In: *The Survey Statistician* 92, pp. 30–44.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. In: *Journal of Official Statistics* 9.2, pp. 461–468.

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

Seeman, J. (2023). “Theoretical and Applied Problems in Partially Private Data”. Doctoral Dissertation. University Park, PA: The Pennsylvania State University. URL: https://etda.libraries.psu.edu/catalog/23008jhs5496.

Seeman, J., M. Reimherr, and A. Slavković (2022). Formal Privacy for Partially Private Data. In: *ArXiv* abs/2204.01102. URL: https://arxiv.org/abs/2204.01102.

Seeman, J., Y. Si, and J. P. Reiter (2026). “Differentially Private Population Quantity Estimates via Survey Weight Regularization”. In: *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and New Findings*. University of Chicago Press. Chap. 6. URL: https://www.nber.org/books-and-chapters/data-privacy-protection-and-conduct-applied-research-methods-approaches-and-new-findings/differentially-private-population-quantity-estimates-survey-weight-regularization.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov (May 2017). “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. DOI: 10.1109/SP.2017.41.

Skinner, C. J. (2009). “Statistical Disclosure Control for Survey Data”. In: *Handbook of Statistics*. Vol. 29. Elsevier, pp. 381–396.

Slavkovic, A. and J. Lee (May 2010). Synthetic two-way contingency tables that preserve conditional frequencies. In: *Statistical Methodology* 7, pp. 225–239. DOI: 10.1016/j.stamet.2009.11.002.

Slavković, A. and J. Seeman (2023). Statistical Data Privacy: A Song of Privacy and Utility. In: *Annual Review of Statistics and Its Application* 10, pp. 189–218. DOI: 10.1146/annurev-statistics-033121-112921. URL: https://doi.org/10.1146/annurev-statistics-033121-112921.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. In: *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. In: *arXiv preprint arXiv:2304.06588*.

Wang, Y.-X. (2018). Revisiting Differentially Private Linear Regression: Optimal and Adaptive Prediction & Estimation in Unbounded Domain. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Webb, K., P. Protivash, J. Durrell, D. Toth, A. Slavković, and D. Kifer (2026). Statistics-Friendly Confidentiality Protection for Establishment Data, with Applications to the QCEW. In: *PoPETS*. To appear. arXiv:2509.01597. arXiv: 2509.01597 [cs.CR]. URL: https://arxiv.org/abs/2509.01597.

Woo, Y. M. J. and A. B. Slavkovic (2012). “Logistic Regression with Variables Subject to Post Randomization Method”. In: *Privacy in Statistical Databases*, pp. 116–130. URL: https://api.semanticscholar.org/CorpusID:5871044.

Wuttke, A., M. Aßenmacher, C. Klamm, M. M. Lang, Q. Würschinger, and F. Kreuter (May 2025). AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* LaTeCH-CLfL 2025, pp. 179–204. DOI: 10.18653/v1/2025.latechclfl-1.17.

Xiao, Z., M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, and H. Yang (June 2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. In: *ACM Transactions on Computer-Human Interaction* 27.3. ISSN: 1073-0516. DOI: 10.1145/3381804.

High-Dimensional Variance Estimation for the Generalized Regression Estimator

Kalil Bouhadra[1] **and Mehdi Dagdoug**[2]

McGill University, Department of Mathematics and Statistics, Montreal, Canada
[1]kalil.bouhadra@mcgill.ca, [2] mehdi.dagdoug@mcgill.ca

Abstract

In survey sampling, the goal is to estimate finite population parameters such as totals, means, and proportions. At the estimation stage, it is common to have access to auxiliary information in the form of covariates, known either in aggregate form or for each population unit. These covariates are often used, through models relating them to the variable of interest, to improve efficiency; this approach is known as model-assisted estimation. Modern applications increasingly involve settings where a large number of covariates are observed, sometimes of the same order as the sample size. While this setting offers greater modeling flexibility, it also creates important challenges for inference. In this article, we study variance estimation for the generalized regression (GREG) estimator in high-dimensional regimes. We derive new theoretical results that characterize the high-dimensional asymptotic bias of commonly used variance estimators, including those based on Taylor linearization. Furthermore, under suitable distributional assumptions on the covariates, we show that a cross-validated variance estimator is naturally asymptotically unbiased.

Keywords: finite population sampling; model-assisted estimation; variance estimation; high-dimensional asymptotics; cross-fitting.

1 Introduction

Model-assisted methods are widely used in survey sampling to improve the efficiency of point estimators by leveraging auxiliary information. In particular, the generalized regression estimator (GREG) provides a flexible framework for incorporating covariates through linear modeling; see, e.g., Cassel, Särndal, and Wretman (1977) and Särndal and Wright (1984) for foundational contributions and Särndal, Swensson, and Wretman (1992) for a pedagogical treatment. Beyond point estimation, variance estimation plays a central role in practice, as it allows for the construction of confidence intervals and other measures of uncertainty routinely reported by national statistical offices.

A variety of methods have been proposed for variance estimation of the GREG estimator, including approaches based on Taylor linearization and its g-weighted version (Särndal, Swensson, and Wretman, 1989; Valliant, 2002), as well as resampling techniques such as the jackknife (Duchesne, 2000; Berger and Skinner, 2005). More recent contributions on the topic include, among others, Stefan and Hidiroglou (2023) and Stefan and Hidiroglou (2024), which further develop bootstrap and jackknife methodologies in this context.

Classically, the properties of the GREG estimator and its associated variance estimators have been studied within a low-dimensional asymptotic framework, in which the number of covariates is assumed to be negligible relative to the sample size (Robinson and Särndal, 1983; Kott, 1990). More precisely, this framework considers a regime where the number of covariates p is fixed, while the sample size n and population size N tend to infinity. Such an approximation is appropriate to model practical situations when the ratio p/n is small, that is, $p/n \approx 0$.

In many modern applications, however, practitioners are confronted with settings where the number of covariates is no longer negligible compared to the sample size. For instance, when $n = 100$ and $p = 30$, the ratio $\kappa = p/n$, here equal to $\kappa = 0.3$, is non negligible. A large number of covariates may also originates from a nonlinear low-dimensional setup via basis expansions of the original covariates (e.g., adding powers and interactions of the original covariates). This has led to a growing interest in high-dimensional regimes in survey sampling, where the number of covariates p increases with the sample size n. In this context, Cardot, Goga, and Shehzad (2017) studied calibration based on principal components, while Ta et al. (2020), Chauvet and Goga (2022), and Dagdoug, Goga, and Haziza (2023) investigated the high-dimensional properties of the GREG estimator. More recently, Eustache, Dagdoug, and Haziza (2025) highlighted important limitations of classical variance estimators in such settings. In particular, they showed that standard procedures based on Taylor linearization tend to underestimate the variance, whereas resampling methods such as the jackknife tend to overestimate it. These biases can be substantial and persist asymptotically when p/n does not vanish.

Roughly speaking, these phenomena can be traced back to the high-dimensional behavior of several key quantities, which deviate markedly from their classical low-dimensional counterparts (e.g., residuals, leverages, and g-weights). For instance, the underestimation exhibited by Taylor-based variance estimators can be attributed to an underestimation of the variability of the regression residuals in high dimensions. Similar issues have been documented in classical statistics; see, for example, El Karoui and Purdom (2018) and Zhao and Candes (2022).

These effects are closely related to overfitting, which arises when a model-assisted estimator relies on a highly flexible regression function. In such cases, residuals tend to be artificially small, and when plugged into Taylor-type variance estimators, this leads to a systematic underestimation of the true variance (Dagdoug, Goga, and Haziza, 2023). To address this issue, Dagdoug, Goga, and Haziza (2023) proposed a cross-validated variance estimator, which replaces the in-sample residuals with residuals obtained through cross-validation. This idea initially originated from Opsomer and Miller (2005), who introduced related techniques in the context of hyper-parameter tuning for model-assisted estimators based on local polynomials. Although developed from a different perspective, it is also closely connected to the cross-fitted variance estimator studied in An, Dagdoug, and Haziza (2026).

Although Eustache, Dagdoug, and Haziza (2025) highlighted important issues of the classical variance estimators of the GREG in high dimensions, several aspects are only partially understood. For example, the high-dimensional bias of the Taylor variance estimator depends on the population average of the so-called g-weights, denoted $\overline{G}_N$. However, in a fixed-design setting, the high-dimensional behavior of $\overline{G}_N$ is difficult to characterize and is generally unknown. In addition, these results were established under high-dimensional assumptions on the sample leverage scores, whose behavior also depends on the distribution of the covariates. These assumptions were supported empirically, but their theoretical validity was not studied. Finally, under Bernoulli sampling, the bias formulas of Eustache, Dagdoug, and Haziza (2025) can be used to construct debiased variance estimators. It is less clear, however, whether these debiased estimators would continue to perform well beyond the Bernoulli setting. This motivates the search for a variance estimator that is *naturally* asymptotically unbiased and valid in all dimensions, rather than one obtained through adjustments specific to a particular setting.

Contributions. In this paper, we further investigate the problem of variance estimation for the GREG estimator in high-dimensional regimes. First, we show that, under suitable conditions on the covariates, some of the assumptions used in Eustache, Dagdoug, and Haziza (2025) hold in classical settings. This includes characterizing the high-dimensional behavior of the g-weights mean $\overline{G}_N$ and of the

sample leverages. Second, we study the high-dimensional asymptotic bias of a cross-validated variance estimator and establish that, under appropriate conditions on the covariates and the sampling design, the estimator is asymptotically unbiased in all regimes, and therefore does not require any bias correction. We also provide closed-form expressions for the asymptotic biases of the Taylor variance estimator and its g-weighted version. The empirical results presented confirm that these expressions describe well empirical phenomena.

Outline. The remainder of the paper is organized as follows. In Section 2, we introduce the model-assisted framework and formally define the problem of interest. Section 3 presents the main theoretical results. In Section 4, we illustrate these findings through a simulation study. Finally, Section 5 concludes with a discussion of the limitations of our approach and directions for future research. All proofs are postponed to the Appendix.

2 Basic setup

2.1 Model-assisted estimation

Consider a finite population $U_N := \{1, 2, ..., N\}$ of size N. The measurements of a survey variable Y are denoted y_i for $i \in U_N$. We aim to estimate the finite population mean

$$\mu := \frac{1}{N} \sum_{i \in U_N} y_i.$$

A random sample S_N of size n_N is drawn from U_N using a sampling design $\mathcal{P}_N$. The sample S_N is equivalently characterized by the tuple $(I_i)_{i \in U_N}$ of sampling indicators satisfying $I_i := 1$ if $i \in S_N$ and $I_i := 0$, otherwise. The first-order and second-order inclusion probabilities are defined respectively as

$$\pi_i := \mathbb{P}(i \in S_N), \quad \pi_{ij} := \mathbb{P}(i, j \in S_N), \qquad i, j \in U_N.$$

We assume that $\pi_i > 0$ for all $i \in U_N$, a condition under which the Horvitz-Thompson estimator $\widehat{\mu}_\pi$ of μ defined by

$$\widehat{\mu}_\pi := \frac{1}{N} \sum_{i \in S_N} \frac{y_i}{\pi_i},$$

is design-unbiased for μ. We write $\mathbb{E}_p$ and $\mathbb{V}_p$ to denote the expectation and variance with respect to the sampling design, respectively. We denote by $\Delta_{ij} := \pi_{ij} - \pi_i \pi_j$ the sampling covariance of elements $i, j \in U_N$.

We consider a framework where p_N covariates $X_1, X_2, ..., X_{p_N}$ are observed for every population element; we denote by $\boldsymbol{x}_i$ the measurements of these covariates for element $i \in U_N$. Without loss of generality, we assume that the intercept is included in the first position, thus $X_0 = 1$. Although the overall number of covariates is $p_N + 1$ with the intercept, we sometimes write p_N instead, for simplicity.

Model-assisted estimation is commonly used to leverage the predictive power of covariates for the variable of interest. In the particular case of linear regression, the resulting estimator is the so-called Generalized REGression estimator Särndal, Swensson, and Wretman, 1992, GREG defined by

$$\widehat{\mu}_{greg} := \frac{1}{N} \left(\sum_{i \in U_N} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_N + \sum_{i \in S_N} \frac{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_N}{\pi_i} \right), \tag{1}$$

where, assuming that $\boldsymbol{A}_\Pi := \sum_{i\in S_N} \pi_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top$ is invertible, the weighted least squares coefficients $\widehat{\boldsymbol{\beta}}_N$ are given by

$$\widehat{\boldsymbol{\beta}}_N := \left(\sum_{i\in S_N} \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\pi_i}\right)^{-1} \sum_{i\in S_N} \frac{\boldsymbol{x}_i y_i}{\pi_i}.$$

In what follows, we assume for simplicity that $\boldsymbol{A}_\Pi$ is invertible, an assumption in line with the current literature, see, e.g., Chauvet and Goga (2022).

High-dimensional asymptotics

We are interested in studying the behavior of variance estimators of $\mathbb{V}_p(\widehat{\mu}_{greg})$ in situations where the number of covariates p_N is smaller than the sample size n_N, but of similar order, that is, $p_N/n_N \not\approx 0$. To model this situation, we embed it into an appropriate sequence of similar configurations. Specifically, we adapt the framework of Isaki and Fuller (1982) to accommodate for high-dimensional limiting cases. We consider a sequence of increasing populations $(U_N)_{N\in\mathbb{N}}$. In each population, a random sample S_N of size n_N is selected using a sampling design $\mathcal{P}_N$. Although we require the populations $(U_N)_{N\in\mathbb{N}}$ to be embedded, the samples $(S_N)_{N\in\mathbb{N}}$ are random and need not be. Based on each sample S_N and p_N covariates, a model-assisted estimator $\widehat{\mu}_{greg}$ is defined. This framework, therefore, allows for a number of covariates p_N that is increasing as N increases. More specifically, with $\kappa_N := p_N/n_N$, we consider cases where $(p_N)_{N\in\mathbb{N}}$ satisfies

$$\lim_{N\to\infty} \frac{p_N}{n_N} = \lim_{N\to\infty} \kappa_N := \kappa_\star \in [0, 1).$$

This framework includes: (i) the *low-dimensional* case, with $\kappa_\star = 0$, where the number of covariates is asymptotically negligible with respect of the sample size n_N; (ii) the *high-dimensional* case, with $\kappa_\star > 0$, where the number of covariates grows at the same order as n_N. The setup we consider, however, does not include the ultra-high dimensional case where $\kappa_* \geq 1$, which would require a different treatment.

A joint framework

The aim of this article is to analyze the high-dimensional bias of design-based variance estimators. However, deriving meaningful theoretical results in a purely design-based setup is challenging. To circumvent this difficulty, we follow the technique of Eustache, Dagdoug, and Haziza (2025), which considers a joint framework in which both the survey variable Y and the sample S_N are treated as random. Specifically, we assume that the measurements of the survey variable $(y_i)_{i\in U_N}$ are independent random variables satisfying the following linear model

$$y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i, \qquad i \in U_N,$$

where ϵ_i satisfies $\mathbb{E}_m[\epsilon_i] = 0$ and $\mathbb{E}_m[\epsilon_i^2] := \sigma^2$. We denote by $\mathbb{E}_m$ and $\mathbb{V}_m$ the expectation and variance with respect the distribution of the survey variable Y (treating the covariates X fixed), respectively.

The decomposition

$$\mathbb{V}_{mp}(\widehat{\mu}_{greg}) = \mathbb{E}_m\left[\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)\right] + \mathbb{V}_m\left(\mathbb{E}_p\left[\widehat{\mu}_{greg}\right]\right), \tag{2}$$

motivates variance estimators of the type

$$\widehat{V}_{mp} = \widehat{V}_{design} + \frac{\widehat{\sigma}^2}{N}, \tag{3}$$

where $\widehat{V}_{design}$ represents an estimator of $\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)$ and $\widehat{\sigma}^2$ an estimator of $\sigma^2 := \mathbb{V}_m(y_i)$. Given that σ^2 can be estimated unbiasedly in all configurations (n_N, p_N) with $p_N < n_N$, we assume without loss of generality that σ^2 is known.

