Book of Abstracts

Small Area Estimation, Survey and Data Science 2026

SAE 2026 Conference

15–19 June 2026

Faculty of Business and Administration, University of Bucharest

Bucharest, Romania

Institutional partners

https://sae2026.faa.ro/

smallareaestimation2026@gmail.com

Publication note

The abstracts included in this volume are published as submitted by the authors, with minor editorial adjustments where necessary. The views expressed are those of the authors and do not necessarily reflect those of the organisers or affiliated institutions.

Accepted abstracts may be followed by full paper submissions to the special issue of the *Romanian Statistical Review*.

Organising information

Advisory Committee

Name	Affiliation
Monica Pratesi	University of Pisa – Department of Economics and Management, Italy
Jiming Jiang	University of California, Davis, USA
Malay Ghosh	Distinguished Professor, University of Florida, USA
Raymond Chambers	University of Wollongong, Australia
Isabel Molina Peralta	Faculty of Mathematical Sciences, Complutense University of Madrid, Spain
Partha Lahiri	University of Maryland, College Park
Ralf Münnich	Economic and Social Statistics, Trier University, Germany

Program Committee

Co-Chairs

Name	Affiliation
Andreea Erciulescu	Westat Maryland, USA
Domingo Morales González	Universidad Miguel Hernández de Elche, Spain

Members

Name	Affiliation
Emily Berg	Iowa State University, USA
Daniel Bonnery	National Institute of Geographic and Forest Information, France

Jan van den Brakel	Statistics Netherlands, Department of Research and Innovation, Heerlen, the Netherlands; Maastricht University School of Business and Economics, Department of Quantitative Economics, Maastricht, the Netherlands
Jan Pablo Burgard	Trier University, Germany
Snigdhansu Chatterjee	Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA
Sanjay Chaudhuri	Department of Statistics, University of Nebraska-Lincoln
Haoyi Chen	Coordinator of the Inter-Secretariat Working Group on Household Surveys and co-Chair of the Collaborative on Citizen Data
Angelo Cozzubo	NORC; University of Maryland; PUC Peru
Gauri Datta	Department of Statistics, University of Georgia, Athens, GA, USA; CSRM, U.S. Census Bureau, USA
Enrico Fabrizi	Università Cattolica del S. Cuore, Milan, Italy
Carolina Franco	NORC at the University of Chicago, USA
David Haziza	Department of Mathematics and Statistics, University of Ottawa, Canada
Scott Holan	Department of Statistics and Data Science, University of Missouri, USA
David Newhouse	Development Economics Data Group of the World Bank Group
Thuan Nguyen	Oregon Health & Science University, USA
Silvia Pacei	Department of Statistical Sciences, University of Bologna, Italy
Nicola Salvati	University of Pisa, Department of Economics and Management
Marius Stefan	Faculty of Applied Sciences and Fundamental Sciences Applied in Engineering – Research Center, National University of Science and Technology Politehnica, Bucharest
Jiraphan Suntornchost	Chulalongkorn University, Thailand
Marcin Szymkowiak	Poznań University of Economics and Business; Statistics Poland
Matthias Templ	FHNW University of Applied Sciences and Arts Northwestern Switzerland, School of Business, Institute for Competitiveness and Communication
Mahmoud Torabi	College of Community and Global Health, University of Manitoba, Canada

Local Committee

Co-Chairs

Name	Affiliation
Ana-Maria Ciuhu	Institute of National Economy, Romanian Academy and National Institute of Statistics, Romania
Marius Jula	Faculty of Business and Administration, University of Bucharest, Romania

Members

Name	Affiliation
Bogdan Oancea	Faculty of Business and Administration, University of Bucharest, Romania
Valentina Vasile	Institute of National Economy, Romanian Academy, Romania
Andreea Constantinescu	Institute of National Economy, Romanian Academy, Romania

Institutional partners

Foreword

It is a great honour for Romania, for the University of Bucharest, and for the Faculty of Business and Administration to host the *Small Area Estimation, Survey and Data Science 2026 Conference* in Bucharest. Bringing this international scientific event to Romania is a valuable opportunity to strengthen dialogue between researchers, official statisticians, survey methodologists, data scientists, and practitioners working at the intersection of small area estimation, survey statistics, and modern data-driven approaches.

A rich and exciting programme awaits at SAE 2026. This international conference will serve as a bridge among statisticians, survey methodologists, economists, engineers, mathematicians, computer scientists, and others interested in combining information from multiple sources to make reliable inferences at granular levels. In addition to traditional small area estimation topics, the conference will cover emerging topics in survey and official statistics, including non-probability sampling, record linkage, data fusion, machine learning and artificial intelligence, disclosure control, and synthetic data.

We are delighted to have a programme comprising 16 invited sessions, 6 contributed sessions, 3 keynote addresses, and 2 short courses. The scientific programme reflects the richness and diversity of the field, covering methodological advances, practical applications, and emerging challenges in the production of reliable statistics for small domains and populations. The conference also provides a forum for discussing how small area estimation and related methods can support high-quality, policy-relevant statistical evidence in increasingly complex data environments.

The conference is hosted by the Rectorate of University of Bucharest, located at 90 Panduri Road, Sector 5, Bucharest. The venue offers an academic setting with convenient access to transport connections, accommodation options, and cultural landmarks, providing an excellent environment for scientific exchange, collaboration, and networking.

Bucharest, the capital of Romania, offers participants the opportunity to experience a dynamic European city with a rich historical, cultural, and academic tradition. We hope that, alongside the scientific sessions, participants will enjoy discovering the city, its architecture, museums, parks, and vibrant cultural life. Hosting SAE 2026 in Romania is also an opportunity to highlight the role of statistical science in supporting evidence-based decision-making at local, regional, national, and international levels.

We would like to express our sincere gratitude to all participants who submitted strong session proposals and paper abstracts. We also thank all those who volunteered to serve as session chairs, as well as the keynote speakers, invited session organisers, presenters, short-course lecturers, reviewers, committee members, institutional partners, and all participants whose contributions make this conference possible. Special thanks are due to the Programme Committee and the Ad-

visory Committee for their valuable feedback and support on various scientific and organisational matters, and to the Local Organising Committee for its dedication in preparing the conference.

We hope that this Book of Abstracts will serve as a useful guide to the scientific content of SAE 2026 and as a record of the research, ideas, and collaborations presented during the conference.

Enjoy the conference!

The SAE 2026 Organising Committee

SAE conferences

One of the earliest physical gatherings of researchers dedicated to Small Area Estimation was the 1978 *Workshop on Synthetic Estimates for Small Areas*. Sponsored by the U.S. National Institute on Drug Abuse and the National Center for Health Statistics, the workshop produced a highly important monograph at the time. The formalization of SAE into a collaborative global community grew throughout the 1980s and 1990s. Notably, the 1985 *International Symposium on Small Area Statistics* in Ottawa, Canada, helped lay the field's mathematical groundwork. Following this, the 1993 symposium, *Small Area Statistics in Public Health: Design, Analysis, Graphic and Spatial Methods*, became a watershed moment for epidemiology, public health, and spatial statistics. The proceedings from both landmark symposia were published by Wiley.