The analysis technique we follow consists of studying the joint bias of estimators with the structure of $\widehat{V}_{mp}$. Our main underlying interest, however, remains the design bias of $\widehat{V}_{design}$. Since studying this bias directly is mathematically challenging, we instead analyze the joint bias of $\widehat{V}_{mp}$, which includes both a design component and a model component, with the understanding that the main contribution to the bias is expected to come from the design component.

Remark 1. *The rationale for the structure of the variance estimator $\widehat{V}_{mp}$ in* (3) *is based on the intuition that $\widehat{V}_{design}$ estimates the term $\mathbb{E}_m\left[\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)\right]$, and that $\widehat{\sigma}^2/N$ provides a reasonable estimator of $\mathbb{V}_m\left(\mathbb{E}_p\left[\widehat{\mu}_{greg}\right]\right)$. However, this implicitly relies on the idea that $\mathbb{E}_p\left[\widehat{\mu}_{greg}\right] = \mu$, so that $\mathbb{V}_m\left(\mathbb{E}_p\left[\widehat{\mu}_{greg}\right]\right) = \sigma^2/N$. Since the GREG estimator is biased, this argument is usually justified through asymptotic approximations: in the traditional low-dimensional asymptotic framework, the design bias of the GREG estimator is asymptotically negligible (Robinson and Särndal, 1983). However, this result typically holds when the number of covariates is fixed. Therefore, the validity of the estimator structure in* (3) *rests on the implicit assumption that the design bias of the GREG estimator remains asymptotically negligible, that is, $\mathbb{E}_m[\mathbb{E}_p[\widehat{\mu}_{greg} - \mu]^2] = o(N^{-1})$, even when $\kappa_\star > 0$. This assumption is supported by existing empirical work (see, for example, the simulation section of Dagdoug, Goga, and Haziza, 2023), but a formal proof is still lacking. This is, however, beyond the scope of the article, and we do not investigate it further. The above bias analysis technique, based on this idea, proved to work very well to model empirical phenomena, as shown Eustache, Dagdoug, and Haziza (2025) and in our simulations, see Section 4.*

2.2 The leave-one-out variance estimator

In the literature, many estimators $\widehat{V}_{design}$ of the design variance $\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)$ have been suggested. In Eustache, Dagdoug, and Haziza (2025), the following estimators were investigated.

1. The Taylor variance estimator, defined by

$$\widehat{V}_{tay} = \frac{1}{N^2} \sum_{i \in S_N} \sum_{j \in S_N} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\widehat{\epsilon}_i}{\pi_i} \frac{\widehat{\epsilon}_j}{\pi_j}, \tag{4}$$

   with $\widehat{\epsilon}_i := y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_N$ for $i \in S_N$.

2. The g-weighted Taylor variance estimator, defined by

$$\widehat{V}_g = \frac{1}{N^2} \sum_{i \in S_N} \sum_{j \in S_N} \frac{\Delta_{ij}}{\pi_{ij}} \frac{g_{i,N}\widehat{\epsilon}_i}{\pi_i} \frac{g_{\ell}\widehat{\epsilon}_j}{\pi_j}, \tag{5}$$

   where for $i \in S_N$, $g_{i,N} := \mathbf{t}_{\boldsymbol{x}}^\top \boldsymbol{A}_\Pi^{-1} \boldsymbol{x}_i$ with $\mathbf{t}_{\boldsymbol{x}} := \sum_{i \in U_N} \boldsymbol{x}_i$.

3. The Generalized Jackknife variance estimator (Berger and Skinner, 2005), which has the following closed-form solution

$$\widehat{V}_{jk} = \frac{1}{N^2} \sum_{i \in S_N} \sum_{j \in S_N} \frac{\Delta_{ij}}{\pi_{ij}} \frac{(1-w_i)\, g_{i,N}\, \widehat{\epsilon}_i}{(1-h_{ii,N})\, \pi_i} \frac{(1-w_j)\, g_j\, \widehat{\epsilon}_j}{(1-h_{jj})\pi_j}, \tag{6}$$

where $w_i := (N\pi_i)^{-1}$ and the weighted leverages are defined by

$$h_{ii,N} := \boldsymbol{x}_i^\top \boldsymbol{A}_\Pi^{-1} \pi_i^{-1} \boldsymbol{x}_i, \qquad i \in S_N. \tag{7}$$

Each of these estimators exhibits significant biases in high dimensions. Indeed, the Taylor variance estimators $\widehat{V}_{tay}$ in (4) and its g-weighted version $\widehat{V}_g$ in (5) suffer from important negative biases, while the Jackknife variance estimator $\widehat{V}_{jk}$ in (6) exhibits a large positive bias (Eustache, Dagdoug, and Haziza, 2025). These three estimators are algebraically very similar, yet show very different behaviors. Informally, there are three components providing a partial explanation to these phenomena. First, the high-dimensional behavior of the residuals $(\widehat{\epsilon}_i)_{i\in S_N}$. In high dimensions, these residuals tend to be *artificially* underestimated. Second, the high-dimensional behavior of the leverages differs from their classical, low-dimensional behavior. Specifically, when $p_N = p$ is fixed, then it can be shown under mild conditions that $\max_{i\in S_N} h_{ii,N} \to 0$ in probability as $N \to \infty$. This does not hold in high dimension anymore since

$$\max_{i\in S_N} h_{ii,N} \geq \frac{1}{n_N} \sum_{i\in S_N} h_{ii,N} = \kappa_N \xrightarrow[N\to\infty]{\mathbb{P}} \kappa_\star > 0.$$

For example, in the extreme case where $p_N = n_N$, then $\widehat{\epsilon}_i = 0$ for all $i \in S_N$. This holds even in cases where all covariates $X_1, ..., X_p$ are independent of Y and thus have no "real predictive power". This is intimately linked to the behavior of the leverages. Indeed, for the sake of illustration, consider the unweighted OLS estimator where $\mathbb{V}_m(\widehat{\epsilon}_i) = \sigma^2(1 - h_{ii,N})$. Then, unless $h_{ii,N} \to 0$ as $N \to \infty$, which may not hold in high dimensions, then the second moment of $\widehat{\epsilon}_i$ does not match that of ϵ_i since $\mathbb{V}_m(\epsilon_i) = \sigma^2$; in a low-dimensional setting with $p_N = p$ fixed, their second moment would match, asymptotically. Finally, the high-dimensional behavior of the GREG estimator, in particular its variance, is fundamentally different from that in the low-dimensional case; see, for example, the simulation study of Dagdoug, Goga, and Haziza (2023).

An informal summary reads as follows: naive $\widehat{\epsilon}_i$ leads to an underestimation (Taylor), multiplying them by $g_{i,N}$ (g-weighted) still leads to an underestimation, although less severe, and, if we were to divide by $1 - h_{ii,N}$ (Jackknife – albeit a negligible factor), we would obtain an overestimation. This suggests considering an estimator that corrects the in-sample residuals by the factor $1 - h_{ii,N}$, while avoiding the additional g-weighting present in the jackknife estimator. This approach would lead to the following estimator of the design variance:

$$\widehat{V}_{loo} = \frac{1}{N^2} \sum_{i\in S_N} \sum_{j\in S_N} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\widehat{\epsilon}_i}{(1 - h_{ii,N})\pi_i} \frac{\widehat{\epsilon}_j}{(1 - h_{jj,N})\pi_j}. \tag{8}$$

Although our discussion motivating this estimator was informal, it is a very natural estimator in this regime and is known in the literature. Specifically, recognizing that the residual $\widehat{\epsilon}_i^{(i)}$ of element $i \in S_N$ that would have been obtained if $\widehat{\boldsymbol{\beta}}_N$ were fitted without element i satisfies $\widehat{\epsilon}_i^{(i)} = \widehat{\epsilon}_i/(1 - h_{ii,N})$, we observe that $\widehat{V}_{loo}$ corresponds the Leave-One-Out (LOO) variance estimator replacing potentially overfitted residuals $(\widehat{\epsilon}_i)_{i\in S_N}$ by the leave-one-out cross-validation residuals $(\widehat{\epsilon}_i^{(i)})_{i\in S_N}$. This estimator traces back to Opsomer and Miller (2005) which was then used for hyper-parameter tuning. This also corresponds to a particular case of the cross-validated variance estimator in Dagdoug, Goga, and Haziza (2023) and the crossfitted variance estimator in An, Dagdoug, and Haziza (2026). The high-dimensional behavior of the LOO one of the primary interests of this article.

3 Main results

3.1 Regularity conditions and uniform convergence of sample leverages

The results presented in this section will be proved under the following regularity conditions.

(A1) The sequence of sampling designs $(\mathcal{P}_N)_{N\in\mathbb{N}}$ is a sequence of Bernoulli designs with parameters $(\pi_N)_{N\in\mathbb{N}}$ with $\lim_{N\to\infty}\pi_N := \pi_\star$ and $\lim_{N\to\infty} N\pi_N = \infty$.

(A2) The leverages $(h_{ii,N})_{i\in S_N, N\in\mathbb{N}}$ satisfy

$$\max_{i\in S_N}\left|h_{ii,N} - \kappa_N\right| \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Assumption (A1) is essentially used to make mathematical arguments more tractable. Although restrictive, we believe that similar conclusions to those presented below would hold in closely related sampling designs such as simple random sampling without replacement. This is supported by the simulations presented in Section 4. In general, the aim of the article is to improve our understanding of the phenomena studied in settings where interpretable theoretical results can be derived.

Assumption (A2) is a statement about the uniform convergence of the leverages to $\kappa_\star$ and was initially formulated in Eustache, Dagdoug, and Haziza (2025). In a low-dimensional setting where $p_N = p$ is fixed, this result is easily established. In a high-dimensional setting, it is known that, for each $i \in S_N$, $h_{ii,N} = \kappa_N + o_{\mathbb{P}}(1)$ (El Karoui and Purdom, 2018). A uniform statement, such as that of Assumption (A2), is stronger and allows for a deeper investigation of the asymptotic bias of the LOO variance estimator. When $p_N^{3/2}\log(n_N)/n_N \to 0$ but $p_N/\log(n_N) \to \infty$ as $N \to \infty$, Lemma 3.2 of Portnoy (1987) shows that Assumption (A2) holds in elliptical distributions. However, the above conditions imply $p_N/n_N \to 0$ as $N \to \infty$, which thus does not include the case $\kappa_\star > 0$. We investigate this scenario in the next lemma. To do so, we also consider the following assumption on the distribution of the covariates.

(C1) The covariates include an intercept X_0 and p_N covariates $X_1, X_2, ..., X_{p_N}$ that are independent with Gaussian distribution $\mathcal{N}(0,1)$.

Assumption (C1) considers the case where each covariate is independent of the others, with a standard normal distribution. Although restrictive, these types of distributional assumptions on the distribution of the covariates are often needed to establish the high-dimensional behavior of various statistics and are common in the high-dimensional literature. Similar approaches were used, for example, in Portnoy (1987) or Jiang et al. (2025), more recently. We note that Assumption (C1) will only be used for some of our results. We conjecture that our results might hold for other distributions, although different proof strategies would be needed. For example, our simulation study includes covariates drawn from a uniform distribution, for which the empirical behavior of the estimators remains consistent with our theoretical findings.

Lemma 1. *Assume (A1) and (C1). Then, (A2) holds, that is,*

$$\max_{i\in S_N}\left|h_{ii,N} - \kappa_N\right| \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Proof. See Appendix C.1. □

Lemma 1, which could be of independent interest, shows that uniform convergence of the leverages extends to the case $\kappa_\star > 0$, assuming a Gaussian design. In particular, this shows that Assumption (A2) holds under appropriate settings.

3.2 Asymptotically unbiased variance estimation

In the next result, we determine a closed-form expression for the asymptotic bias of the LOO variance estimator.

Result 1. *Assume (A1) and (A2). Then,*

$$\frac{\mathbb{E}_m\left(\widehat{V}_{loo}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = \left(\frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star\right)\left(\frac{1}{N}\sum_{i\in U_N} g_{i,N}\right)^{-1} + o_{\mathbb{P}}(1). \tag{9}$$

Proof. See Appendix B.1. □

Although interesting, the expression (9) is difficult to interpret since it depends on the behavior of the population average of the g-weights

$$\overline{G}_N := \frac{1}{N}\sum_{i\in U_N} g_{i,N}.$$

This quantity arises naturally in the high-dimensional asymptotic analysis of various variance estimators (Eustache, Dagdoug, and Haziza, 2025). Our next lemma analyzes its low and high-dimensional behavior.

Lemma 2. *Assume (A1). Then, the following statements hold.*

(a) The average g-weights $\overline{G}_N$ satisfies

$$\liminf_{N\to\infty} \overline{G}_N \geq 1 + o_{\mathbb{P}}(1).$$

(b) Assume that

$$\left\| \frac{1}{N}\sum_{i\in S_N}\frac{\boldsymbol{x}_i}{\pi_i} - \frac{1}{N}\sum_{i\in U_N}\boldsymbol{x}_i \right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{p_N}{N}}\right), \tag{10}$$

and that there exists a constant $c > 0$ such that

$$\lim_{N\to\infty} \mathbb{P}\left(\lambda_{\min}(\boldsymbol{A}_\Pi/N) \geq c\right) = 1.$$

Then, provided that $\sqrt{\kappa_N}\max_{i\in U_N}||\boldsymbol{x}_i|| = o_{\mathbb{P}}(1)$, it holds that

$$\max_{i\in U_N}|g_{i,N} - 1| \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Consequently,

$$\overline{G}_N \xrightarrow[N\to\infty]{\mathbb{P}} 1.$$

(c) Assume (C1). Then,

$$\overline{G}_N \xrightarrow[N\to\infty]{\mathbb{P}} \frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star.$$

Proof. See Appendix C.2. □

Remark 2. *Part (b) relies on the condition* $\sqrt{\kappa_N}\max_{i\in U_N}\|\boldsymbol{x}_i\| = o_{\mathbb{P}}(1)$

which holds in low-dimensional regimes where $\kappa_N \to 0$*, but fails when* $\kappa_N \to \kappa_\star > 0$*. In the Gaussian setting of (b), using Theorem 3.1.1 of Vershynin (2026), it can be shown that*

$$\max_{1\le i\le N}\|x_i\|_2 = \sqrt{p_N} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\log N}\right).$$

so that $\sqrt{\kappa_N}\max_{i\in U_N}\|\boldsymbol{x}_i\|$ *vanishes only when* $p_N^2/n_N \to 0$*, which does not hold when* $\kappa_* > 0$*. Part (c) characterizes the limit of* $\overline{G}_N$ *in this regime.*

Lemma 2 shows that the g-weights have different behaviors in low and high dimensions. In general, (a) shows that the (low and high-dimensional) limit of $\overline{G}_N$ is always greater or equal to 1. In the low-dimensional case, the g-weights converge uniformly to 1, which explains why the Taylor variance estimator $\widehat{V}_{tay}$ and its g-weighted version $\widehat{V}_g$ in (4) and (5), respectively, share the same asymptotic behavior and performances. When $\kappa_\star > 0$, on the other hand, the behavior of the average weights $\overline{G}_N$ changes from converging to 1 to converging to a function of $\kappa_\star$, often greater than 1; this is highlighted in (ii) in the Gaussian case.

Our next corollary leverages the high-dimensional characterization of $\overline{G}_N$ to derive closed-form formulas for the asymptotic biases of $\widehat{V}_{tay}$,$\widehat{V}_g$ and $\widehat{V}_{loo}$.

Corollary 1. *Assume (A1) and (C1). Then, the following statements hold.*

(i) The estimator $\widehat{V}_{tay}$ *satisfies*

$$\frac{\mathbb{E}_m\left(\widehat{V}_{tay}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} \xrightarrow[N\to\infty]{\mathbb{P}} \frac{(1-\kappa_\star)(1-\kappa_\star(1-\pi_\star))}{1-\pi_\star\kappa_\star}. \tag{11}$$

(ii) The estimator $\widehat{V}_g$ *satisfies*

$$\frac{\mathbb{E}_m\left(\widehat{V}_g\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} \xrightarrow[N\to\infty]{\mathbb{P}} \frac{(1-\kappa_\star)(1-\pi_\star\kappa_\star(1-\pi_\star))}{1-\pi_\star\kappa_\star}. \tag{12}$$

(iii) The estimator $\widehat{V}_{loo}$ *satisfies*

$$\frac{\mathbb{E}_m\left(\widehat{V}_{loo}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} \xrightarrow[N\to\infty]{\mathbb{P}} 1.$$

Proof. See Appendix B.2. □

The above corollary, in (iii), shows that, in the high-dimensional Gaussian regime, the LOO variance estimator is asymptotically unbiased. $\overline{G}_N$ precisely cancels the bias term in (9). This is therefore an attractive feature of this estimator, which is known to work well in similar scenarios where residuals

overfitting may happen (Opsomer and Miller, 2005; Dagdoug, Goga, and Haziza, 2023). On the other hand, (i) and (ii) highlight that, in general, when $\kappa_\star > 0$, the Taylor variance estimator $\widehat{V}_{tay}$ and its g-weighted version $\widehat{V}_g$ are negatively biased. Although their bias ratios may be somewhat difficult to interpret, the following observations can be made. If either $\kappa_\star = 0$ (low-dimensional case) or $\pi_\star = 1$ (only a negligible part of the population is non-sampled), then both bias ratios are equal to 1, meaning that the estimators are asymptotically unbiased in these settings. However, as soon as $\pi_* < 1$ and $\kappa_\star > 0$, both bias ratios are strictly less than 1, implying that both estimators $\widehat{V}_{tay}$ and $\widehat{V}_g$ are asymptotically (negatively) biased, and, under appropriate conditions, inconsistent. Moreover, for any $\pi_\star < 1$, both bias ratios are decreasing as $\kappa_\star$ increases, showing that higher dimensional problems lead to more severe variance underestimations for $\widehat{V}_{tay}$ and $\widehat{V}_g$.

4 A simulation study: empirical and theoretical behaviors

In this section, we present the results of a simulation study comparing the finite-sample performance of the variance estimators $\widehat{V}_{\text{tay}}$, $\widehat{V}_{\text{g}}$, and $\widehat{V}_{loo}$ with the asymptotic benchmark developed in this article.

We generated a finite population of size $N = 1000$ with $X_1, \ldots, X_{p_N} \overset{i.i.d.}{\sim} \mathcal{U}([-1, 1])$. To study the effect of the dimension, we considered scenarios with a varying number of covariates, $p_N = \lfloor \kappa_N \cdot \pi_N \cdot N \rfloor$, where $\kappa_N \in \{0.1, 0.2, \ldots, 0.8\}$ and $\pi_N = n_N/N = 0.3$.

For each scenario, we performed a Monte Carlo simulation study with $B = 1000$ repetitions of the following steps.

(i) Generating the variable of interest Y according to

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i \in \mathcal{U}_N,$$

where $\boldsymbol{\beta} = \mathbf{1}_{p+1}$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and $\boldsymbol{x}_i \perp\!\!\!\perp \epsilon_i$ for $i \in \mathcal{U}_N$.

(ii) Drawing a sample S_N according to either Simple Random Sampling Without Replacement (SRSWOR) or Bernoulli sampling.

(iii) Computing the GREG estimator $\widehat{\mu}_{greg}$ and the three variance estimators considered: the leave-one-out estimator $\widehat{V}_{loo}$, the standard Taylor estimator $\widehat{V}_{tay}$, and the g-weighted Taylor estimator $\widehat{V}_g$.

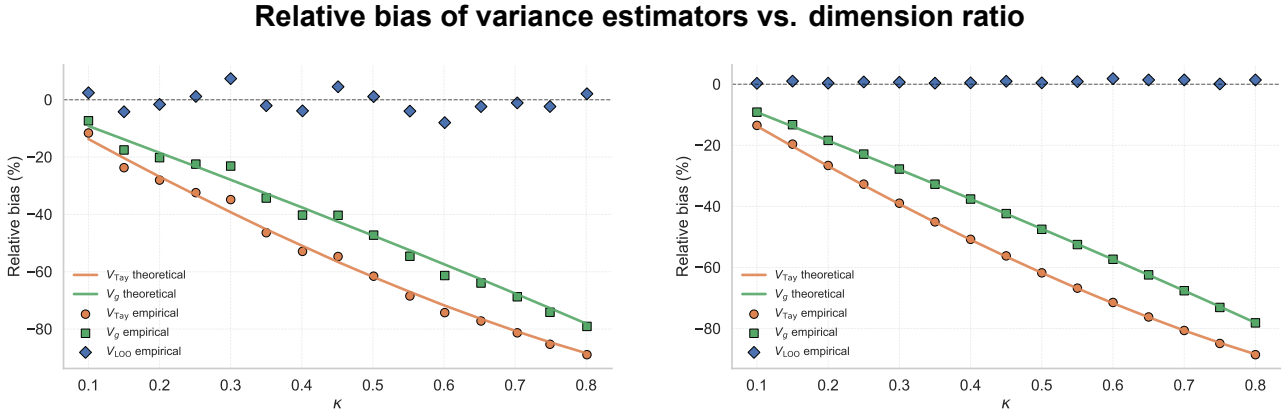
All three estimators were evaluated relative to the design-based variance of $\hat{\mu}_{\text{greg}}$. For each estimator $\widehat{V} \in \{\widehat{V}_{loo}, \widehat{V}_{tay}, \widehat{V}_g\}$, we computed their Monte-Carlo Relative Bias (RB) defined as

$$\text{RB}(\widehat{V}) = 100 \times \frac{1}{B} \sum_{b=1}^{B} \frac{\widehat{V}^{(b)} - V_{MC}}{V_{MC}} \quad (\%),$$

where V_{MC} denotes the Monte Carlo (MC) approximation of the total variance of $\hat{\mu}_{\text{greg}}$, computed using another MC experiment using B repetitions. A value $\text{RB}(\widehat{V}) = 0$ corresponds to empirical unbiasedness, while negative values indicate underestimation and positive values indicate overestimation, all expressed in percentages.

To assess the quality of the technique of analysis used in the article, and the usefulness of the formulas provided by Corollary 1, we compared the empirical biases of the variance estimators with the theoretical asymptotic biases given in Corollary 1. Figure 1 displays $\text{RB}(\widehat{V}_{loo})$, $\text{RB}(\widehat{V}_{tay})$, and $\text{RB}(\widehat{V}_g)$

as functions of κ_N.



(a) Bernoulli Sampling

(b) SRSWOR Sampling

Figure 1: Relative bias (in %) of the variance estimators $\widehat{V}_{loo}$, $\widehat{V}_{tay}$, and $\widehat{V}_g$ as a function of $\kappa_N = p_N/\mathbb{E}(n_N)$ under Bernoulli and SRSWOR sampling designs. The empirical relative biases were represented by the points, while the solid lines correspond to the theoretical curves derived from the asymptotic expressions in (11) and (12).

Overall, the results for both sampling designs were in line with the asymptotic results summarized in Corollary 1. The relative bias of $\widehat{V}_{loo}$ remained close to zero for all values of κ_N, without requiring any debiasing step depending on the dimension. This seems to confirm that, for SRSWOR and Bernoulli, $\widehat{V}_{loo}$ is a reliable variance estimator of $\widehat{\mu}_{greg}$, independently of the dimension. In contrast, $\widehat{V}_{tay}$ and $\widehat{V}_g$ both showed an increasingly negative bias as κ_N increased. In particular, the solid curves in Figure 1, which represent the theoretical formulas for the asymptotic biases of $\widehat{V}_{tay}$ and $\widehat{V}_g$, agreed well with their empirical behavior. This held for both Bernoulli sampling and SRSWOR sampling.

Finally, we note that similar overall patterns were observed for both sampling designs when using uniform covariates, even though the theory was proved under the Gaussian setting. This suggests that our results may be robust to more general settings.

5 Final remarks

In this article, we studied variance estimation for the GREG estimator in regimes where the number of covariates is not necessarily negligible compared with the sample size. Our results show that the three variance estimators considered, although closely related in form, can have very different high-dimensional behaviors. Under the assumptions of Corollary 1, the leave-one-out estimator $\widehat{V}_{loo}$ is asymptotically unbiased in all dimensional regimes considered. In particular, it does not require any debiasing step depending on the dimension. This is to be contrasted with the standard Taylor estimator $\widehat{V}_{tay}$ and the g-weighted Taylor estimator $\widehat{V}_g$, which are negatively biased as soon as $\kappa^\star > 0$ and $\pi^\star < 1$, and this bias becomes more severe as the dimension increases.

The simulation study supports these theoretical findings. It also suggests that the conclusions may be more robust than what is covered by our current proofs. Indeed, similar patterns were observed with uniform covariates, although the main theoretical results were proved under Gaussian covariates. Moreover, the same qualitative behavior was obtained for Bernoulli sampling and SRSWOR. This leads us to believe that the interpretation of the results may extend, at least in spirit, to more general distributional settings and to other sampling designs with equal inclusion probabilities.

Several questions remain open. In particular, the present analysis does not cover unequal inclusion probability designs. In such settings, the behavior of the leverages, the g-weights, and the leave-one-out residuals may be quite different, and it is unclear whether the cancellation mechanism leading to the unbiasedness of $\widehat{V}_{loo}$ would still hold. This is a promising research direction for future work.

References

An, Z., M. Dagdoug, and D. Haziza (2026). Agnostic Model-Assisted Estimation with Machine Learning for Survey Data. Preprint available upon request.

Berger, Y. G. and C. J. Skinner (2005). A jackknife variance estimator for unequal probability sampling. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.1, pp. 79–89.

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

Cassel, C.-M., C.-E. Särndal, and J.-H. Wretman (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons.

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

Dagdoug, M., C. Goga, and D. Haziza (2023). Model-assisted estimation through random forests in finite population sampling. In: *Journal of the American Statistical Association* 118.542, pp. 1234–1251.

Duchesne, P. (2000). A note on jackknife variance estimation for the general regression estimator. In: *Journal of Official Statistics* 16.2, p. 133.

El Karoui, N. and E. Purdom (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. In: *Journal of Machine Learning Research* 19.5, pp. 1–66.

Eustache, E., M. Dagdoug, and D. Haziza (2025). On high-dimensional variance estimation in survey sampling. In: *Scandinavian Journal of Statistics* 52.2, pp. 924–959.

Isaki, C. T. and W. A. Fuller (1982). Survey design under the regression superpopulation model. In: *Journal of the American Statistical Association* 77.377, pp. 89–96.

Jiang, K., R. Mukherjee, S. Sen, and P. Sur (2025). A new central limit theorem for the augmented IPW estimator: Variance inflation, cross-fit covariance and beyond. In: *The Annals of Statistics* 53.2, pp. 647–675.

Kott, P. S. (1990). Estimating the conditional variance of a design consistent regression estimator. In: *Journal of Statistical Planning and Inference* 24.3, pp. 287–296.

Opsomer, J. and C. Miller (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. In: *Nonparametric Statistics* 17.5, pp. 593–611.

Portnoy, S. (1987). A central limit theorem applicable to robust regression estimators. In: *Journal of multivariate analysis* 22.1, pp. 24–50.

Robinson, P. and C.-E. Särndal (1983). Asymptotic properties of the generalized regression estimator in probability sampling. In: *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 240–248.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. New York, NY: Springer-Verlag.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. In: *Biometrika* 76.3, pp. 527–537.

Särndal, C.-E. and R. L. Wright (1984). Cosmetic form of estimators in survey sampling. In: *Scandinavian Journal of Statistics*, pp. 146–156.

Skorski, M. (2023). Bernstein-type bounds for beta distribution. In: *Modern Stochastics: Theory and Applications* 10.2, pp. 211–228.

Stefan, M. and M. A. Hidiroglou (2023). A bootstrap variance procedure for the generalised regression estimator. In: *International Statistical Review* 91.2, pp. 294–317.

Stefan, M. and M. A. Hidiroglou (2024). Jackknife bias-corrected generalized regression estimator in survey sampling. In: *Journal of Survey Statistics and Methodology* 12.1, pp. 211–231.

Ta, T., J. Shao, Q. Li, and L. Wang (2020). Generalized regression estimators with high-dimensional covariates. In: *Statistica Sinica* 30.3, p. 1135.

Valliant, R. (2002). Variance estimation for the general regression estimator. In: *Survey methodology* 28.1, pp. 103–108.

Vershynin, R. (2026). *High-Dimensional Probability: An Introduction with Applications in Data Science*. 2nd ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zhao, Q. and E. J. Candes (2022). An adaptively resized parametric bootstrap for inference in high-dimensional generalized linear models. In: *arXiv preprint arXiv:2208.08944*.

A Additional notation

Linear algebra. The i-th canonical basis vector of $\mathbb{R}^n$ is written $\mathbf{e}_i$. We also denote by $\mathbf{1}_n := [1, ..., 1]^\top \in \mathbb{R}^n$ the vector of ones in $\mathbb{R}^n$. The span of a vector $\mathbf{u}$ is written span$(\mathbf{u})$ and the column space of a matrix $\boldsymbol{X}$ is denoted Col$(\boldsymbol{X})$. For a subspace $\mathcal{S} \subset \mathbb{R}^n$, we denote by $\boldsymbol{P}(\mathcal{S})$ the orthogonal projection onto $\mathcal{S}$. In particular, for a matrix $\boldsymbol{X}$, we write $\boldsymbol{P}(\boldsymbol{X}) := \boldsymbol{P}(\text{Col}(\boldsymbol{X})) = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. We denote by $s_{\min}(\boldsymbol{X})$ the smallest singular value of a matrix $\boldsymbol{X}$, and by $\lambda_{\min}(\boldsymbol{X})$ the smallest eigenvalue of a square matrix $\boldsymbol{X}$. We also use the following notation:

$E^\perp := \{x \in V :< x, e >= 0, \forall e \in E\}$, where V is a vector space with inner product $< ., . >$.

$\oplus$ means that two spaces are in direct sum (i.e., every element of the surrounding vector space V can be uniquely decomposed into an element of, say, E and an element of F, if $E \oplus F = V$).

$\mathbf{g} \perp\!\!\!\perp \boldsymbol{H}$ means that the random vector g is independent (in the probabilistic sense) of the random matrix H.

Probability distributions. We denote by $\stackrel{d}{=}$ equality in distribution. We write: $Ber(p)$ for the Bernoulli distribution with parameter p; $Bin(n, p)$ for the Binomial distribution with parameters (n, p); $\mathcal{N}(\boldsymbol{\mu},)$ for the multivariate normal distribution; χ^2_k for the chi-square distribution with k degrees of freedom; $\mathcal{B}(\alpha, \beta)$ for the Beta distribution with parameters (α, β); $\mathcal{W}_p(n,)$ for the Wishart distribution with n degrees of freedom and scale matrix ; F_{d_1,d_2} for the Fisher distribution with (d_1, d_2) degrees of freedom; and $T^2_{p,n}$ for the Hotelling distribution with parameters (p, n).