The 1992 *International Scientific Conference on Small Area Statistics and Survey Designs* in Warsaw, Poland, is widely regarded as the first major international SAE conference in Europe. Jointly organized by the International Association of Survey Statisticians (IASS), Poland's Central Statistical Office, and the Polish Statistical Association with support from Eurostat and the World Bank, its selected papers were published in Statistics in Transition (1993, 1994). Later, the 1999 Riga, Latvia, Conference on Small Area Estimation served as an official IASS Satellite Conference to the 52nd ISI World Statistics Congress.

The 2001 *International Conference on Small Area Estimation and Related Topics* in Potomac, Maryland—supported by leading U.S. federal statistical agencies, private survey organizations, the Washington Statistical Society, and the ASA—catalyzed the modern, rotating international SAE conference series. Leaving behind localized national formats, the series launched as a rotating global forum with SAE 2005 in Jyväskylä, Finland, co-sponsored by the IASS. The conference frequently shifts among Europe, the Americas, and Asia:

- Jyväskylä, Finland, 2005
- Pisa, Italy, 2007
- Elche, Spain, 2009
- Trier, Germany, 2011
- Bangkok, Thailand, 2013
- Poznan, Poland, 2014
- Santiago, Chile, 2015
- Maastricht, The Netherlands, 2016

- Paris, France, 2017
- Shanghai, China, 2018
- Naples, Italy, 2021 (Virtual)
- College Park, Maryland, USA, 2022 (Hybrid)
- Lima, Peru, 2024
- Turin, Italy, 2025
- Bucharest, Romania, 2026

The series has featured unique formats and milestones over the years. In 2009, a separate Small Area Conference with exclusively plenary and poster sessions was held during a Rhine River Cruise in Germany. In 2019, the conference expanded as a satellite event to the World Congress in Kuala Lumpur, Malaysia, forming part of a month-long data integration program at the National University of Singapore's Institute of Mathematical Sciences. Another major milestone occurred in 2022, when the International Association for Official Statistics (IAOS) became a formal co-sponsor.

Several SAE conferences have been designated as World Statistics Congress satellite meetings. Selected papers from some of these conferences are published in special issues of peer-reviewed journals, such as the *Calcutta Statistical Association Bulletin*, *Journal of the Royal Statistical Society Series A*, *Statistics in Transition New Series*, and *Survey Methodology*.

Furthermore, since 2017, the conference has presented SAE Awards during a special ceremony to recognize outstanding contributions to research, application, and education in the field. The past recipients of this award include J.N.K. Rao (2017), Danny Pfeffermann (2018), Malay Ghosh (2019), Partha Lahiri (2020), Wayne Fuller (2021), Robert Fay (2022), Jiming Jiang (2024), and Ray Chambers (2025).

Future editions of the SAE conference will continue to unite the global statistical community, advancing the field through cutting-edge research and international cooperation.

Partha Lahiri

Contents

General conference programme

Day	Overview
Monday, 15 June 2026	Registration; Short Course: Entity resolution; Welcome Reception / Networking.
Tuesday, 16 June 2026	Welcome ceremony; Keynote Lecture 1; IS1–IS5; Contributed Sessions 1–3.
Wednesday, 17 June 2026	IS6–IS11; Keynote Lecture 2; Contributed Sessions 4–6.
Thursday, 18 June 2026	IS12–IS16; Keynote Lecture 3; Contributed Session 7; Closing Ceremony.
Friday, 19 June 2026	Short Course: Bayesian small area estimation, with an emphasis on low- and middle-income countries.

Keynote lectures

Isabel Molina

Affiliation:	Faculty of Mathematical Sciences, Complutense University of Madrid, Spain
Chair:	Partha Lahiri
Keynote title:	*Conciliation: the key to success in small area estimation*
Room:	SAE 1

Abstract

Conciliation of design-based and model-based approaches to statistical inference is key to success in Small Area Estimation, as it provides a means to "borrow strength" across areas by combining data from different areas and integrating different data sources. Procedures that incorporate the sampling design into model-based estimators are reviewed, both for area means and for more general indicators, including poverty and/or inequality measures. Approaches that reconcile the two frameworks to estimate the mean squared error of model-based small-area estimators are also outlined. Finally, other research directions that combine both approaches are discussed.

Gauri Datta

Affiliation:	Department of Statistics, University of Georgia, Athens, GA; Center for Statistical Research and Methodology, U.S. Census Bureau, Suitland, MD
Chair:	Malay Ghosh
Keynote title:	*A Bayesian framework for multi-goals small area inference: estimation, ranking and benchmarking*
Room:	SAE 1

Abstract

In small area inference, estimation of subpopulation means is often the primary goal of national statistics offices. For example, information on poverty, income, and access to primary healthcare at disaggregated levels is required in welfare programs. However, inference on the overall ranking of entities, such as small areas, hospitals, school districts, etc., is equally critical in planning, policy making, and advocacy related to such programs.

Ranking draws attention to unusually high or low performing subpopulations and provides investigators and policy makers with useful tools to establish priorities. In human development, where resources are limited, authorities need to identify the most underprivileged or impoverished subpopulations to deliver relief.

Estimates of ranks constructed exclusively from point estimates of parameters lack uncertainty quantification and may lead to imbalances and inequities. This is especially true in small area statistics, where there may be only a limited amount of directly observed data from each area, and the point estimates are subject to large estimation error. To address this deficiency, Klein et al. developed frequentist confidence sets for the rank vector. As an alternative to this and another recent frequentist solution, we propose a novel Bayesian approach. Our solutions are built on strengths of the Bayesian paradigm: they identify the likely ranks for entities along with a probability distribution on the identified plausible ranks.

The proposed solutions significantly outperform the state-of-the-art frequentist alternatives and, unlike their frequentist counterparts, can borrow from covariates as well as from other small areas. We evaluate our proposed Bayesian algorithms in terms of accuracy and stability using a simulation study and two applications of interest to the U.S. Census Bureau.

In a related problem, either by policy necessity or to ensure against possible model failure, model-based estimates of small area means are often required to be modified so that certain well-defined aggregates of these modified values agree with corresponding more reliable and direct benchmark values. Many existing benchmarked solutions are obtained by modifying the regular Bayesian estimates of the small area means in order to comply with the aggregation requirement, but variances of the benchmarked estimates are still computed under the regular model. Sugasawa et al. argued that incorporating the benchmark constraints perturbs the regular posterior distribution of the small area means to a new posterior distribution that automatically satisfies the constraints, produces benchmarked estimates, and more accurately measures uncertainty. Using samples from this modified posterior distribution, we carry out point and interval estimation for the means and ranks of the small areas.

Cristina-Rodica Boboc

Affiliation:	Bucharest University of Economic Studies and Institute of National Economy, Romanian Academy, Romania
Chair:	Valentina Vasile
Keynote title:	*Measuring Skill Mismatch In The Romanian Labour Market. Are Small Area Estimation Methods A Solution?*
Room:	SAE 1

Abstract

Romania's labour market has a problem that unemployment figures do not capture: the right number of workers, but in the wrong jobs, with the wrong qualifications. Half of workers in elementary occupations are overqualified. Over half of agricultural workers lack formal certification. And these imbalances have persisted for a decade. Yet when a local policymaker asks what is happening in their county, the data simply is not there. This keynote explores what we know about skill mismatch in Romania, where our measurement tools fall short, and whether small area estimation can finally give us the local picture we need.