Specific notation. The covariates $\boldsymbol{x}_i = (1, X_1, \ldots, X_p)$ decomposes as $\boldsymbol{x}_i = (1, \widetilde{\boldsymbol{x}}_i)$. Define the matrix $\boldsymbol{X}_S \in \mathbb{R}^{n\times(p+1)}$ as the matrix whose i-th row is $\boldsymbol{x}_i$ and $\widetilde{\boldsymbol{X}}_S \in \mathbb{R}^{n\times p}$ as the matrix whose i-th row is $\widetilde{\boldsymbol{x}}_i$. Similarly, define $\mathbf{t}_{x,S_N} = \sum_{i\in S_N} \boldsymbol{x}_i$, $\mathbf{t}_{x,S^c_N} = \sum_{i\in S^c_N} \boldsymbol{x}_i$, $\boldsymbol{A}_S = \sum_{i\in S_N} \boldsymbol{x}_i\boldsymbol{x}_i^\top$ and $\tilde{\mathbf{t}}_{x,S} = \sum_{i\in S_N} \widetilde{\boldsymbol{x}}_i$, $\widetilde{\boldsymbol{A}}_S = \sum_{i\in S_N} \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^\top$, with $S^c_N = U_N \setminus S_N$.

Remark 3. *In the following, although the sample size n_N is random, a law of large numbers argument shows that $n_N/n_{exp,N} \xrightarrow[N\to\infty]{\mathbb{P}} 1$, where $n_{exp,N} := N\pi_N$ denotes the expected sample size. For simplicity*

of notation, we sometimes give asymptotic orders in terms of n_N, instead of its deterministic equivalent $n_{exp,N}$. This abuse of notation however does not affect our results.

B Proofs of the main results

B.1 Proof of Result 1

Recall from Eustache, Dagdoug, and Haziza (2025) that

$$\mathbb{V}_m\left(\widehat{\mu}_{\text{greg}}\right) = \frac{\sigma^2}{\pi_N N^2} \sum_{i \in U_N} g_{i,N}.$$

Under Bernoulli sampling, the estimator $\widehat{V}_{loo}$ reduces to

$$\widehat{V}_{loo} = \frac{1}{N^2} \frac{1-\pi_N}{\pi_N^2} \sum_{i \in S_N} \frac{\widehat{\epsilon}_i^2}{(1-h_{ii,N})^2} + \frac{\sigma^2}{N}.$$

Moreover, $\mathbb{V}_m(\widehat{\epsilon}_i) = \sigma^2(1-h_{ii,N})$, so that, taking expectation on both sides yields

$$\mathbb{E}_m\left(\widehat{V}_{loo}\right) = \frac{\sigma^2}{N^2} \frac{1-\pi_N}{\pi_N^2} \sum_{i \in S_N} \frac{1}{1-h_{ii,N}} + \frac{\sigma^2}{N}. \tag{13}$$

For $i \in S_N$, a first-order Taylor expansion of $(1-h_{ii,N})^{-1}$ around κ_N gives

$$\frac{1}{1-h_{ii,N}} = \frac{1}{1-\kappa_N} + \frac{h_{ii,N}-\kappa_N}{(1-\kappa_N)^2} + \mathcal{O}_{\mathbb{P}}\big(\max_{i \in S_N} |h_{ii,N}-\kappa_N|^2\big).$$

Averaging over $i \in S_N$ gives

$$\frac{1}{n_N} \sum_{i \in S_N} \frac{1}{1-h_{ii,N}} = \frac{1}{1-\kappa_N} + \frac{1}{n_N} \sum_{i \in S_N} \frac{h_{ii,N}-\kappa_N}{(1-\kappa_N)^2} + \mathcal{O}_{\mathbb{P}}\big(\max_{i \in S_N} |h_{ii,N}-\kappa_N|^2\big).$$

Under (A2) and using that $\kappa_N \xrightarrow[N\to\infty]{} \kappa_\star$, this reduces to

$$\frac{1}{n_N} \sum_{i \in S_N} \frac{1}{1-h_{ii,N}} = \frac{1}{1-\kappa_\star} + o_{\mathbb{P}}\left(1\right).$$

Consequently, using (A1), equation (13) can be written

$$\mathbb{E}_m\left(\widehat{V}_{loo}\right) = \frac{\sigma^2}{\pi_N N} \left(\frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star\right) + o_{\mathbb{P}}\left(N^{-1}\right).$$

Finally, combining the two previous expressions leads to

$$\frac{\mathbb{E}_m\left(\widehat{V}_{loo}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = \left(\frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star\right) \left(\frac{1}{N} \sum_{i \in U_N} g_{i,N}\right)^{-1} + o_{\mathbb{P}}(1),$$

since $(N^{-1}\sum_{i \in U_N} g_{i,N})^{-1} = \mathcal{O}_{\mathbb{P}}(1)$ by an application of Lemma 2.

B.2 Proof of Corollary 1

B.2.1 Proof of (i)

From Result 4.1 of Eustache, Dagdoug, and Haziza (2025), we have

$$\frac{\mathbb{E}_m\left(\widehat{V}_{tay}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = \frac{(1-\pi_\star)(1-\kappa_\star)+\pi_\star}{\overline{G}_N} + o_{\mathbb{P}}(1).$$

Combining this with Lemma 2 (c) yields (i).

B.2.2 Proof of (ii)

From Result 4.1 of Eustache, Dagdoug, and Haziza (2025), we have

$$\frac{\mathbb{E}_m\left(\widehat{V}_{g}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = (1-\pi_\star)\left(1 - \frac{n_N^{-1}\sum_{i\in S_N} g_{i,N}^2 h_{ii,N}}{\overline{G}_N}\right) + \frac{\pi_\star}{\overline{G}_N} + o_{\mathbb{P}}(1). \tag{14}$$

Let us study the term $n_N^{-1}\sum_{i\in S_N} g_{i,N}^2 h_{ii,N}$. From (A2), we can write

$$h_{ii,N} = \kappa_N + r_{i,N}, \qquad \text{where } r_{i,N} := h_{ii,N} - \kappa_N, \quad \text{with} \quad \max_{i\in S_N}|r_{i,N}| = o_{\mathbb{P}}(1).$$

Hence

$$\frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2 h_{ii,N} = \kappa_N \frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2 + \frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2 r_{i,N}.$$

Moreover,

$$\left|\frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2 r_{i,N}\right| \le \max_{i\in S_N}|r_{i,N}|\,\frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2.$$

By Technical Lemma 2 of Eustache, Dagdoug, and Haziza (2025),

$$\frac{1}{n_N}\sum_{i\in S_N} g_{i,N}^2 = \frac{\pi_N}{n_N}\sum_{i\in U_N} g_{i,N} = \overline{G}_N.$$

Since in our setting $\overline{G}_N = \dfrac{1-\pi_\star}{1-\kappa_\star} + \pi_\star + o_{\mathbb{P}}(1)$, we have

$$n_N^{-1}\sum_{i\in S_N} g_{i,N}^2 h_{ii,N} = \kappa_N\left(\frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star\right) + o_{\mathbb{P}}(1).$$

Finally, replacing $\overline{G}_N$ and $n_N^{-1}\sum_{i\in S_N} g_{i,N}^2 h_{ii,N}$ in (14), we obtain

$$\frac{\mathbb{E}_m\left(\widehat{V}_{g}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = \frac{(1-\kappa_\star)(1-\pi_\star\kappa_\star(1-\pi_\star))}{1-\pi_\star\kappa_\star} + o_{\mathbb{P}}(1).$$

B.2.3 Proof of (iii)

The result follows directly by combining Result 1 and Lemma 2(c).

C Proof of lemmas

C.1 Proof of Lemma 1

Recall that $\boldsymbol{X}_S \in \mathbb{R}^{n\times(p+1)}$ denotes the matrix whose i-th row is $\boldsymbol{x}_i$, so that $\boldsymbol{x}_i = \boldsymbol{X}_S^\top \boldsymbol{e}_i$. Therefore, for $i \in S_N$,

$$h_{ii,N} = \boldsymbol{x}_i^\top (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1} \boldsymbol{x}_i = \boldsymbol{e}_i^\top \boldsymbol{X}_S (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^\top \boldsymbol{e}_i = \boldsymbol{e}_i^\top \boldsymbol{P}(\boldsymbol{X}_S)\boldsymbol{e}_i.$$

Since $\boldsymbol{X}_S = [\mathbf{1}_n, \widetilde{\boldsymbol{X}}_S]$, we can apply Lemma 5 with $E = \text{span}(\mathbf{1}_n)$ of dimension 1, $\boldsymbol{G} = \widetilde{\boldsymbol{X}}_S$ and $\mathbf{a} = \mathbf{e}_i$. Since $\boldsymbol{P}(E)\mathbf{e}_i = \mathbf{1}_n/n_N$ we have $\|\boldsymbol{P}(E)\mathbf{e}_i\|_2^2 = 1/n_N$. Moreover, by the Pythagorean theorem,

$$\|\boldsymbol{P}(E^\perp)\mathbf{e}_i\|_2^2 = \|\mathbf{e}_i\|_2^2 - \|\boldsymbol{P}(E)\mathbf{e}_i\|_2^2 = 1 - \frac{1}{n_N}.$$

Therefore, Lemma 5 gives

$$h_{ii,N} \overset{d}{=} \frac{1}{n_N} + \left(1 - \frac{1}{n_N}\right) B_i, \qquad \text{where } B_i \mid n_N \sim \mathcal{B}\left(\frac{p_N}{2}, \frac{n_N - 1 - p_N}{2}\right).$$

It remains to bound $\max_{i\in S_N} \left|h_{ii,N} - p_N/n_N\right|$ to conclude the proof. Note that

$$\begin{aligned} h_{ii,N} - \frac{p_N}{n_N} &= \frac{1}{n_N} + \left(1 - \frac{1}{n_N}\right) B_i - \frac{p_N}{n_N} \\ &= \frac{1}{n_N} + \left(1 - \frac{1}{n_N}\right) (B_i - \mathbb{E}(B_i \mid n_N)), \quad \text{where } \mathbb{E}(B_i \mid n_N) = \frac{p_N}{n_N - 1}. \end{aligned}$$

Therefore,

$$\max_{i\in S_N} \left|h_{ii,N} - \frac{p_N}{n_N}\right| \le \frac{1}{n_N} + \left(1 - \frac{1}{n_N}\right) \max_{i\in S_N} |B_i - \mathbb{E}(B_i \mid n_N)|.$$

So it suffices to show that $\max_{i\in S_N} |B_i - \mathbb{E}(B_i \mid n_N)| = o_{\mathbb{P}}(1)$. It follows from Theorem 1 of Skorski (2023) that

$$\mathbb{P}\{|B_i - \mathbb{E}(B_i \mid n_N)| > \epsilon\} \le 2\exp\left(-\frac{\epsilon^2}{2\left(v + \frac{|c|\epsilon}{3}\right)}\right).$$

where

$$v = \frac{p_N(n_N - 1 - p_N)}{(n_N - 1)^2(n_N + 1)} = \mathcal{O}(n_N^{-1}), \qquad c = \frac{4(n_N - 1 - 2p_N)}{(n_N - 1)(n_N + 3)} = \mathcal{O}(n_N^{-1}).$$

Therefore, for every $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ such that

$$\mathbb{P}\left(|B_i - \mathbb{E}(B_i \mid n_N)| > \epsilon \mid n_N\right) \le e^{-C_\epsilon n_N}.$$

Thus, by the union bound over the n_N sampled units,

$$\mathbb{P}\left(\max_{i\in S_N} |B_i - \mathbb{E}(B_i \mid n_N)| > \epsilon \mid n_N\right) \le n_N \mathbb{P}\left(|B_1 - \mathbb{E}(B_1 \mid n_N)| > \epsilon \mid n_N\right) \le n_N e^{-C_\epsilon n_N}.$$

Taking expectation with respect to the sampling design, and using the fact that $n_N \sim Bin(N, \pi_N)$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{i\in S_N}\left|B_i - \mathbb{E}(B_i \mid n_N)\right| > \epsilon\right) &= \mathbb{E}\left[\mathbb{P}\left(\max_{i\in S_N}\left|B_i - \mathbb{E}(B_i \mid n_N)\right| > \epsilon \,\middle|\, n_N\right)\right] \\
&\leq \mathbb{E}\left[n_N e^{-C_\epsilon n_N}\right] \\
&= \sum_{k=0}^{N} k e^{-C_\epsilon k}\binom{N}{k}\pi_N^k(1-\pi_N)^{N-k} \\
&= N\pi_N e^{-C_\epsilon}\sum_{l=0}^{N-1}\binom{N-1}{l}(\pi_N e^{-C_\epsilon})^l(1-\pi_N)^{N-1-l} \\
&= N\pi_N e^{-C_\epsilon}\left(1-\pi_N+\pi_N e^{-C_\epsilon}\right)^{N-1},
\end{aligned}$$

which converges to zero as $N \to \infty$. Hence, $\max_{i\in S_N}|B_i - \mathbb{E}(B_i \mid n_N)| = o_{\mathbb{P}}(1)$. This concludes the proof.

C.2 Proof of Lemma 2

C.2.1 Proof of (a)

We study the term $\overline{G}_N = N^{-1}\mathbf{t}_{\boldsymbol{x}}^\top \boldsymbol{A}_\Pi^{-1}\mathbf{t}_{\boldsymbol{x}}$. To that aim, let $\mathbf{e}_0 := [1, 0, ..., 0]^\top \in \mathbb{R}^N$. Since the intercept is included in the covariates, the following equalities hold:

$$\mathbf{e}_0^\top \mathbf{t}_{\boldsymbol{x}} = N, \qquad \mathbf{e}_0^\top \boldsymbol{A}_\Pi^{-1}\mathbf{e}_0 = \frac{n_N}{\pi_N}.$$

Therefore, by the Cauchy-Schwarz inequality,

$$\left(\mathbf{e}_0^\top \mathbf{t}_{\boldsymbol{x}}\right)^2 = \left\{\left(\boldsymbol{A}_\Pi^{1/2}\mathbf{e}_0\right)^\top \boldsymbol{A}_\Pi^{-1/2}\mathbf{t}_{\boldsymbol{x}}\right\}^2 \leqslant \|\boldsymbol{A}_\Pi^{1/2}\mathbf{e}_0\|_2^2 \|\boldsymbol{A}_\Pi^{-1/2}\mathbf{t}_{\boldsymbol{x}}\|_2^2 = \frac{n_N}{\pi_N}\mathbf{t}_{\boldsymbol{x}}^\top \boldsymbol{A}_\Pi^{-1}\mathbf{t}_{\boldsymbol{x}}.$$

Since $\left(\mathbf{e}_0^\top \mathbf{t}_{\boldsymbol{x}}\right)^2 = N^2$, we obtain

$$\overline{G}_N \geq \frac{N\pi_N}{n_N} = 1 + o_{\mathbb{P}}(1),$$

which shows the result.

C.2.2 Proof of (b)

Recall that from Särndal, Swensson, and Wretman (1992), $g_{i,N} = 1 + (\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}})^\top \boldsymbol{A}_\Pi^{-1}\boldsymbol{x}_i$ where $\widehat{\mathbf{t}}_{\boldsymbol{x}} = \sum_{j\in S_N}\boldsymbol{x}_j/\pi_j$. Therefore,

$$\begin{aligned}
\max_{i\in U_N}|g_{i,N} - 1| &= \max_{i\in U_N}\left|(\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}})^\top \boldsymbol{A}_\Pi^{-1}\boldsymbol{x}_i\right| \\
&\leq \max_{i\in U_N}\|\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}}\| \cdot \|\boldsymbol{A}_\Pi^{-1}\boldsymbol{x}_i\| \\
&\leq \|\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}}\| \cdot \|\boldsymbol{A}_\Pi^{-1}\| \cdot \max_{i\in U_N}\|\boldsymbol{x}_i\|.
\end{aligned}$$

By hypothesis, $\|\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}}\| = \mathcal{O}_{\mathbb{P}}(\sqrt{N p_N})$ and, for N large enough,

$$\|\boldsymbol{A}_{\Pi}^{-1}\| = \frac{1}{\lambda_{\min}(\boldsymbol{A}_{\Pi})} \leq \frac{1}{c\,N} = \mathcal{O}(1/N).$$

Therefore, we obtain $\|\mathbf{t}_{\boldsymbol{x}} - \widehat{\mathbf{t}}_{\boldsymbol{x}}\| \cdot \|\boldsymbol{A}_{\Pi}^{-1}\| = \mathcal{O}_{\mathbb{P}}(\sqrt{\kappa_N})$, which yields

$$\max_{i \in U_N} |g_{i,N} - 1| \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Concerning $\overline{G}_N$, it follows from the triangle inequality that

$$|\overline{G}_N - 1| \leq \frac{1}{N} \sum_{i \in U_N} |g_{i,N} - 1| \leq \max_{i \in U_N} |g_{i,N} - 1| = o_{\mathbb{P}}(1).$$

Therefore, $\overline{G}_N \xrightarrow[N\to\infty]{\mathbb{P}} 1$.

C.2.3 Proof of (c)

Under Bernoulli sampling, we have

$$\begin{aligned}
\overline{G}_N &= \frac{1}{N} \sum_{i \in U_N} \mathbf{t}_{\boldsymbol{x}}^{\top} \boldsymbol{A}_{\Pi}^{-1} \boldsymbol{x}_i \\
&= \frac{\pi_N}{N} \mathbf{t}_{\boldsymbol{x}}^{\top} \boldsymbol{A}_S^{-1} \mathbf{t}_{\boldsymbol{x}} \\
&= \frac{\pi_N}{N} \left(\mathbf{t}_{x,S_N}^{\top} + \mathbf{t}_{x,S_N^c}^{\top} \right) \left(\sum_{i \in S_N} \boldsymbol{x}_i \boldsymbol{x}_i^{\top} \right)^{-1} \left(\mathbf{t}_{x,S_N} + \mathbf{t}_{x,S_N^c} \right) \\
&= \frac{\pi_N}{N} \mathbf{t}_{x,S_N}^{\top} \boldsymbol{A}_S^{-1} \mathbf{t}_{x,S_N} + \frac{2\pi_N}{N} \mathbf{t}_{x,S_N}^{\top} \boldsymbol{A}_S^{-1} \mathbf{t}_{x,S_N^c} + \frac{\pi_N}{N} \mathbf{t}_{x,S_N^c}^{\top} \boldsymbol{A}_S^{-1} \mathbf{t}_{x,S_N^c} \\
&=: A_1(S_N) + A_2(S_N, S_N^c) + A_3(S_N^c).
\end{aligned}$$

We now control each of these terms separately. The terms $A_1(S_N)$ and $A_2(S_N, S_N^c)$ are relatively straightforward algebraically, whereas $A_3(S_N^c)$ will require a more detailed analysis.

First and second term: $A_1(S_N)$ and $A_2(S_N, S_N^c)$. Since $\mathbf{e}_1^{\top} \boldsymbol{x}_j = 1$, for all $j \in U_N$, we obtain

$$A_1(S_N) = \frac{\pi_N}{N} \sum_{j \in S_N} \boldsymbol{x}_j^{\top} \left(\sum_{\ell \in S_N} \boldsymbol{x}_\ell \boldsymbol{x}_\ell^{\top} \right)^{-1} \sum_{i \in S_N} \boldsymbol{x}_i = \frac{\pi_N}{N} \sum_{i \in S_N} \mathbf{e}_1^{\top} \boldsymbol{x}_i = \pi_\star^2 + o_{\mathbb{P}}(1).$$

Similarly, write

$$A_2(S_N, S_N^c) = 2\frac{\pi_N}{N} \sum_{j \in S_N} \boldsymbol{x}_j^{\top} \left(\sum_{\ell \in S_N} \boldsymbol{x}_\ell \boldsymbol{x}_\ell^{\top} \right)^{-1} \sum_{i \in S_N^c} \boldsymbol{x}_i = 2\frac{\pi_N}{N} \sum_{i \in S_N^c} \mathbf{e}_1^{\top} \boldsymbol{x}_i = 2\pi_\star(1 - \pi_\star) + o_{\mathbb{P}}(1).$$

Third term: $A_3(S_N^c)$**.** Recall that $A_3(S_N^c) = \pi_N N^{-1} \mathbf{t}_{x,S_N^c}^\top \boldsymbol{A}_S^{-1} \mathbf{t}_{x,S_N^c}$ and

$$\mathbf{t}_{x,S_N^c} = \sum_{i \in S_N^c} \boldsymbol{x}_i = \begin{pmatrix} N - n_N \\ \sum_{i \in S_N^c} \widetilde{\boldsymbol{x}}_i \end{pmatrix}.$$

We first study $\mathbf{t}_{x,S_N^c}$. Since the covariates $\{\widetilde{\boldsymbol{x}}_i\}_{i \in U_N}$ are independent with distribution $\mathcal{N}(0, \mathbf{I}_p)$, conditionally on n_N, we have

$$\sum_{i \in S_N^c} \widetilde{\boldsymbol{x}}_i \overset{d}{=} \sqrt{N - n_N}\, \widetilde{\mathbf{w}}, \quad \text{where } \widetilde{\mathbf{w}} \sim \mathcal{N}_p(0, \mathbf{I}_p).$$

Thus, defining

$$\mathbf{w} := \begin{pmatrix} 1 \\ \widetilde{\mathbf{w}} \end{pmatrix}, \qquad \mathbf{a}_N := \begin{pmatrix} \sqrt{N(1-\pi_\star)} - 1 \\ \mathbf{0}_p \end{pmatrix},$$

we obtain

$$\begin{aligned} \mathbf{t}_{x,S_N^c} &= \begin{pmatrix} N - n_N \\ \sqrt{N - n_N}\, \widetilde{\mathbf{w}} \end{pmatrix} \\ &= \sqrt{N - n_N} \left(\begin{pmatrix} 1 \\ \widetilde{\mathbf{w}} \end{pmatrix} + \begin{pmatrix} \sqrt{N - n_N} - 1 \\ \mathbf{0}_p \end{pmatrix} \right) \\ &= \sqrt{N(1-\pi_\star)}\, (\mathbf{w} + \mathbf{a}_N) + o_{\mathbb{P}}(N). \end{aligned}$$

Therefore,

$$\begin{aligned} A_3(S_N^c) &= \frac{\pi_\star}{N}\, \mathbf{t}_{x,S_N^c}^\top \boldsymbol{A}_S^{-1} \mathbf{t}_{x,S_N^c} + o_{\mathbb{P}}(1) \\ &= \pi_\star(1-\pi_\star)\, (\mathbf{w} + \mathbf{a}_N)^\top \boldsymbol{A}_S^{-1} (\mathbf{w} + \mathbf{a}_N) + o_{\mathbb{P}}(1) \\ &= \pi_\star(1-\pi_\star) \Big(\mathbf{w}^\top \boldsymbol{A}_S^{-1} \mathbf{w} + 2\, \mathbf{a}_N^\top \boldsymbol{A}_S^{-1} \mathbf{w} + \mathbf{a}_N^\top \boldsymbol{A}_S^{-1} \mathbf{a}_N \Big) + o_{\mathbb{P}}(1) \\ &= \pi_\star(1-\pi_\star)(B_1 + B_2 + B_3) + o_{\mathbb{P}}(1). \end{aligned}$$

Each of these terms will also be studied independently.

First term: B_1**.** It follows directly from Lemma 4 that

$$B_1 = \frac{\kappa_\star}{1 - \kappa_\star} + o_{\mathbb{P}}(1).$$

Second term: B_2**.** Using (15), B_2 simplifies as follows

$$\begin{aligned} B_2 &= \begin{pmatrix} \sqrt{N(1-\pi_\star)} - 1 \\ \mathbf{0}_p \end{pmatrix}^\top \begin{pmatrix} \Delta_N^{-1} & -\Delta_N^{-1} \widetilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \\ -\widetilde{\boldsymbol{A}}_S^{-1} \widetilde{\mathbf{t}}_{x,S} \Delta_N^{-1} & \widetilde{\boldsymbol{A}}_S^{-1} + \widetilde{\boldsymbol{A}}_S^{-1} \widetilde{\mathbf{t}}_{x,S} \Delta_N^{-1} \widetilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \widetilde{\mathbf{w}} \end{pmatrix} \\ &= \big(\sqrt{N(1-\pi_\star)} - 1\big) \Delta_N^{-1} - \big(\sqrt{N(1-\pi_\star)} - 1\big) \Delta_N^{-1} \widetilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \widetilde{\mathbf{w}} \\ &= o_{\mathbb{P}}(1), \end{aligned}$$

where the last equality follows since $\Delta_N^{-1} = (\boldsymbol{A}_S^{-1})_{11} = \mathcal{O}_{\mathbb{P}}(1/n_N)$ by Lemma 3, and $\widetilde{\mathbf{t}}_{x,S}^\top \boldsymbol{A}_S^{-1} \widetilde{\mathbf{w}} = \mathcal{O}_{\mathbb{P}}(1)$, as shown in the proof of Lemma 4.

Third term: B_3**.** By (15) and Lemma 3, B_3 satisfies

$$\begin{aligned} B_3 &= \begin{pmatrix} \sqrt{N(1-\pi_\star)} - 1 \\ \mathbf{0}_p \end{pmatrix}^\top \begin{pmatrix} \Delta_N^{-1} & -\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \\ -\widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\Delta_N^{-1} & \widetilde{\boldsymbol{A}}_S^{-1} + \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \end{pmatrix} \begin{pmatrix} \sqrt{N(1-\pi_\star)} - 1 \\ \mathbf{0}_p \end{pmatrix} \\ &= N(1-\pi_\star)\cdot \Delta_N^{-1} + o_{\mathbb{P}}(1) \\ &= \frac{(1-\pi_\star)}{\pi_N(1-\kappa_\star)} + o_{\mathbb{P}}(1). \end{aligned}$$

Hence,

$$A_3(S_N^c) = \frac{(1-\pi_\star)^2}{1-\kappa_\star} + \pi_\star(1-\pi_\star)\frac{\kappa_\star}{1-\kappa_\star} + o_{\mathbb{P}}(1).$$

Conclusion. Summing the three contributions, we obtain

$$\begin{aligned} \overline{G}_N &= \pi_\star^2 + 2\pi_\star(1-\pi_\star) + \frac{(1-\pi_\star)^2}{1-\kappa_\star} + \pi_\star(1-\pi_\star)\frac{\kappa_\star}{1-\kappa_\star} + o_{\mathbb{P}}(1) \\ &= \frac{1-\pi_\star}{1-\kappa_\star} + \pi_\star + o_{\mathbb{P}}(1). \end{aligned}$$

D Auxiliary lemmas

Lemma 3. *Assume (A1) and (C1). Then,*

$$(\boldsymbol{A}_S^{-1})_{11} = \frac{1}{n_N(1-\kappa_\star)} + o_{\mathbb{P}}\left(n_N^{-1}\right).$$

Proof. We have

$$\boldsymbol{A}_S = \begin{pmatrix} n_N & \tilde{\mathbf{t}}_{x,S}^\top \\ \tilde{\mathbf{t}}_{x,S} & \widetilde{\boldsymbol{A}}_S \end{pmatrix}.$$

By the Schur complement formula, $(\boldsymbol{A}_S^{-1})_{11} = (n_N - \tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S})^{-1}$. Therefore, it suffices to determine the asymptotic behavior of $\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}$. Since $\tilde{\mathbf{t}}_{x,S} = \widetilde{\boldsymbol{X}}_S^\top \mathbf{1}_n$ and $\widetilde{\boldsymbol{A}}_S = \widetilde{\boldsymbol{X}}_S^\top \widetilde{\boldsymbol{X}}_S$, we have $\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S} = \mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n$. Hence, the problem reduces to characterizing the distribution of $\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n$.

Applying Lemma 5 with $E = \{0\}$, $\boldsymbol{G} = \widetilde{\boldsymbol{X}}_S$ and $\mathbf{a} = \mathbf{1}_n$, we get

$$\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n | n_N \overset{d}{=} n_N \mathcal{B}\left(\frac{p_N}{2}, \frac{n_N - p_N}{2}\right),$$

since $\|\boldsymbol{P}(E^\perp)\mathbf{1}_n\|^2 = \|\mathbf{1}_n\|^2 = n_N$. Using the moments of the Beta distribution, we have

$$\mathbb{E}\left(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n \mid n_N\right) = p_N, \quad \mathbb{V}\left(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n \mid n_N\right) = \frac{2\kappa_\star(1-\kappa_\star)n_N^2}{n_N+2} = 2\kappa_\star(1-\kappa_\star)n_N + o_{\mathbb{P}}(n_N).$$

Using the law of total expectation, we obtain $\mathbb{E}(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n) = p_N$. Similarly, by the law of total variance,

$$\begin{aligned} \mathbb{V}(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n) = &\mathbb{V}\left(\mathbb{E}\left(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n \mid n_N\right)\right) \\ &+ \mathbb{E}\left(\mathbb{V}\left(\mathbf{1}_n^\top \boldsymbol{P}(\widetilde{\boldsymbol{X}}_S)\mathbf{1}_n \mid n_N\right)\right) \end{aligned}$$

$$= \mathbb{E}\left(2\kappa_\star(1-\kappa_\star)n_N + o_{\mathbb{P}}(n_N)\right)$$
$$= 2\kappa_\star(1-\kappa_\star)N\pi_\star + o(N), \quad \text{as } n_N \sim Bin(N, \pi_N).$$

Then, by Chebyshev's inequality, it follows that

$$\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S} = p_N + \mathcal{O}_{\mathbb{P}}(n_N^{1/2}), \quad \text{and} \quad n_N - \tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S} = n_N(1-\kappa_\star) + o_{\mathbb{P}}(n_N).$$

Therefore,

$$(\boldsymbol{A}_S^{-1})_{11} = \frac{1}{n_N(1-\kappa_\star)} + o_{\mathbb{P}}(n_N^{-1}).$$

□

Lemma 4. *Assume (A1) and (C1). Define* $\mathbf{z} = [1\ \widetilde{\mathbf{z}}^\top]^\top$ *where* $\widetilde{\mathbf{z}} \sim \mathcal{N}(0, \mathbf{I}_p)$ *and* $\widetilde{\mathbf{z}} \perp\!\!\!\perp \{\boldsymbol{x}_i\}_{i\in U_N}$. *Then,*

$$\mathbf{z}^\top \boldsymbol{A}_S^{-1}\mathbf{z} = \frac{\kappa_\star}{1-\kappa_\star} + o_{\mathbb{P}}(1).$$

Proof. Let $\Delta_N = n_N - \tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}$ be the Schur complement of $\widetilde{\boldsymbol{A}}_S$ in $\boldsymbol{A}_S$. By the block inverse formula,

$$\boldsymbol{A}_S^{-1} = \begin{pmatrix} \Delta_N^{-1} & -\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \\ -\widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\Delta_N^{-1} & \widetilde{\boldsymbol{A}}_S^{-1} + \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1} \end{pmatrix}. \tag{15}$$

It follows that

$$\mathbf{z}^\top \boldsymbol{A}_S^{-1}\mathbf{z} = \Delta_N^{-1} + \widetilde{\mathbf{z}}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}} + \Delta_N^{-1}\left(\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}}\right)^2 - 2\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}} := T_1 + T_2 + T_3 + T_4.$$

We study each term separately and show that, $T_2 = \kappa_\star/(1-\kappa_\star) + o_{\mathbb{P}}(1)$ and that T_1, T_3 and T_4 all converge to zero in probability, from which the result will follow.

First term: $T_1 = \Delta_N^{-1}$**.** A direct application of Lemma 3 shows that $\Delta_N^{-1} = o_{\mathbb{P}}(1)$.