Short courses

Entity resolution

Lecturer:	Ted Enamorado
Affiliation:	Department of Political Science, Washington University in St. Louis
Date:	Monday, 15 June 2026
Format:	Slot A: 14:00–15:30; Slot B: 16:00–17:30
Room:	TBA

Course description

This course focuses on the common task of identifying and merging records from diverse data sources that correspond to the same entities. Whether you are working with large-scale databases, healthcare records, or customer datasets, the ability to accurately link records is crucial for data integration, analysis, and decision-making. Our goal is to provide you with a comprehensive introduction to both the theoretical foundations and practical applications of record linkage, equipping you with the skills to address real-world data challenges effectively.

Throughout this course, we will explore various methodologies and algorithms that underpin entity resolution, including deterministic and probabilistic approaches, machine-learning techniques, and strategies for handling complex scenarios. You will gain a solid understanding of the principles behind these methods, as well as the strengths and limitations of different techniques. A key element of this course is on the production of code to facilitate entity resolution tasks. Practical sessions will guide you through the implementation of record linkage algorithms using the fastLink package. You will learn to write efficient, scalable code that can handle large datasets, perform data cleaning and preprocessing, and accurately link records. By the end of this course, you will not only have a strong theoretical grounding but also hands-on experience in developing and deploying practical solutions, empowering you to drive data integration projects with confidence.

Target audience / prerequisites

Anyone interested in learning more about entity resolution. Familiarity with the R statistical software. A laptop with R and fastLink installed.

Instructor's biography Ted Enamorado is an Associate Professor of Political Science at Wash-

ington University in St. Louis, where he is affiliated with the Center for the Study of Race, Ethnicity & Equity, the Division of Computational & Data Sciences, and the Political Data Science Lab. He holds a Ph.D. in Politics from Princeton University, where I specialized in Political Economy and Political Methodology. His research focuses on improving probabilistic methods, particularly in the context of record linkage and data integration. These methods have recently been applied to various areas, such as examining inequities in the criminal justice system, evaluating social desirability in survey research, and analyzing political campaign contributions. Before pursuing his graduate studies, he worked as a Research Fellow at the Inter-American Development Bank and as a Consultant at the World Bank.

Bayesian small area estimation, with an emphasis on low- and middle-income countries

Lecturer:	Jon Wakefield
Affiliation:	School of Public Health, University of Washington
Date:	Friday, 19 June 2026
Format:	Slot A: 9:00–10:30; Slot B: 11:00–12:30
Room:	TBA

Course description

Small area estimation is of crucial importance in low- and middle-income countries (LMICs). A modern Bayesian treatment will be presented and illustrated using a range of examples. Area-level (Fay-Herriot) and unit-level models will be presented. Unit-level models for both linear and generalized linear models will be discussed. Fast computation is carried out with the Integrated Nested Laplace Approximation (INLA) method, which is embedded within the SUMMER and surveyPrev R implementation. Hyperprior specification is via penalized complexity priors. Between-area variation will be modeled using independent and spatial random effects. For the latter, the Besag, York, Mollié model will be described.

The course covers the following learning outcomes: An introduction to Bayesian statistics; Limitations of direct estimates; A comprehensive description of area-level and unit-level models, including advantages and disadvantages; Borrowing of strength and the bias/variance trade-off; Mixed effects models; Spatial models; Regression modeling; Parameter interpretation; Bayesian computation; Prior specification.

Target audience / prerequisites

Anyone with an interest in learning about modern Bayesian methods, with an emphasis on LMICs, though the methods are generally applicable. Some knowledge of basic probability and linear and logistic regression models will be useful.

Preparatory material, including software

The site: https://sae4health.stat.uw.edu/ contains links to software. We will, discuss the

surveyPrev and SUMMER R packages which carry out spatial and spatio-temporal modeling, respectively. These packages have high-quality mapping functions. Pre-modeled estimates are available at https://sae4lmic.stat.uw.edu/ while users can analyze DHS data from X countries from over XX surveys at https://rsc.stat.washington.edu/sae4health/

Instructor's biography

Jon Wakefield (https://faculty.washington.edu/jonno/) is a professor in the Departments of Statistics and Biostatistics at the University of Washington in Seattle. He is a statistician with interests in small area estimation, demography, spatial epidemiology, global health, and exploring the links between Bayesian and frequentist estimation procedures. He wrote the book "Bayesian and Frequentist Regression Models", which was published in 2013 by Springer and has published 187 peer-reviewed papers and 29 book chapters, with a h-index of 71 and over 20K citations. He has extensive experience in SAE in LMICs, working closely with the UN, WHO and Gates, and has given workshops in multiple countries including Nigeria, Rwanda, South Africa, Malawi, Ecuador. Dr Wakefield's work is focused on putting tools into the hands of local researchers. He has also taught multiple short courses on Spatial Epidemiology and on SAE, as part of the UW summer school program. He has graduated 31 PhD students and runs the Space Time Analysis Bayes (STAB) working group (https://alanamcgovern.github.io/stablab/). Dr Wakefield received the Guy Medal in Bronze in 2000 from the Royal Statistical Society (one award is made per year and is given for excellence of research appearing in its journals and/or conferences). He was elected a Fellow of the American Statistical Association in 2007 (each year, no more than 1/300 of the society's membership can be selected for this honor). Dr Wakefield's group produced subnational estimates for under-5 mortality and neonatal mortality (death in the first month of life) for the UN Inter-Agency Group for Mortality Estimation (IGME), for around 30 LIMCs (https://childmortality.org/)

Invited session abstracts

IS1: Advances in robust inference for small area estimation

Date, time, room: Tuesday, 16 June 2026, 11:00–12:30, SAE 1
Organisation: Organizer/chair: David Haziza; Chair: Marius Stefan

Integrating big data and alternative sources into small area estimation for poverty estimation

Speaker: Monica Pratesi
Authors: Monica Pratesi, Francesco Schirripa-Spagnolo, Caterina Giusti, Nicola Salvati, Antonella D'Agostino

Abstract

Nowadays, the availability of non-traditional data sources, including administrative records, satellite imagery, mobile phone data, and citizen-generated data, commonly referred to as big data, is steadily increasing. At the same time, their unprecedented spatial granularity offers new opportunities in the context of Small Area Estimation (SAE) to infer characteristics for very small domains. However, data obtained from big data sources are often generated through non-probability sampling processes, making adjustment for selection bias a critical practical challenge. This presentation explores the integration of big and alternative data sources into SAE models, with a focus on methodological innovations, practical challenges, and real-world applications. We discuss how these data sources can be incorporated as auxiliary information, how issues of representativeness and selection bias can be addressed, and how to balance model complexity with interpretability. Finally, we propose a novel approach to reducing selection bias in big data sources within the SAE framework. Our method is based on data integration, combining information from a big data sample with that from a probability sample.