Second term: $T_2 = \widetilde{\mathbf{z}}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}}$**.** Recall that, conditional on S_N, the rows of $\widetilde{\boldsymbol{X}}_S$ are independent $\mathcal{N}(0, \mathbf{I}_p)$ under (C1). Therefore, $\widetilde{\boldsymbol{A}}_S = \widetilde{\boldsymbol{X}}_S^\top \widetilde{\boldsymbol{X}}_S \sim \mathcal{W}_p(\boldsymbol{I}_p, n_N)$ which implies that

$$n_N\, \widetilde{\mathbf{z}}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}} \sim T^2_{p_N, n_N}, \quad \text{and} \quad T_2 \sim \frac{p_N}{n_N - p_N + 1} F_{p_N, n_N - p_N + 1}.$$

Consequently,

$$\mathbb{E}[T_2] = \frac{p_N}{n_N - p_N + 1} \cdot \frac{n_N - p_N + 1}{n_N - p_N - 1} = \frac{\kappa_\star}{1-\kappa_\star} + o(1),$$

$$\mathbb{V}(T_2) = \frac{p_N^2}{(n_N - p_N + 1)^2} \cdot \frac{2(n_N - p_N + 1)^2\left(p_N + n_N - p_N + 1 - 2\right)}{p_N\,(n_N - p_N - 1)^2\,(n_N - p_N - 3)} = \frac{2}{n_N}\,\frac{\kappa_\star}{(1-\kappa_\star)^3} + o(n_N^{-1}).$$

An application of Chebyshev's inequality gives

$$T_2 = \frac{\kappa_\star}{1-\kappa_\star} + o_{\mathbb{P}}(1).$$

Third term: $T_3 = \Delta_N^{-1}\left(\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}}\right)^2$**.** We now show that $T_3 = o_{\mathbb{P}}(1)$. From Lemma 3, we have $\Delta_N^{-1} = o_{\mathbb{P}}(1)$, so it suffices to show that $T_3' = \tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\widetilde{\mathbf{z}} = \mathcal{O}_{\mathbb{P}}(1)$. Note that $T_3' \mid \{\boldsymbol{x}_i\}_{i\in S_N} \sim \mathcal{N}(0, \|\widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\|^2)$.

We therefore study the quantity $\|\widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\|$. First, observe that

$$\tilde{\mathbf{t}}_{x,S} = \sum_{i\in S_N} \widetilde{\boldsymbol{X}_i} \sim \mathcal{N}(0, n_N \mathbf{I}_p).$$

Hence, $n_N^{-1}\|\tilde{\mathbf{t}}_{x,S}\|^2 \sim \chi^2_p$, which implies, by Chebyshev's inequality, that $\|\tilde{\mathbf{t}}_{x,S}\| = \mathcal{O}_{\mathbb{P}}(\sqrt{p_N n_N}) = \mathcal{O}_{\mathbb{P}}(n_N)$. Moreover, since $\widetilde{\boldsymbol{A}}_S = \widetilde{\boldsymbol{X}}_S^\top \widetilde{\boldsymbol{X}}_S$, we have

$$\|\widetilde{\boldsymbol{A}}_S^{-1}\| = \lambda_{\min}^{-1}(\widetilde{\boldsymbol{A}}_S) = s_{\min}^{-2}(\widetilde{\boldsymbol{X}}_S).$$

Under (C1), Exercise 7.13 of Vershynin (2026) gives that, for any $t \geq 0$,

$$\mathbb{P}\left(s_{\min}(\widetilde{\boldsymbol{X}}_S) \geq \sqrt{n_N} - \sqrt{p_N} - t\right) \geq 1 - 2\exp(-t^2).$$

Let $\epsilon > 0$ be small enough and take $t = \epsilon\sqrt{n_N}$. Then there exists a constant $c = c(\kappa_N, \epsilon) > 0$ such that

$$\mathbb{P}\left(s_{\min}(\widetilde{\boldsymbol{X}}_S) \geq c\sqrt{n_N}\right) \geq 1 - 2\exp(-\epsilon^2 n_N).$$

Consequently,

$$\mathbb{P}\left(\|\widetilde{\boldsymbol{A}}_S^{-1}\| \leq \frac{1}{c^2 n_N}\right) \geq 1 - 2\exp(-\epsilon^2 n_N) \xrightarrow[N\to\infty]{} 1,$$

so that $\|\widetilde{\boldsymbol{A}}_S^{-1}\| = \mathcal{O}_{\mathbb{P}}(n_N^{-1})$. Combining both bounds, we obtain

$$\|\widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{t}}_{x,S}\| \leq \|\widetilde{\boldsymbol{A}}_S^{-1}\|\,\|\tilde{\mathbf{t}}_{x,S}\| = \mathcal{O}_{\mathbb{P}}(1).$$

Therefore, by Chebyshev's inequality, $T_3' \mid \{\boldsymbol{x}_i\}_{i\in S_N} = \mathcal{O}_{\mathbb{P}}(1)$ from which it can be shown that the unconditional tightness also holds. Hence, $T_3 = o_{\mathbb{P}}(1)$.

Fourth term: $T_4 = -2\Delta_N^{-1}\tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{z}}$**.** Since $T_3' = \tilde{\mathbf{t}}_{x,S}^\top \widetilde{\boldsymbol{A}}_S^{-1}\tilde{\mathbf{z}} = \mathcal{O}_{\mathbb{P}}(1)$, and by Lemma 3, $\Delta_N^{-1} = o_{\mathbb{P}}(1)$, their product is negligible in probability and thus, $T_4 = o_{\mathbb{P}}(1)$.

□

Lemma 5. *Let $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ be a random matrix with independent $\mathcal{N}(0,1)$ entries. Let $E \subset \mathbb{R}^n$ be a fixed deterministic subspace of dimension q, such that $p + q < n$. Then, for any vector $\mathbf{a} \in \mathbb{R}^n$,*

$$\mathbf{a}^\top \boldsymbol{P}\left(E + \mathit{Col}(G)\right)\mathbf{a} \overset{d}{=} ||\boldsymbol{P}(E)\,\mathbf{a}||^2 + ||\boldsymbol{P}(E^\perp)\mathbf{a}||^2 \cdot \mathcal{B}\left(\frac{p}{2}, \frac{n-q-p}{2}\right).$$

Proof. We have $E + \mathsf{Col}(G) = E \oplus \mathsf{Col}\left(\boldsymbol{P}(E^\perp)\boldsymbol{G}\right)$ in orthogonal direct sum, so that

$$\boldsymbol{P}\left(E + \mathsf{Col}(G)\right) = \boldsymbol{P}(E) + \boldsymbol{P}\left(\boldsymbol{P}(E^\perp)\boldsymbol{G}\right).$$

Moreover, decomposing $\mathbf{a} = \boldsymbol{P}(E)\mathbf{a} + \boldsymbol{P}(E^\perp)\mathbf{a}$ and noting that $\boldsymbol{P}(E^\perp)\mathbf{a} \in E^\perp$, we get

$$\mathbf{a}^\top \boldsymbol{P}\left(E + \mathsf{Col}(G)\right)\mathbf{a} = ||\boldsymbol{P}(E)\,\mathbf{a}||^2 + \left(\boldsymbol{P}(E^\perp)\mathbf{a}\right)^\top \boldsymbol{P}\left(\boldsymbol{P}(E^\perp)\boldsymbol{G}\right)\left(\boldsymbol{P}(E^\perp)\mathbf{a}\right).$$

We now study the second term. Let $d = n - q$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$ be a matrix with columns being an

orthonormal basis of $E^{\perp}$. Then, $\boldsymbol{P}(E^{\perp}) = \boldsymbol{Q}\boldsymbol{Q}^{\top}$, then $\boldsymbol{P}(E^{\perp})\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{Q}^{\top}\boldsymbol{G} := \boldsymbol{Q}\boldsymbol{H}$ where $\boldsymbol{H} = \boldsymbol{Q}^{\top}\boldsymbol{G}$ is standard Gaussian matrix since for each column $\mathbf{g}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, $\boldsymbol{Q}^{\top}\mathbf{g}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{\top}\boldsymbol{Q}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Note also that

$$\boldsymbol{P}\left(\boldsymbol{P}(E^{\perp})\boldsymbol{G}\right) = \boldsymbol{P}\left(\boldsymbol{Q}\boldsymbol{H}\right) = \boldsymbol{Q}\boldsymbol{P}\left(\boldsymbol{H}\right)\boldsymbol{Q}^{\top},$$

since $\boldsymbol{Q}$ is orthonormal. Therefore, we can write

$$\left(\boldsymbol{P}(E^{\perp})\mathbf{a}\right)^{\top}\boldsymbol{P}\left(\boldsymbol{P}(E^{\perp})\boldsymbol{G}\right)\left(\boldsymbol{P}(E^{\perp})\mathbf{a}\right) = \mathbf{b}^{\top}\boldsymbol{P}(\boldsymbol{H})\mathbf{b},$$

with $\mathbf{b} := \boldsymbol{Q}^{\top}\mathbf{a}$ a deterministic vector. Since $\boldsymbol{H}$ is standard Gaussian, its column space is rotationally invariant in $\mathbb{R}^d$, which implies that the distribution of the quadratic form $\mathbf{b}^{\top}\boldsymbol{P}(\boldsymbol{H})\mathbf{b}$ depends on $\mathbf{b}$ only through its norm (see, e.g., the proof of Theorem 3.3.9. of Vershynin, 2026). It follows that, choosing $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ independent of $\boldsymbol{H}$, then $\mathbf{g}/||\mathbf{g}||$ is a unit vector and thus,

$$\mathbf{b}^{\top}\boldsymbol{P}(\boldsymbol{H})\mathbf{b} \overset{d}{=} \|\mathbf{b}\|^2 \cdot \frac{\mathbf{g}^{\top}\boldsymbol{P}(\boldsymbol{H})\mathbf{g}}{\mathbf{g}^{\top}\mathbf{g}}.$$

Note that $p < d$ and $\boldsymbol{H} \in \mathbb{R}^{d\times p}$ is standard Gaussian, so it has rank p, and by the spectral theorem there exists an orthogonal matrix $\boldsymbol{R}$ such that

$$\boldsymbol{P}(\boldsymbol{H}) = \boldsymbol{R}\begin{pmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\boldsymbol{R}^{\top}.$$

Using rotational invariance again and the fact that $\mathbf{g} \perp\!\!\!\perp \boldsymbol{H}$, we have $\boldsymbol{R}^{\top}\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, so that,

$$\frac{\mathbf{g}^{\top}\boldsymbol{P}(\boldsymbol{H})\mathbf{g}}{\mathbf{g}^{\top}\mathbf{g}} = \frac{\sum_{j=1}^{p} z_j^2}{\sum_{j=1}^{p} z_j^2 + \sum_{j=p+1}^{d} z_j^2} \sim \frac{\chi_p^2}{\chi_p^2 + \chi_{d-p}^2} = \mathcal{B}\left(\frac{p}{2}, \frac{d-p}{2}\right),$$

where $z_1, ..., z_p$ are independent $\mathcal{N}(0,1)$. Therefore, noting that $d - p = n - q - p$, the result follows.

□

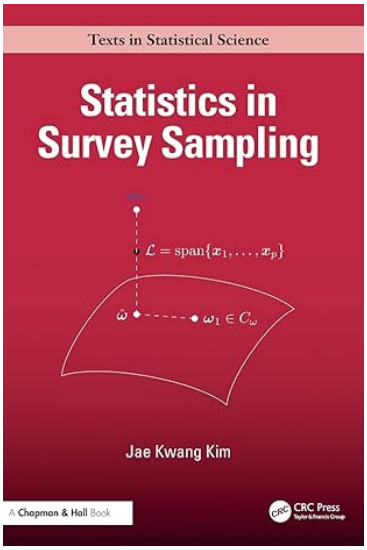
Book Review

Statistics in Survey Sampling

Changbao Wu

University of Waterloo, Canada, cbwu@uwaterloo.ca

Kim, J. K. (2025). *Statistics in Survey Sampling* (1st ed.). Chapman and Hall/CRC, ISBN 9781032997766, DOI: 10.1201/9781003606147, 284 pages.



The title of a book often leads readers to infer its content. This book conveys a clear message through its title, *Statistics in Survey Sampling*, indicating that the focus is on statistical theory and methods used in survey sampling. The author, Dr. Jae Kwang Kim, is a Professor in the Department of Statistics at Iowa State University, a powerhouse in survey sampling research and education with a tradition that spans many decades. Professor Kim is a leading researcher in the field of survey sampling and an innovative theoretical statistician. The content of this book reflects the materials he has used for teaching and some of the research topics he has pursued over the past 25 years. The book is suitable as a textbook or reference for an advanced graduate level course in survey sampling and as a reference for researchers working in the field or on related topics.

The book can be roughly divided into three parts. The first part consists of Chapters 1–7 and covers basic topics in survey design and sampling methods. These include simple and systematic sampling designs described in Chapter 3, stratified sampling in Chapter 4, sampling with unequal probabilities in Chapter 5, single-stage and two-stage cluster sampling in Chapters 6 and 7. The most unique feature of the presentation of this part is the early introduction of the Horvitz-Thompson estimator in Chapter 2.

Chapter 2 starts with the mathematical definition of a probability sample and the probability sampling design through the so-called sampling distribution $P(\cdot)$, which is the probability measure for the sampling design. Design-based expectation, variance, and mean squared error of an estimator, along with first-order and second-order inclusion probabilities, are defined using $P(\cdot)$. Theoretical properties of the Horvitz-Thompson estimator for the population total are then presented for a general sampling design characterized by the first- and second-order inclusion probabilities. Estimation of more general parameters, including those defined through census estimating equations, is briefly discussed. This provides a more rigorous approach to the introduction of survey sampling and is aimed at an advanced level for graduate students in a statistics program.

Chapters 3–7 contain material that goes beyond the basic presentations one often sees in traditional textbooks and may be of interest to students with a strong mathematical and statistical background. Examples include the random sorting method and the reservoir sampling method for implementing simple random sampling (SRS), stratum boundary determination, entropy for sampling designs and

maximum entropy sampling, and balanced sampling.

The second part of the book consists of Chapters 8, 9 and 10 and covers topics on point and variance estimation for the population total and mean. Chapter 8 presents standard material on ratio and regression estimators in the presence of population auxiliary information. Chapter 9 focuses mainly on calibration methods for estimation with survey data and describes several calibration techniques with theoretical details, including the conventional calibration methods, the so-called model-assisted calibration, generalized entropy calibration, and soft calibration. Chapter 10 discusses methods and techniques for variance estimation, including linearization methods and replication methods such as random grouping, the jackknife, and the bootstrap.

The third part of the book covers more specialized topics in Chapters 11–16. Chapter 11 discusses several aspects of two-phase sampling and how it is used to provide auxiliary information for estimation. Chapters 12 and 13 deal with missing survey data, focusing on how to adjust for unit nonresponse and how to impute item nonresponse. Chapter 14 addresses inference for analytic model parameters using survey data, including Bayesian inference. Chapter 15 deals with so-called predictive inference under multi-level models and over time. Chapter 16 includes material on current research topics in the analysis of non-probability survey samples.

Dr. Jae Kwang Kim has worked on many of the topics presented in the second and third parts of the book. He has the insight and technical expertise to introduce the research problems, describe the methodological approaches, and develop theoretical results. These materials are a rich resource for graduate students in statistics and young researchers who are interested in pursuing research on related topics. Each chapter of the book is also followed by a set of practice problems to facilitate teaching and learning, making it easy to use as a textbook.

The book *Statistics in Survey Sampling* contains many references of historical importance to various research topics, some of which remain active today. It is a valuable addition to the existing textbooks and research monographs used for teaching and research in survey sampling, including Cochran (1977), Fuller (2009), Lohr (2022), and Wu and Thompson (2020).

References

Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed., Wiley, New York.

Fuller, W. A. (2009) *Sampling Statistics*. Wiley, Hoboken, NJ.

Lohr, S. L. (2022) *Sampling: Design and Analysis*, 3rd ed., Chapman & Hall/CRC, Boca Raton.

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

Argentina

Reporting: **Verónica Beritich**

INDEC chaired the 28th Specialized Meeting on Statistics of Mercosur

The National Institute of Statistics and Censuses (INDEC) chaired the 28th Specialized Meeting on Statistics of Mercosur (REES), held virtually on May 27, 2025, under Argentina's pro tempore presidency. The meeting aimed to review progress on the 2025–2026 work program and the activities undertaken by the REES commissions and working group.