Double robust small area estimation

Speaker: Jiming Jiang
Authors: Haiqiang Ma, Zhiyan Sheng and Jiming Jiang

Abstract

In the context of robust small area estimation, there are two types of robustness considerations, robustness against model misspecification and robustness against outliers. We propose a method of SAE that has both types of robustness features. The method combines the idea of observed best prediction, which is known to be more robust against model misspecification than the traditional best linear unbiased prediction method, and the method of density power divergence, which is known to be more robust against outliers than the EBLUP. The double robust predictor is developed under an area-level model with normal or normal-mixture sampling errors, and under a unit-level model. Another advantage of the DRP method is that it provides a natural estimator of a tuning parameter involved in the DPD. We develop theory about the proposed method, and

demonstrate empirical performance of the proposed DRP and its comparison to EBLUP, OBP, a robust version of the EBLUP, and predictors based on the DPD. A second-order unbiased estimator of the mean squared prediction error of the DRP is developed and its performance is evaluated. A real-data example is discussed.

Approximately conformal shrinkage interval for small area estimation

Speaker: Li-Chun Zhang
Authors: Li-Chun Zhang

Abstract

Provided IID or exchangeable distributions, conformal interval is an assumption-lean method of predictive inference, whose coverage does not depend on a chosen predictor of the target outcome. For finite-population small-area estimation based on area-level direct estimators, we consider inference only with respect to repeated sampling, and develop approximately conformal shrinkage intervals, where the target small-area parameters are treated as fixed constants instead of random variables. In addition, modelling the fixed-effect expectation of the area-level direct estimators can help to reduce the interval width. The proposed ACSI is more assumption-lean, simpler to implement and more resilient against model misspecification, compared to model-based inference of empirical best linear unbiased prediction that is perhaps the most common in practice.

IS2: Contributions to model-based small area estimation

Date, time, room: Tuesday, 16 June 2026, 11:00–12:30, SAE 2
Organisation: Organizer/chair: Domingo Morales

Generalized M-Quantile small area estimation under working models

Speaker: María Bugallo
Authors: María Bugallo, Ray Chambers, Nicola Salvati

Abstract

M-quantile regression provides a robust alternative to mixed-model approaches for Small Area Estimation, but standard plug-in M-quantile predictors often exhibit bias. Although smearing-type corrections can be applied, they typically lead to a loss of efficiency. In this work, we propose a new distribution-based M-quantile predictor that incorporates bias correction in a principled and efficient way. Our approach relies on the Generalised Asymmetric Least Informative distribution as a working model for M-quantile prediction errors. By exploiting its closed-form first two moments, we derive explicit area-specific bias corrections that yield unbiased and efficient predictors under the working model, together with robustified versions to ensure numerical stability. The proposed framework also stabilises the area-specific quantile index in settings with small within-area sample sizes. An extension to binary data is developed through a Generalised Asymmetric Bernoulli specification. Simulation studies highlight substantial improvements in bias reduction and predictive accuracy, particularly for small areas.

Shapley value-based variable selection for Fay-Herriot models

Speaker: Esteban Cabello
Authors: Esteban Cabello, Juan Carlos Gonçalves-Dosantos, Domingo Morales, Joaquín Sánchez-Soriano

Abstract

Optimal model specification is a critical challenge in Small Area Estimation, particularly for area-level models such as the Fay-Herriot. Conventional selection procedures often rely on stepwise algorithms and information criteria, which tend to evaluate the model globally and fail to isolate the specific contribution of individual auxiliary variables. To address this issue, we present a framework based on cooperative game theory that uses the Shapley value to quantify the marginal contribution and average importance of each predictor. By assessing all potential covariate subsets, this method provides a robust measure of influence. We conducted simulation experiments to investigate the new methodology's performance and effectiveness in detecting the model's generating variables and reconstructing the data-generating model. We applied the new methodology to the 2022 Spanish Living Conditions Survey to identify the best predictors for estimating poverty proportions in Spanish provinces by sex.

Divide and conquer strategy for multivariate Fay-Herriot models

Speaker: Alexandro Aneiros-Batista
Authors: Alexandro Aneiros-Batista, Esther López-Vizcaíno, María José Lombardía

Abstract

Area-level models remain central to small area estimation, yet their practical use becomes computationally demanding when the number of areas is large, especially in the actual world of Big Data. To address this challenge, we introduce a Divide and Conquer framework for multivariate Fay–Herriot models that decomposes the estimation problem into smaller, independent subsystems and subsequently integrates the partial results through a principled aggregation step. The proposed method preserves the statistical properties of the full model while substantially reducing computational cost and improving numerical stability in high-dimensional or ill-conditioned settings. Simulation studies based on bi-, tri- and tetravariate Fay–Herriot populations confirm that the Divide and Conquer approach yields nearly unbiased point estimates comparable to those of the global estimator, while reducing execution time. The methodology is further illustrated using data from the 2024 second-quarter Spanish Quarterly Labour Cost Survey, providing efficient small-area predictions of mean wage costs and effective hours by economic sector and firm size. The proposed strategy offers a general, scalable framework for modern small area estimation applications where dimensionality and heterogeneity make traditional joint estimation impractical.

IS3: New developments in small area estimation

Date, time, room: Tuesday, 16 June 2026, 14:00–15:30, SAE 1
Organisation: Organizer: Gauri S. Datta; Chair: Scott Holan

Bias corrected variance stabilizing transformation for small area estimation

Speaker: Malay Ghosh
Authors: Masayo Hirose, Malay Ghosh, Mayumi Oka

Abstract

Small area estimation models are typically based on the normality assumption of response variables. More recently, attention has been drawn to the transformation of the original variables to justify the assumption of normality. Variance stabilizing transformation of observations serves the dual purpose of reaching closer to normality, as well as known variance of the transformed variables in contrast to the assumption of known variances of the original variables, the latter needed to avoid non-identifiability. However, the existing literature on the topic ignores a certain bias introduced in the seemingly correct back transformation. The present paper rectifies this deficiency by introducing asymptotically unbiased empirical Bayes (EB) estimators of small area means. Mean squared errors (MSEs) and estimated MSEs of such estimators are provided. The theoretical results are accompanied with simulations and data analysis. A somewhat surprising phenomenon is a finding which connects one of our results to the natural exponential family quadratic variance function (NEF-QVF) family of distributions introduced by Morris (1982,1983).

Considerations for fitting unit-level generalized linear mixed effects models to complex survey data in small area estimation

Speaker: Yajuan Si
Authors: Yajuan Si, Katherine Li, Brady T. West, Elina T. Page, Xingyou Zhang

Abstract

Complex sample survey data are often used to produce small area estimates (SAEs) that guide policy decisions for specific geographic or demographic groups. Unit-level, model-based approaches have the potential to enhance the precision and accuracy of SAEs by relating outcomes to predictors at various group and individual levels. However, accurately estimating SAEs while accounting for the complexities of sampling design remains challenging. Our goal is to assess SAE methods that are standardizable across statistical software, easily accessible to analysts, and straightforward to implement. We compared two model-based approaches for calculating SAEs under a two-stage stratified-cluster sampling design. Both approaches share a common process: point estimation using (weighted/unweighted) unit-level generalized linear mixed effects models (GLMMs), and standard error estimation via resampling. The first method employs a weighted GLMM, addressing design features (including clusters and strata) through the jackknife repeated

replication procedure. The second method utilizes the weighted finite population Bayesian bootstrap (WFPBB), incorporating all design features with synthetic populations. We found that both methods tend to overestimate the variance of random intercepts. Furthermore, the SAEs generated from GLMM predictions under WFPBB were similar to weighted direct estimates. Our results highlight the need for caution when fitting unit-level GLMMs to complex sample survey data, as current methods may not provide valid estimates of conditional effects or SAEs.

Bias asymptotics for estimating-equation solutions, with application to the Fay-Herriot SAE model

Speaker: Eric Slud
Authors: Eric Slud, Gauri Datta, Kyle Irimata, Jerry Maples

Abstract

Motivated by the FH model with non-normal area effects, this paper develops the regularity conditions needed to establish rigorously the bias and variance expansions for parameter estimates and predictors. The regularity conditions involve moments of the area effects and empirical moments of the predictor variables, and apply to any of the estimating equations commonly used to estimate the FH model parameters. The bias expansions are needed to ensure that sample moments in Monte Carlo simulations of estimates and area predictors conform closely to theoretical moment expressions. An essential step is to bound with remainder the probability of the exceptional set on which theoretical expressions appearing in limit theorems fail to behave as asserted in those theorems. These theoretical estimates are closely related to the empirical frequency with which the expected numerical convergence of estimating-equation solutions fails in simulations. Beyond these theoretical considerations, the talk will present simulation results indicating in some moderate-sample FH applications that the biases realized in practice for non-normal SAE estimates and predictions can be much worse than those that have been reported in past published simulations.

IS4: Recent approaches in neural methods and spatial modeling for small area estimation and synthetic population generation

Date, time, room: Wednesday, 17 June 2026, 9:00–10:30, SAE 1
Organisation: Organizer: Scott H. Holan; Chair: Gauri Datta

A Variational Bayesian neural diffusion approach for synthetic population generation and hierarchical downscaling

Speaker: Christopher K. Wikle
Authors: Christopher K. Wikle

Abstract

We develop a Variational Bayesian diffusion modeling framework for mixed-type tabular data that incorporates principled epistemic uncertainty into both continuous and categorical generative diffusion processes. Our approach augments standard Denoising diffusion Probabilistic Models (DDPM) and multinomial diffusion models with Bayesian linear prediction heads while keeping the diffusion backbones deterministic, yielding a tractable variational posterior. This posterior induces a distribution, enabling posterior predictive inference and uncertainty-aware synthetic data generation. By coupling the diffusion mechanisms with Variational Bayes and a novel learnable weighting scheme for heterogeneous features, the model provides calibrated uncertainty quantification and improved fidelity for complex mixed-type data. Coupled with copula-based hierarchical downscaling, this provides a theoretical advance in population synthesis by replacing point-estimate generators with a posterior-predictive ensemble that maintains high-order attribute dependencies, marginal summaries, and spatial fidelity. The ultimate goal is to provide public use synthetic populations over geographies that can be used to facilitate modeling decision making in disaster evacuation scenarios. We demonstrate this on American Community Survey Public Use Microdata applied to lower-level geographies in California.

Small area estimation of wealth quantiles over time, a Bayesian unit-level modeling approach

Speaker: Daniel Vedensky
Authors: Daniel Vedensky

Abstract

Modeling wealth quantiles over time and space and across fine-grained demographic classifications is of vital importance to policy makers. However, such data present unique challenges for standard small area estimation methodologies because they are highly skewed and heavy-tailed with both positive and negative values. Traditional area-level modeling and design-based approaches struggle to capture these features. Meanwhile, existing unit-level methodology does not apply to longitudinal surveys and frequently treats the survey design as ignorable. Often the analysis

is restricted to alternative response variables that are non-negative, such as poverty indicators or income. To overcome these limitations, we propose a Bayesian, asymmetric Laplace pseudo-likelihood model that is able to capture the major features of wealth while adjusting for survey weights and accounting for both spatial and temporal dependence. We illustrate the proposed methodology with an application to the Survey of Income and Program Participation.

Variational autoencoded multivariate spatial Fay-Herriot models

Speaker: Scott H. Holan
Authors: Scott H. Holan

Abstract

Small area estimation models are essential for estimating population characteristics in regions with limited sample sizes, thereby supporting policy decisions, demographic studies, and resource allocation, among other use cases. The spatial Fay-Herriot model is one such approach that incorporates spatial dependence to improve estimation by borrowing strength from neighboring regions. However, this approach often requires substantial computational resources, limiting its scalability for high-dimensional datasets, especially when considering multiple (multivariate) responses. This presentation proposes two methods that integrate the multivariate spatial Fay-Herriot model with spatial random effects, learned through variational autoencoders, to efficiently leverage spatial structure. Importantly, after training the variational autoencoder to represent spatial dependence for a given set of geographies, it may be used again in future modeling efforts, without the need for retraining. Additionally, the use of the variational autoencoder to represent spatial dependence results in extreme improvements in computational efficiency, even for massive datasets. We demonstrate the effectiveness of our approach using 5-year period estimates from the American Community Survey over all census tracts in California.

IS5: New advances in small area estimation and the analysis of longitudinal data

Date, time, room: Wednesday, 17 June 2026, 9:00–10:30, SAE 2
Organisation: Organizer: Thuan Nguyen; Chair: Jiming Jiang

Continuation ratio bridging models for local estimation of dementia prevalence

Speaker: Carolina Franco
Authors: Carolina Franco

Abstract

Dementia is widely recognized as one of the most pressing public health challenges today, and reliable, geographically disaggregated prevalence estimates for demographic subgroups are essential for planning, resource allocation, and targeted surveillance. Existing dementia data are insufficient to reliably measure prevalence below the national level. In this study, we estimate dementia prevalence at the state level by age (65–79, 80+), race/ethnicity, and sex by integrating information from the Health and Retirement Study (HRS), administrative records, and other publicly-available sources. HRS data linked to Medicare records support the estimation of diagnosed, undiagnosed, and misdiagnosed dementia using detailed cognitive assessments and diagnosis codes, but its sample size limits subnational inference. Administrative records derived from Medicare— a U.S. federal health insurance program with near-universal coverage of adults aged 65 and older—enables granular estimation but captures only diagnosed dementia and is subject to differential diagnostic bias across populations (e.g., Giannattasio et al. 2019). We address these challenges using extensions of continuation-ratio (CR) models, introduced to the small area estimation literature by Slud, Franco, and Hall (2024) and Rein, Franco et al. (2024), fit within a hierarchical Bayesian framework. These models exploit the natural ordering of dementia diagnostic states observed in HRS and bridge this information to local Medicare-based estimates of diagnosed dementia, borrowing strength across areas and additional auxiliary administrative data sources to produce calibrated small area prevalence estimates.

A random slopes two-fold Fay-Herriot model for small area income estimation

Speaker: Naomi Diz Rosales
Authors: Naomi Diz Rosales. María José Lombardía. Domingo Morales

Abstract

This work presents a temporal two-fold Fay-Herriot model with random slopes to improve small area income estimation. By combining random slopes with an AR(1) structure, the model captures complex domain-specific trends and enhances the borrowing of strength across time and space. We provide a comprehensive framework for REML estimation and EBLUP prediction, including a robust analytic MSE estimator that optimizes the bias-computation trade-off. A key

contribution of this work is the development of a novel R package, designed to implement these advanced estimators efficiently for practitioners. Simulations confirm that our approach significantly increases precision over standard models in small-sample scenarios. Finally, the methodology is applied to Spanish disposable income data (2013–2022), revealing critical regional disparities and a widening post-pandemic gender gap.