In addition to INDEC specialists, the meeting was attended by representatives from the Brazilian Institute of Geography and Statistics (IBGE), the National Institute of Statistics (INE) of Paraguay, the National Institute of Statistics (INE) of Uruguay, and the National Institute of Statistics (INE) of Bolivia.

During the session, presentations were delivered by representatives of the Commission on the Inventory of Statistical Operations, the Commission on Income and Employment Statistics, the Permanent Commission on National Accounts, the Commission on the Use of Administrative Records, the Permanent Commission on Communication and Dissemination of Statistical Information, and the Working Group on the Mercosur Community Statistical Plan. They presented progress and proposals for the period.

In presenting INDEC's new statistical operations, it was noted that work is underway on the design of a new Urban Master Sample of Dwellings of the Argentine Republic (MMUVRA), based on data from the 2022 Census, and that population projections are in the final stage of preparation. Among other topics, new technical reports released this year were highlighted, such as the Business Tendency Survey for supermarkets and wholesale self-service stores and the Informality Indicators (Permanent Household Survey). It was also announced that INDEC will incorporate additional information into its set of indicators, including national accounts, environmental accounts, and a series of governance statistics. Finally, the ongoing collaboration between INDEC and the Economic Commission for Latin America and the Caribbean (ECLAC) was underscored, particularly to implement improvements in the development of the geo-statistical portal and the Redatam system for accessing census data.

Additionally, Brazil proposed that REES reinstate the working commission on classifications and nomenclatures, which had concluded its functions in 2020; Paraguay shared its experience with experimental statistics using new data sources such as digital platforms, mobile mobility data, social media, and the internet; and Uruguay noted that it is preparing new population projections through 2079 and will update the statistical register of the resident population to produce a biennial population count.

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: **Joseph Chien**

Streamlining microdata output protection in the ABS DataLab

National Statistical Offices (NSOs) operate secure research environments - commonly referred to as research data centres - to enable approved researchers to access sensitive microdata while safeguarding confidentiality. The Australian Bureau of Statistics (ABS) DataLab is one such environment, providing access to a wide range of individual, household, and business datasets. Since 2020, DataLab usage has grown substantially, with annual sessions increasing from approximately 13,000 to a peak exceeding 38,000 in 2023, and remaining well above pre-2020 levels thereafter.

This rapid growth has intensified a key operational challenge: output vetting. All analytical outputs must be assessed prior to release to ensure they do not disclose information about individual data providers. At present, this process is largely manual, making it resource-intensive, difficult to scale, and susceptible to human error. Increasing output volumes also heighten the risk of differencing, whereby multiple outputs that individually satisfy disclosure rules may collectively reveal sensitive information.

To address these challenges, the ABS has developed a suite of prototype tools—Fortified, Assured, Streamlined, Trusted (FAST). These tools enable approved researchers to apply disclosure protection directly within the DataLab prior to submission for clearance. Implemented in R and R Shiny, FAST provides both programmatic and graphical interfaces, supporting users with varying levels of technical expertise.

The methodology underpinning FAST extends the existing ABS perturbation framework, which comprises two components. The first is a cellkey method that assigns a unique random key to each record and applies consistent noise based on aggregated keys within each output cell. This ensures that outputs derived from the same underlying records receive coherent perturbation, thereby mitigating differencing risks across outputs and projects. The second component is an entropy maximisation approach that increases uncertainty in the applied noise, reducing the risk of reverse engineering while preserving analytical utility.

FAST introduces several enhancements to this framework. To better utilise the noise distribution in the cellkey method, it implements a memory-efficient quantile sampling approach, improving representation of tail behaviour—an important requirement for differential privacy guarantees. For entropy maximisation, FAST develops a novel sequential descent optimisation algorithm for constructing noise distributions. Unlike earlier approaches, this method supports both symmetric and asymmetric distributions, which is critical for outputs bounded at zero, where symmetric noise may produce invalid negative values. The algorithm also incorporates (ε,δ)-differential privacy parameters, providing a principled and flexible mechanism for balancing disclosure risk and data utility.

FAST is complemented by FASTmanager, a system designed for ABS output vetting staff. FASTmanager supports the generation and management of record keys and noise distributions, and provides automated validation that outputs have been produced using approved methods prior to release.

The FAST prototype has undergone initial usability testing within the DataLab, with broader rollout planned. Early feedback indicates strong user acceptance, and the approach is expected to significantly reduce manual vetting effort while improving scalability, consistency, and robustness in output protection.

Brazil

Reporting: **Andrea Diniz da Silva and Clara Oliviera Silveira**

Artificial Intelligence for Public Policies: the IAPP Initiative in Brazil

Recent advances in artificial intelligence (AI), combined with the growing availability of data, have significantly expanded the potential for improving public policies, particularly in terms of producing more accurate diagnostics and strengthening evidence-based decision-making. In this context, the Artificial Intelligence for Public Policies (IAPP) project was created, an effort focused on developing innovative solutions such as conversational AI systems applied to the field of public policies, with a focus on enhancing state action.

The IAPP was launched in 2023 as a collaborative initiative involving IBGE, the State University of Campinas, and the Federal University of Goiás, also bringing together additional academic and institutional partners. The project is structured around an integrated modular architecture, composed of different components that combine technological development, knowledge curation, documentation of practices, and capacity-building in the public sector.

From a strategic perspective, the IAPP is guided by a set of interdependent objectives aimed at developing AI solutions aligned with the public interest, strengthening the management capacities of federative entities — particularly at the municipal level — and expanding access to qualified technical and academic knowledge, anchored in the Centre for Artificial Intelligence in Public Policies (CIAP) as a governance, institutional coordination, and network integration hub. These objectives also include the development of a curated national repository, the systematization of public programs and best practices, and the articulation of an institutional network dedicated to research, innovation, and knowledge dissemination in the field of public policies. Together, these fronts constitute an integrated approach that combines technological innovation, capacity-building, and institutional cooperation as pillars for improving government action.

More than a technological initiative, the IAPP represents a structured effort of institutional cooperation aimed at improving public policies in Brazil. By articulating data production, technological development, training, and knowledge exchange, the project contributes to strengthening state capacity and to the design of more effective, inclusive, and public interest-oriented policies. In this sense, initiatives such as the IAPP reaffirm the importance of promoting more qualified government action, grounded in evidence and supported by the responsible use of artificial intelligence.

Canada

Reporting: **Peter Wright**

Improving the Canadian Labour Force Survey

The Labour Force Survey (LFS) is the official monthly source for labour statistics in Canada. Its main objective is to provide labour information on the employed, unemployed and those not in the labour force. Recent improvements at Statistics Canada have led to greater efficiency in the collection of data and the quality of estimates, and a series of improvements to respondent communications have led to improve efficiency in the collection process.

One major change involved the sampling design. Throughout most of Canada, the LFS employs a six-month rotating panel design with a stratified two-stage sampling methodology. Given that the first stage has a small sampling fraction and resembles with-replacement sampling, variance estimation since 2015 has involved the bootstrap variance estimator proposed by Rao, Wu and Yue. In 2025,

the two-stage sample design changed from a method proposed by Rao, Hartley and Cochran (RHC) to a randomized probability proportional-to-size (RPPSS) approach. In studying the impact of each of the two approaches, RPPSS yielded lower estimates of variance in less-populated geographic areas that were not calibrated to known population control totals. In larger geographic areas where the weights are calibrated, RPPSS yields lower estimated variances for certain key estimates such as total employment.

Another major change was to introduce a pre-collection contact period. This contact period takes place during the month preceding the first of six months of the rotating panel. That way, during the first month of collection much less effort needs to be spent on contacting the dwelling and more effort is spent on collecting data. Given that the monthly collection period lasts only 10 days, the efficiency of collection during the first month is considerably improved.

Other recent improvements include an imputation module for visible minorities and new questions that improve the measure of self-employed earnings. To fulfil the need for timely indicators of the dynamics of the Canadian labour market, particularly during and after the pandemic, the Labour Market Indicators survey was introduced in 2022 as a supplementary survey to the LFS.

Finally, Statistics Canada is regularly testing improvements designed to improve the response rate as well as to increase response by means of the LFS online questionnaire compared to the more traditional collection modes of computer-assisted telephone or personal interviews. For example, some tests involve adjustments to the content of contact letters, the content and timing of reminders sent in SMS texts and email messages, and increasing the collection period to 11 days by starting one day earlier. Initial results indicate that some of these initiatives result in responses earlier during collection period, thereby reducing the need for interviewer follow-up by telephone or in person.

To find out more information about the Canadian Labour Force Survey, please visit https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3701.

References:

Rao, J. N. K., Hartley, H. O., & Cochran, W. G. (1962). On a Simple Procedure of Unequal Probability Sampling Without Replacement. *Journal of the Royal Statistical Society: Series B (Methodological), 24*(2), 482–491.

Rao, J. N. K., C. F. J. Wu, and K. Yue (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18 (2), 209-217.

Croatia

Reporting: **Ksenija Dumičić**

Croatia's KnowledgePanel: A Probability-Based Survey Panel

Croatia recently introduced a probability-based online survey infrastructure through the Ipsos KnowledgePanel, marking an important methodological improvement for academic and applied research. Unlike conventional opt-in web panels, where respondents volunteer themselves, the Croatian panel is based on invitation-only recruitment using probability sampling. According to Ivan Burić (2025), members were recruited through a combination of Random Digit Dialling (RDD) and random selection from local telephone-number databases, with the panel initially reaching around 8,000 members. The likely design is two-stage. First, the panel is constructed through random recruitment from telephone frames. Second, for each survey, a stratified subsample of eligible members is selected according to key characteristics such as sex, age, education, region, and settlement size. Final survey weights are then calibrated to official population distributions (Ipsos, 2025). This makes the panel suitable for estimating population parameters, monitoring trends, and

conducting longitudinal studies. For Croatia, this approach is especially relevant because digital exclusion still affects some older, rural, and lower-income groups. By allowing telephone or face-to-face participation for non-internet users, the panel improves coverage. In addition, Croatia's mobile-dominated telecommunications market requires careful dual-frame sampling and weighting adjustments (HAKOM, 2026). Overall, KnowledgePanel combines the speed of online research with stronger statistical validity than volunteer web panels.

References

Burić, I. (2025) *Probabilistički online panel u Hrvatskoj*. Sociologija.hr. Available at: https://sociologija.hr/probabilisticki-online-panel-u-hrvatskoj/

HAKOM (2026) *Quarterly comparative data on electronic communications market, Q4 2025*. Croatian Regulatory Authority for Network Industries. Available at: https://www.hakom.hr/

Ipsos (2025) *KnowledgePanel Croatia: Probability-based online research methodology*. Ipsos Croatia. Available at: https://www.ipsos.com/hr-hr/ipsos-knowledgepanel-r

The Netherlands

Reporting: **Ger Snijkers**

Cognitive Friendly Statistics

Statistics Netherlands (CBS) is responsible for compiling and publishing reliable and high-quality statistics of the Dutch society. In every step of the statistical production process considerable effort goes into ensuring data quality. However, CBS has recognized the need to extend this to the dissemination stage: even though compiled with quality in mind, little systematic attention has been paid to whether the final published output (like visualizations, articles and other outlets) is actually understood correctly by intended users. Effective communication of official statistics and identifying its quality aspects are considered important. To address this, the research program 'Statistical Information Communication' has been launched, aiming for evidence-based communication design guidelines extracted from empirical user-centric studies. The central research question is how can statistical information be effectively visualized and communicated so that it is understandable and usable by the intended audience? In this paper the first studies conducted in the context of this program are presented. A series of exploratory user evaluation studies have been carried out, examining how different types of users perceive and interpret a range of statistical communication products, including infographics, bar charts, dashboards and written articles.

The studies draw on methods from multiple disciplines — cognitive psychology, survey methodology, information visualisation and human-computer interaction — and were conducted in innovative settings such as pop-up tests in public spaces and school visits to the CBS Userlab. Using cognitive interviewing techniques, color palette experiments and mixed-method usability studies, several types of infographics were examined with 42 participants aged 12 to 72 across varying backgrounds and levels of statistical literacy.

The findings consistently showed that there is no one-size-fits-all approach to communicating statistics. User comprehension varied significantly depending on design choices such as color selection, legend placement, icon design, label and title wording, and the use of accompanying photographs. Some general guidelines could be summarized like the necessity of using legends and fully labelled symbols. Decorative elements intended to make infographics more appealing, frequently caused information overload or distracted users from the core statistical message. Users also tended to personalize statistics — searching for themselves within a graph — highlighting the importance of inclusive and representative design.

On the basis of these findings, the authors introduce the concept of "cognitive friendly statistics," defined as statistical information communicated in a way that aligns with natural human cognitive processes, thereby reducing the chance of bias by misinterpretations and misuse. The parallel can be drawn with pre-testing and evaluation of survey questions aimed at reducing bias in survey data caused by respondent misinterpretation of these questions and measurement errors. As such, user studies are a fruitful way to develop design guidelines for statistical output products. The authors call on other National Statistical Institutes to join efforts in building shared best practices for and insights on cognitive friendly statistical communication.

Reference

Meertens, V., Snijkers, G., Stoel, R.D., Tennekes, M., Krol, N., Bongers, N., Jonge, de E., & Puts, M. (2026). Past Insights and Future Innovations in Statistical Communication Design: Towards Cognitive Friendly Statistics. *Statistical Journal of the IAOS* 42(1): 170-182. *(*https://doi.org/10.1177/18747655251414390*)*

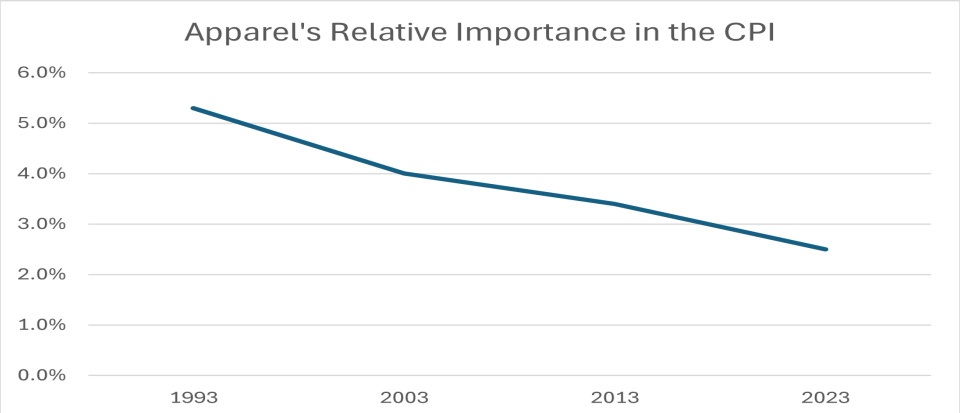
United States

Reporting: **Jeffrey Hill and Sarah Van Giezen (BLS)** and **Andreea L. Erciulescu (Westat)**

U.S. Consumer Price Index improves to reflect price change for secondhand apparel

The U.S. Bureau of Labor Statistics (BLS) now reflects price change for secondhand apparel in the U.S. Consumer Price Index (CPI). The CPI is the most widely used measure of inflation in the United States, measuring the average monthly change in prices paid by urban consumers. It historically limited its observed market basket of goods and services to new or unused items in good condition in order to maintain a market basket with consistent item quality. Yet, for many years, BLS has included used cars and trucks in the CPI because consistent quality is easily determined for these items. Now, given consumer preferences shifting towards secondhand goods and the expanded availability and improved quality control of secondhand apparel, BLS determined how to incorporate secondhand apparel into the CPI.