A random-effects approach to generalized linear mixed model analysis of incomplete longitudinal data

Speaker: Thuan Nguyen
Authors: Thuan Nguyen

Abstract

Longitudinal data are widely encountered in health and medical studies. It is quite common that the original data collected from a longitudinal study involve missing values, which can significantly complicate the statistical analysis. For example, even if a single component of the data record is missing, the data record is incomplete. On the other hand, even if some components of the data are missing, the remaining components could still provide useful information; discarding such information would be wasteful. We propose a random-effects approach to missing values for generalized linear mixed model (GLMM) analysis of longitudinal data. The method converts a GLMM with missing covariates to another GLMM without missing covariates. The standard GLMM analysis tools then apply. The method applies, in particular, to the cases of linear mixed models and logistic regression. Performance of the method is evaluated empirically, and compared with alternative approaches, including the popular MICE procedure of multiple imputation. Theoretical justification of the method is given, and explained, for the patterns observed in the simulation studies. Two real-data examples from healthcare studies are discussed.

IS6: Small area statistics with multiscale data

Date, time, room: Wednesday, 17 June 2026, 14:00–15:30, SAE 1
Organisation: Organizer/chair: Sanjay Chaudhuri

A Bayesian Approach to Produce Subnational Population Estimates Using a Population Base Statistical Register

Speaker: Snigdhansu Chatterjee
Authors: Snigdhansu Chatterjee

Abstract

Subnational Population Estimates (SPE) in Latin America are useful to implement new public policies in subnational areas with internal armed conflicts or difficult to access. In this work, we propose to combine a Population Base Statistical Register (PBSR) and the Official Population Projections (OPP) using a Bayesian approach to produce SPE. Our proposed procedures are useful for computing SPE of the population size or the SPE of the population size in percentage SPE (%). However, we focused on SPE (%) due to some data restrictions and to ensure data confidentiality. In this article, the PBSR is constructed using multiple administrative sources with registers from the health, education, vital statistics systems, tax registration, and, more importantly, the registers of the victims of the current internal armed conflict in Colombia. We also propose new fast Markov chain Monte Carlo algorithms to produce SPE (%) using data augmentation procedures to address the complications caused by the resulting joint posterior containing gamma functions. We implement our proposal to compute SPE (%) by age and sex groups in the municipality of Jamundí in Colombia which is currently affected by poverty, forced displacement, and the internal armed conflict and evaluate the accuracy with a Population Census.

Improving small area estimation through robust prior specification

Speaker: Jairo A. Fuquene Patino
Authors: Jairo A. Fuquene Patino

Abstract

In this work, we propose robust priors within unit-level area models to address heterogeneity and apply the proposed approach to estimate mortality at subnational levels in low- and middle-income countries.

Empirical Likelihood-based Methods for Multiscale Data Integration

Speaker: Sanjay Chaudhuri
Authors: Sanjay Chaudhuri

Abstract

It is well-known that the empirical likelihood-based methods provide an easy way of data integration in a vast array of real-life problems. In this talk, we will discuss the application of such methods to multiscale data integration. Basic theory of empirical likelihood-based data integration will be introduced. Based on these preliminaries, we will discuss new methods for the integration of multiscale data. The methods would be illustrated with real-life examples.

IS7: Innovations in small area estimation for modern data challenges

Date, time, room: Wednesday, 17 June 2026, 14:00–15:30, SAE 2
Organisation: Organizer: Mahmoud Torabi; Chair: Gauri Datta

Post-2024 U.S. presidential election analysis: real-life validation of prediction via small area estimation and uncertainty quantification

Speaker: Jiming Jiang
Authors: Jiming Jiang, Zheshi Zheng, Yuanyuan Li, Peter X.K. Song

Abstract

We carry out a post-election analysis of the 2024 U.S. Presidential Election (USPE) using a prediction model derived from the Small Area Estimation (SAE) methodology. With pollster data obtained one week prior to the election day, retrospectively, our SAE-based prediction model can perfectly predict the Electoral College election results in all 44 states where polling data were available. In addition to such desirable prediction accuracy, we introduce the probability of incorrect prediction (PoIP) to rigorously analyze prediction uncertainty. Since the standard bootstrap method appears inadequate for estimating PoIP, we propose a conformal inference method that yields reliable uncertainty quantification. We further investigate potential pollster biases by the means of sensitivity analyses and conclude that swing states are particularly vulnerable to polling bias in the prediction of the 2024 USPE.

Small area models for proportions: an area-level EFD model

Speaker: Serena Arima
Authors: Serena Arima

Abstract

In this contribution, we focus on small-area compositional data. These data are defined as vectors whose elements are strictly positive and sum to one (e.g., proportions). Compositional data arise in various fields, including medicine, economics, psychology, and environmetrics. They are defined on the D-part simplex (S^D) and require complex techniques for proper analysis.

A traditional approach involves applying log-ratio transformations, which map the D-part simplex onto a $(D-1)$-dimensional real space. However, this approach has several limitations: parameter estimates are interpretable only in the transformed space, and issues such as skewness, heteroscedasticity, non-normality, and outliers may bias inference. An alternative solution for regression with compositional responses is the Dirichlet model. It is often implemented using a multinomial logit function to link the response mean vector to covariates, allowing for a straightforward interpretation of regression coefficients in terms of log-odds ratios. Nevertheless, the Dirichlet model has significant drawbacks: it uses only a single parameter to describe the

entire variance–covariance matrix (the precision parameter), offers limited flexibility, and implies several forms of simplicial independence. Moreover, it always yields negative covariances, making it unable to model many relevant phenomena (e.g., heavy-tailed or multimodal responses).

We therefore propose using the Extended Flexible Dirichlet (EFD), a structured mixture of Dirichlet components. Unlike general Dirichlet mixture models, the parameters of each EFD component are strictly linked to one another, providing greater flexibility. We employ the EFD for modelling small-area data: we reparameterize the model in terms of mean and precision and incorporate covariates to account for design effects. We estimate the model within a fully Bayesian framework and assess its performance using simulated data.

Benchmarking priors: A fully Bayesian approach to small area benchmarking

Speaker: Anindya Roy
Authors: Anindya Roy

Abstract

We develop a new class of priors for small area models that allows a full Bayesian treatment of small area models under strict benchmarking constraints. The prior is fully supported on the constrained parameter set defined by the benchmarking constraints and thereby provides a proper constrained posterior that is also supported on the restricted set. The posterior naturally constrains the small area estimates to conform to the benchmarking values at the aggregate level. Also, it facilitates valid uncertainty quantification using credible sets that are fully contained within the constrained set. The framework has a wide range of applicability, ranging from the standard Gaussian Fay-Herriot model to more general Fay-Herriot models, such as robust Fay-Herriot models, where the sampling errors are distributed as multivariate-t or Fay-Herriot models where there is a complex nonlinear relationship between the area means and the covariates. We illustrate the usefulness of the proposed methodology using simulation experiments as well as the 2019 American Community Survey Public Use Microdata Sample data from Maryland on Per Capita Income. The numerical results reaffirm the advantages of the proposed benchmarking priors.