American households spent roughly the same amount of money on apparel each year between 1990 and 2022, except in 2020 when they reduced their spending on apparel by 23.8 percent during the early stages of the COVID-19 pandemic, according to the BLS Consumer Expenditure (CE) Surveys. Although the amount Americans spent on clothing was consistent, the amount compared with other expenditure categories decreased due to increased spending in other categories. Apparel's relative importance in the CPI declined from 5.3 percent in 1993 to 4.0 percent in 2003, 3.4 percent in 2013, and 2.5 percent in 2023.



According to various media reports, American consumers have been buying more secondhand clothing rather than shopping at major apparel and department stores, with the shift toward secondhand apparel being driven by consumer concerns about apparel costs and environmental sustainability, particularly within younger generations. The percentage of millennials and members of Generation X and Generation Z who are shopping secondhand apparel has increased steadily between 2016 and 2020, with 42 percent of the two youngest generations shopping secondhand apparel at some point in 2020, according to ThredUP's 2021 Resale Report.

The CE captures consumer expenditures at secondhand stores. As a result of being reported in the CE, secondhand stores are included in the pool of stores eligible for selection into the CPI sample. These stores, both brick-and-mortar and online, today can provide more detailed descriptions of item quality. It is standard practice for secondhand item listings to include a glossary of terms for sellers to include in their listings, from "brand new" to "very used." Many popular secondhand apparel stores have measures in place to control item quality. As a result, BLS follows data collectors' observations and stores' descriptions to maintain consistent quality over time and will continue to exclude items that are damaged.

Note that the CPI apparel indexes will not have a separate breakout for secondhand apparel in the way that New vehicles is separate from Used cars and trucks or that Apparel is further divided into Men's and boys' apparel and Women's and girls' apparel. Rather, secondhand goods are reflected in the probability of selection of sample units. BLS then collects prices each month, and the Apparel CPIs now reflect secondhand price change.

This effort aligns with the BLS strategic goal of improving measurement of the U.S. economy. BLS is also researching ways to include other secondhand goods in the CPI.

AI Day for Federal Statistics

US federal agencies continue to develop best practices for leveraging AI to advance their mission. The second AI Day for Federal Statistics workshop was organized by the Committee on National Statistics (CNSTAT), the Federal Committee on Statistical Methodology (FCSM), and the National Institute of Statistical Sciences, on April 30, 2026, in Washington, DC. Bringing together leaders from federal agencies, research organizations, and the private sector, the workshop provided a discussion and training platform for federal agencies on how AI is reshaping the development of policies and strategies, and the official statistics production workflows.

Uruguay

Reporting: **Diego Aboal and Federico Segui**

How Uruguay reached 60% online Census participation in a Latin American context

Uruguay's 2023 Population Census marked a major step in the modernization of census-taking in Latin America. For the first time, Uruguay implemented a large-scale Computer-Assisted Web Interviewing (CAWI) strategy as a central component of the census operation, rather than as a marginal or recovery mode. The result was remarkable. Approximately 60% of households completed the census through the web questionnaire, a level unmatched in Latin America and comparable to the first successful large-scale digital census experiences in several developed countries.

Several operational factors explain this outcome. A key innovation was the use of the national electricity utility's customer and meter database as census infrastructure. Since electricity coverage in Uruguay reaches virtually all households, the meter number provided a near-universal household identifier. It also served as an authentication mechanism and enabled the automatic geolocation of web responses. Households accessed the census website using their electricity customer information, received a verification code, completed the questionnaire online, and then obtained a completion code to be shown later to the enumerator.

This design solved three common problems in web censuses: how to authenticate households, how to avoid duplicate responses, and how to link online responses to geographic areas. It also allowed the field operation to focus on non-responding households, improving the efficiency of the second phase of data collection.

The web platform was designed with a mobile-first approach. This was essential in a context where smartphone access is broader than desktop computer access, especially among lower-income groups. The questionnaire was optimized for mobile phones, with a simple login process, clear navigation, contextual help, real-time validation, and the possibility of saving and resuming responses. The results confirmed the relevance of this design. Most online responses (75%) were completed using smartphones.

Communication was another decisive factor. The digital census was supported by a phased, multi-channel campaign that treated communication as part of the census operation itself. Peaks in daily responses coincided with specific communication milestones, showing that public awareness and behavioral incentives (the lottery for one year of free electricity consumption and smartphones offered to households that completed the census online) were central to participation. The campaign positioned online completion as the expected and convenient mode, not merely as an alternative.

The web phase also relied on strong preparatory work. A pre-census cartographic operation validated the administrative address frame, confirming a very high level of concordance between field-verified addresses and the electricity utility database. This ensured that the online phase rested on a reliable territorial infrastructure.

Uruguay's experience shows that high digital participation is feasible in Latin America when technological design, administrative data, communication, and field operations are integrated from the beginning. It also suggests that countries do not necessarily need a central population register to implement a successful web census. A high-quality household-level administrative source, such as a utility database, can provide a practical foundation for authentication, geocoding, and coverage monitoring.

Uruguay's 2023 web census represents an important step toward future census modernization. It provides evidence that digital participation can be expanded substantially in developing countries when the census is designed around existing administrative infrastructure, mobile access patterns, and sustained public engagement.

For further information, please contact: diego.aboal@gmail.com or federico.segui@outlook.com

Events on survey statistics and related areas

Joint Statistical Meetings

Date: 1-6 August 2026

Organizers: American Statistical Association and joint associations

Location: Boston MA, USA

Webpage: https://ww2.amstat.org/meetings/jsm/2026/conferenceinfo.cfm

XXIX Brazilian School of Probability (celebrating Professor Maria Eulália Vares)

Date: 3-7 August 2026

Location: Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brazil

Webpage: https://sites.google.com/im.ufrj.br/ebp2026/home

BNU Workshop on survey statistics 2026

Date: 24-28 August 2026

Organizer: Baltic-Nordic-Ukrainian Network on Survey Statistics

Location: House of Science, Riga, Latvia (also online)

Webpage:

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

XII Congress of ALAP

Date: 26-28 August 2026

Organizer: Latin American Population Association (ALAP)

Location: San José, Costa Rica

Webpage: https://www.alapop.org/

uRos2026 – The Use of R in Official Statistics

Date: 18-22 November 2026

Location: INSEE, Paris, France

Webpage: https://r-project.ro/conference2026.html

HILDA Survey Research Conference

Date: 1-2 October 2026

Location: University of Melbourne, Australia

Webpage: https://melbourneinstitute.unimelb.edu.au/conferences/2026-hilda-survey-research-conference

WAPOR 79th Annual Conference

Date: 26-29 October 2026

Organizer: World Association for Public Opinion Research (WAPOR)

Location: Mexico City, Mexico

Webpage: https://wapor.org/events/annual-conference/current-conference/

18th Annual European DDI[1] Users Conference

Date: 30 November – 4 December 2026

Location: Vrije Universiteit, Brussels, Belgium

Webpage: https://events.geant.org/event/2094/overview

[1] The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences.

Journal of Survey Statistics and Methodology

Volume 14, Issue 2, April 2026
https://academic.oup.com/jssam/issue

Survey Methodology

No Introduction Necessary: Face-saving Answer Options Outperform Preambles in Reducing Social Desirability Bias
Emma Zaal, John Hoeks and Yfke Ongena

Asking for Feedback: Innovating Final Comment Questions in Self-Administered Web Surveys
Joshua Claassen, Jan Karem Höhne and Jessica Kuhlmann

Analyzing List-Style Open-Ended Questions: Combining Texts from Individual Answer Boxes Improves Classification with Language Models
Ruben L. Bach, Matthias Schonlau and Katharina Meitinger

Randomized, Controlled Trial Comparing Household Response and Completion Rates Using 1 and 2 US Dollar Unconditional Incentives in a Mail Push-to-Web Survey
Jon Agley, Tiffany Neal, Mariah Benham, Lilian Golzarri-Arroyo and Naomi Swiezy

Assessing the Efficacy of In-the-Moment Surveys Triggered by Geolocation Data: A Comparative Study of Beach Visitors
Carlos Ochoa

He Said, She Said: Gender-of-Interviewer Effects and the Role of the Interviewers' Gender Attitudes in the Hungarian ESS Round 11
Ádám Stefkovics and Vera Messing

Meeting Data Collection Goals Quicker: An Experimental Evaluation to Reduce Fieldwork Duration in a Mixed-mode Panel Study
Katherine A McGonagle and Narayan Sastry

Using Conditional Linear Regression Trees to Identify Potential Construct Validity and Measurement Error Differences in Health Outcomes from a U.S. National Survey
Morgan S. Earp, Lauren M. Rossen, Sarah E. Forrest and Trent D. Buskirk

Survey Statistics

A Beta–Beta Prime Model for Rates and their Precision for Small Area Estimation: An Application to Brazilian Food Insecurity Index
Fernando A. S. Moura, Soraia Pereira and Giovani L. Silva

Comparative Effectiveness of Propensity Score Estimation Methods for Inverse Probability of Treatment Weighting Analysis with Complex Survey Data: A Simulation Study
Lihua Li, Chen Yang, Liangyuan Hu, Wei Zhang, Melissa Aldridge, Bian Liu and Madhu Mazumdar

On the Use of Proxy Variables in Statistical Matching under Complex Sampling Designs
Daniela Marella and Vincenzina Vitale

In Other Journals

Unifying Small Area Estimators Based on Area-Level and Unit-Level Models Through Calibration
William Acero, Isabel Molina and Juan Miguel Marín

Applications

Repeated Attempt Models for Nonresponse Bias: Small Sample Performance, Optimization of Contact Attempts, and an Application to Deer Hunters
Matthew J. Clement and Matthew Karam

Volume 14, Issue 1, February 2026
https://academic.oup.com/jssam/issue

Special Issue: Survey Research on Asia, Africa, Latin America, the Caribbean, and Oceania: P. 2

Survey Methodology

Developing a Customized, Enumeration Area-based Sampling Frame Tailored to a Specific Population Subgroup Using Geospatial Methods
Sarchil Qader, Edith Darin, Ahmadou Hamady Dicko, Hisham Galal, Hyunju Park, Rebeca Moreno Jimenez, Andrew Harfoot and Andrew J. Tatem

Respondent-driven Sampling Online (Web RDS) as a Strategy to Access Hard-to-reach but Non-hidden Populations: The Case of Health Professionals Working in Chilean Schools
Katherine Dinamarca-Aravena, Andrés González Santa Cruz, Sonia Morales Miranda, Teresita Rocha Jiménez and Álvaro Castillo-Carniglia

Variations In Self-reported Happiness Across Different Response Scales: Evidence from 34 Chinese Surveys from 2002 To 2021
Shuaiying Cao, Minglei Wang, Ting Yan, Lirui He and Chan Zhang

Controlling Acquiescent Response Style Through Negated Versus Polar Opposite Items: Design Considerations for Balancing Measurement Scales
Fernanda Alvarado-Leiton, Sunghee Lee and Rachel E. Davis

Analysis of Risk and Protective Factor Surveillance for Noncommunicable Diseases Using a MultiMode Data Collection Approach
Laura Cordeiro Rodrigues, Izabella Paula Araújo Veiga, Letícia De Oliveira Cardoso and Rafael Moreira Claro

Effects of Interviewer Language and Dialect Choice on Questions About Political Trust: Examining the Asian Barometer Survey in China, The Philippines, and Indonesia
Mao Li, Victoria Owens and Fred Conrad

Survey Statistics

Robust Extension of the Multivariate Fay-Herriot Model for Data with Outliers
Annop Angkunsit and Jiraphan Suntornchost

Direct-assisted Bayesian Unit-level Modeling for Small Area Estimation of Rare Event Prevalence
Alana McGovern, Katherine Wilson and Jon Wakefield

Toward a Principled Workflow for Prevalence Mapping Using Household Survey Data
Qianyu Dong, Yunhan Wu, Zehang Richard Li and Jon Wakefield

Journal of Official Statistics

Volume 42, Issue 2, June 2026
https://journals.sagepub.com/toc/jofa/42/2

Articles

Comparison of Small Area Procedures Based on Gamma Distributions with Extension to Informative Sampling
Yanghyeon Cho and Emily

Geographical Rezoning as a Combinatorial Game: Case Study of the Australian Statistical Geography Standard
Filip Juricev-Martincev, Helen Thompson and Gentry White

Assessing the Heterogeneity in Mode Effects on Data Quality, Response Distributions, and Future Participation Across Sociodemographic Subgroups in a Mixed-Mode Panel Study
Heather M. Schroeder, Mary Beth Ofstedal and Brady T. West

Measuring Risk of Re-Identification for a Nonprobability Sample Using a General Reference Sample
Natalie Shlomo, Minsun Riddles and Tom Krenzke

Calibrating Nonresponse Bias: A Cautionary Tale
Raimund Wildner, Volker Bosch and Florian Meinfelder

Hierarchy-Aware Heterogeneous Graph Neural Network for Occupation Title Coding
Yi Xie and Wenbin Zhu

Book Review

Robust Small Area Estimation: Methods, Theory, Applications, and Open Problems, by Jiming Jiang and J. Sunil Rao
Andreea L. Erciulescu

Volume 42, Issue 1, March 2026
https://journals.sagepub.com/toc/jofa/42/1

Articles

How Does Noise Protection Affect the Accuracy of Life Expectancy and Other Demographic Indicators?
Fabian Bach

Multilingual Hierarchical Classification of Job Advertisements for Job Vacancy Statistics
Maciej Beręsewicz, Marek Wydmuch, Herman Cherniaiev and Robert Pater

Multiplicative Decompositions of GEKS-Type Indices into the Contribution of Individual Commodities
Jacek Białek

Defining Ad-Hoc Sampling Designs for Small Area Estimation
Piero Demetrio Falorsi, Stefano Falorsi, Vincenzo Nardelli, Paolo Righi

Estimating Precision of Deterministic Linkage
Yue Ma and James Chipperfield

A Decomposition of the Change in Total Poverty Gap
Andrea Marletta and Mauro Mussini

Research Note

Privacy-Enhancing or Privacy-Elusion Technology? A Critical View of (Pseudo)Synthetic Data Based on Deep Learning
Fabio Ricciato

Survey Research Methods

Volume 20, No.1, 2026
https://ojs.ub.uni-konstanz.de/srm/

Surveying Minoritized Citizens: A Quantitative Study of Identification Versus Categorization
Sanne van Oosten

Expanding Survey Response Options: Combining Dictation and/or Voice Recording With Text to Answer Narrative Open-Ended Survey Questions
Melanie Revilla and Mick P. Couper

Is There a Single Best Way how to Word a Typical English Agree-Disagree Scale in the German Language: Results From a Survey Experiment
Lukas Schick, Dorothée Behr, Cornelia Neuert and Clemens Lechner

Addressing Answer Consistency of Residents in Residential Homes: Experiences From a Mixed Methods Study
Marie-Kristin Döbler and Katrin Drasch

Measuring the Effect of Questionnaire Design on Rating Responses With the Class of CUB Models
Stefania Capecchi, Romina Gambacorta, Rosaria Simone and Domenico Piccolo

Responding as Expected? The Effects of Survey Mode on Estimates of Sensitive Attitudes in Self-Completion and Face-To-Face Interviews of the European Social Survey
Blanka Szeitl, Bence Ságvári and Vera Messing

Survey Design and Quality During the COVID-19 Pandemic in Germany: An Assessment With 686 Social Science Surveys
Karolina von Glasenapp, Thomas Skora, Tobias Gummer and Elias Naumann

Respondents' Preferred Survey Topics: Measurement and Prevalence
Tobias Gummer, Saskia Bartholomäus, Bernd Weiß

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
Dr.	Ariane	Neethling	South Africa
Dr.	Lara	Cleveland	United States
Ms.	Nadia	Lkhoulf	Morocco
Dr.	Faisal	Awartani	Israel
Dr.	Aldo	Gardini	Italy
Dr.	Faustina	Frempong-Ainguah	Ghana
Dr.	Jonathan	Mendelson	United States
Mr.	Joseph	Rodhouse	United States
Dr.	Yan	Li	United States

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