IS8: Time series methods for official statistics

Date, time, room: Wednesday, 17 June 2026, 16:00–17:30, SAE 1
Organisation: Organizer: Jan van den Brakel; Chair: Ika Wulansari

Trend estimates for a detailed breakdown of mobility indicators

Speaker: Harm Jan Boonstra
Authors: Harm Jan Boonstra, Jan van den Brakel

Abstract

The Dutch Travel Survey aims to produce reliable estimates of mobility for the Dutch population. To obtain consistent mobility trends across several breakdowns, we develop multilevel time-series models that account for survey redesigns and other pronounced measurement effects in one particular year. The target indicators are the number of trip legs per person per day, and the distance and duration per trip leg. Annual estimates are produced for breakdowns by trip characteristics (purpose and transport mode) and personal characteristics (gender and age group). The resulting full cross-classification yields 700 trend series for the period 1999-2025. Because of the complex hierarchical structure of trips nested within persons and the large annual sample sizes, we employ area-level models that use transformed direct estimates and smoothed direct standard errors as inputs. The models are specified in a hierarchical Bayesian framework and fitted using Markov Chain Monte Carlo (MCMC). To handle different types of outliers, we use Student-t sampling distributions for the transformed distance and duration variables and global-local priors for selected coefficients, including those representing redesign effects. The models borrow strength over time and across domains - such as adjacent age groups - and accommodate the atypical trend patterns observed during the COVID-19 period. The resulting estimates provide both detailed and aggregated trend series in which discontinuities caused by survey redesigns are effectively corrected.

Time series area level model for the Dutch Safety Monitor

Speaker: M.L.J. Peerlings
Authors: M.L.J. Peerlings, Jan van den Brakel, Harm Jan Boonstra

Abstract

The Dutch Safety Monitor (DSM) of Statistics Netherlands (CBS) aims to provide insight into the safety situation in the Netherlands. It is a periodic survey that collects information about the population's experiences with crime, feelings of insecurity, opinion about police performance, and the impact of crime on people. Until 2019 the DSM was conducted annually, but since 2019 it is conducted bi-annually. Due to the desired level of detail, direct estimates based on the survey weights are too inaccurate for most subpopulations of interest. In addition, there is demand for predictions for years for which no survey data are available because the survey is conducted

bi-annually since 2019. In this paper an area level time series model casted in an hierarchical Bayesian framework is presented to produce model-based estimates on a refined regional level. The model is used as a form of small area estimation that borrows strength over time and space and provides predictions for the years with no survey data. It is also shown how the model accommodates discontinuities that are the result of survey redesign. To quantify effects, the old and new design were conducted in parallel for some period during the year of the change-over. It is shown how the information from this parallel run is incorporated in this time series model through the use of informative priors for the interventions.

Dynamic synchronized models for national accounts data

Speaker: Jan van den Brakel
Authors: Jan van den Brakel, Lucas Harlaar, Siem Jan Koopman

Abstract

In this paper a multivariate time series model is proposed for the complete national accounts dataset containing at least nine variables. The dynamic features of the variables are stylized by a set of unobserved components that represent stochastically evolving trend, cyclical and seasonal effects. The components can be uniquely associated with a single variable or they can be shared by multiple variables simultaneously. The model considers both expenditure and production variables, and includes constraints for a synchronized and consistently-defined gross domestic product variable. The model constraints can be treated naturally within the Kalman filter and smoother recursions. Parameter estimation is based on exact maximum likelihood which is feasible despite the high-dimensional parameter vector. The proposed model-based framework is used primarily for synchronized nowcasting and forecasting of gross domestic product. We show that the statistical precision of signal extraction increases within our framework, when compared to univariate and unconstrained model alternatives. Hence, we provide potentially more accurate methods for trend extraction, seasonal adjustment, and forecasting of macroeconomic variables in national accounts. Finally, when the model includes a common cycle component that is shared by all variables in national accounts, it can be interpreted as the macroeconomic cycle indicator at current prices. We empirically illustrate our model-based methodology for national accounts data from Germany, Italy and the Netherlands. We show that our analyses can be of significance for both official statistics and macroeconomic policy-making.

IS9: Advances in small area estimation of poverty and socioeconomic indicators using integrated and high-dimensional models

Date, time, room: Wednesday, 17 June 2026, 16:00–17:30, SAE 2
Organisation: Organizer/chair: Angelo Cozzubo

An integrated approach to quarterly small area estimation of extreme poverty and design-effects in Brazil using Beta-binomial Models

Speaker: Guilherme Anthony Pinheiro Jacob
Authors: Guilherme Anthony Pinheiro Jacob

Abstract

The Brazilian National Statistical Office (IBGE) uses its main household survey, PNADC, to produce annual estimates of the extreme poverty rate at the country level, as well as for each of its 27 Federative Units (26 states and the Federal District). Although poverty monitoring could benefit from more frequent and spatially disaggregated estimates, those obtained via direct estimation from the PNADC are considered too unreliable for informing public policy due to their large uncertainty. In this article, we review the current Fay-Herriot approach for estimating extreme poverty rates and propose an integrated modeling of direct estimates and design-effects via Beta-binomial model, combining PNADC data with covariate information obtained from an administrative register of beneficiaries of social programs in Brazil (CadUnico). Instead of a preliminary ad-hoc smoothing of variance estimates, this new approach also models the design effects estimates and avoids some of the issues related to the discretization problem faced when using the effective sample size. Another recent small area estimation project at IBGE is also briefly discussed.

Estimating disease prevalence at local level

Speaker: Sonja Stiebahl
Authors: Sonja Stiebahl, Carl Baker

Abstract

We outline a method of estimating local level disease prevalence by combining GP practice level data with patient residence information. Quality and Outcomes Framework (QOF) data for England includes GP practice level information on the prevalence of 21 health conditions among registered patients. Our analysis joins the QOF practice-level prevalence data with further NHS Digital data on patient residence – specifically, Lower layer Super Output Area (LSOAs) where patients registered to each practice live.

The process includes adjustments using Office for National Statistics (ONS) annual mid-year population estimates. QOF registers for some health conditions only relate to particular age groups rather than the whole population. However, published data on GP patients by LSOA

does not record information on patients' age. Hence, we use ONS LSOA figures to model the number of registered patients in each age group.

Other adjustments are applied to address cases where the GP registered population appears notably lower than the resident population - suggesting a substantial proportion of the resident population is not registered with a GP. This may be areas with a large student population, those containing military bases, prisons etc. On a case-by-case basis these are ""paired"" with a neighbouring LSOAs and assigned that LSOA's demographics.

The LSOA estimates produced can be aggregated to other levels. Our method has been used to compile prevalence estimates for Westminster parliamentary constituencies and Middle layer Super Output Area (MSOA) as shown in our published data dashboard: Constituency data: health conditions Our presentation will provide an overview of the methodology to compile local area estimates and highlight key findings.

Note: The analysis covers England only, because equivalent QOF and GP practice populations data isn't available for Wales, Scotland or Northern Ireland.

IS10: Small area estimation for poverty: new data sources and advanced modeling

Date, time, room: Thursday, 18 June 2026, 9:00–10:30, SAE 1
Organisation: Organizer/chair: Nicola Salvati

Small area estimation with machine learning algorithms

Speaker: Caleb Leedy
Authors: Caleb Leedy

Abstract

In this talk, we propose a general purpose methodology that allows for nonparametric mean estimation within the Fay-Herriot model. Using Neyman orthogonalization and sample splitting, we develop small area predictors that are robust to finite-dimensional model assumptions and we include the theory to guarantee second order unbiased estimation of the mean square prediction error.

The use of spatial information in SAE models with arcsine transformation: estimating food affordability in Italy

Speaker: Stefano Marchetti
Authors: Stefano Marchetti

Abstract

In the context of small area estimation, when the target statistic is a proportion, arcsine or logarithmic transformations are often employed. Moreover, in some applications, incorporating spatial information leads to increased efficiency of small area estimates. This paper proposes the incorporation of spatial information through a simultaneously autoregressive process within an area-level model with an arcsine transformation. The estimator accounts for a bias-corrected back-transformation, and its mean squared error (MSE) is estimated using a parametric bootstrap. The proposed method is applied to the estimation of food poverty at the provincial level in Italy. In this application, food poverty is defined as the ability to afford a basket consistent with a healthy and sustainable diet. The cost of the diet is derived from web-scraped data from the Price Observatory of the Ministry of Enterprises and Made in Italy, while food consumption expenditure is obtained from the Household Budget Survey, carried out by Istat.

Empirical best prediction of poverty indicators using nested error regression with high-dimensional parameters

Speaker: Partha Lahiri
Authors: Partha Lahiri

Abstract

This paper extends the nested error regression model with high-dimensional parameters (NERHDP) to address key challenges in small area poverty estimation. Building on the NERHDP framework, we propose a flexible and robust approach for deriving empirical best predictors (EBPs) of small area poverty indicators that accommodates heterogeneity in regression coefficients and sampling variances across areas. To overcome the computational limitations of existing algorithms, we introduce an efficient estimation procedure that substantially reduces computation time, thereby improving scalability to large datasets. In addition, we develop a novel method for producing area-specific poverty estimates for out-of-sample areas, enhancing the reliability of synthetic predictions. Uncertainty is quantified using a parametric bootstrap procedure tailored to the extended model. Design-based simulation studies show that the proposed method outperforms existing approaches in terms of relative bias and relative root mean squared prediction error. Finally, we apply the proposed methodology to household survey data from the 2002 Albania Living Standards Measurement Survey, combined with auxiliary information from the 2001 census, to estimate poverty indicators for 374 municipalities.

IS11: Inference on small area parameters from probability and non-probability samples

Date, time, room: Thursday, 18 June 2026, 9:00–10:30, SAE 2
Organisation: Organizer: Marius Stefan; Chair: Isabel Molina

Small area estimation based on nonprobability samples

Speaker: Danny Pfeffermann
Authors: Danny Pfeffermann, Michael Sverchkov

Abstract

During the last decade, many articles have been published, considering estimation of finite population parameters based on nonprobability samples. Most of these articles, assume that the probability of inclusion in the sample depends on covariates x, but not on the target response variable y. Only few articles consider the case where the probability of inclusion in the nonprobability sample depends also on y, known as informative sampling. In this presentation, we shall consider small area estimation based on nonprobability samples, assuming the case of informative sampling, without knowledge of x values for units outside the sample, except for the population means of some or all of them, and the true area means in some of the areas.

Robust empirical best estimation of complex small area parameters under unit level nested error linear regression models

Speaker: J.N.K. Rao
Authors: Sanjoy K. Sinha, J.N.K. Rao

Abstract

Empirical best (EB) estimation of complex small area parameters, in particular poverty indicators used by the World Bank, was studied by Molina and Rao (2010) under standard nested error linear regression models, assuming normality of random effects. But such EB estimators are generally sensitive to outliers and other departures from standard assumptions. In this paper, we propose and explore robust EB (REB) estimators that are resistant to outliers often encountered in real data. We also study Census REB (CREB) estimators that do not require linking of the sample units to the population units and hence more useful in practice. Bootstrap estimation of mean square error of REB and CREB estimators is also studied. Empirical properties of the proposed estimators are studied using Monte Carlo simulations. An application is also provided using read data from a health study.

Design mean square error estimation for small area means under unit-level models

Speaker: Marius Stefan
Authors: Marius Stefan, J.N.K Rao

Abstract

Model-based empirical best (EB) predictors are widely used in small area estimation (SAE). The model mean square error (mMSE) is generally used to measure the variability of these predictors. However, survey practitioners are more familiar with the design mean square error (dMSE), a measure which accounts for the variability originating from the randomness of the sample selection, with the population values of the survey variable held fixed. The literature on dMSE estimation of model-based predictors is much more limited than that on mMSE estimation. For example, in the case of the EB predictor, no theoretical properties are known on the bias of dMSE estimators proposed earlier in the literature. Besides, most of these dMSE estimators assume the correctness of the small area model, which may be problematic when the model is misspecified. In this article, we propose a new class of model-based predictors called predictors of empirical best type (EBt) constructed under the basic unit-level model. Our simulations show that the new EBt predictors are approximately as efficient as the EB predictor. The EBt predictors are mainly motivated by the possibility to construct estimators of their dMSE which are unbiased under simple random sampling within small areas, irrespective whether the underlying model holds or not. Therefore, an EBt predictor may be an alternative to the standard EB predictor.

IS12: Modern advances in surveys and small area statistics

Date, time, room: Thursday, 18 June 2026, 14:00–15:30, SAE 1
Organisation: Organizer/chair: Ansu Chatterjee

Survey response in a census year: evidence from a natural experiment in Peru

Speaker: Angelo Cozzubo
Authors: Angelo Cozzubo

Abstract

Abstract of the talk to be announced.

Inference in small area estimation: a generative framework via the implicit bootstrap

Speaker: Lorenzo Mori
Authors: Lorenzo Mori, S. Guerrier, M. Karemera, S. Orso, M. R. Ferrante and M.P. Victoria-Feser

Abstract

This presentation introduces an innovative simulation-based framework for Small Area Estimation (SAE) that leverages the Implicit Bootstrap (IB) to construct highly accurate confidence intervals for LMMs and GLMMs. The IB method offers three main advantages: it provides valid inference regardless of the inconsistency or non-normality of the initial estimator; it yields second-order accurate, transformation-respecting intervals that correct for severe finite-sample bias; and it maintains procedural simplicity by relying on a simple initial estimator without requiring complex analytical adjustments. By integrating these properties, our approach ensures theoretically sound and reliable coverage even in challenging scenarios where traditional asymptotic methods fail.

Bayesian Small Area Estimation for Categorical Data: an Application to Official Statistics

Speaker: Fabrizio Solari
Authors: Fabrizio Solari

Abstract

In Italy, the new Population Census framework is based on a sample survey design and requires the production of estimates for highly disaggregated domains of the collected information. Modeling data at such a fine level of detail poses significant methodological challenges, particularly when the variables of interest have categorical nature. Several approaches have been proposed in the

literature to address this issue. We consider models for compositional data within a Bayesian framework. Transformations are often used to map compositional data onto the real space, allowing the use of standard statistical techniques. Conversely, when dealing with compositional count data, one can resort to a Multinomial distribution. Under this approach, a commonly used model is the Dirichlet–Multinomial model, which represents a first attempt to overcome the limitations of the multinomial distribution. However, it still offers limited flexibility in handling complex correlation patterns across components and domains. In this setting, we investigate extensions of the Dirichlet–Multinomial model, in particular based on work by Ascari, Migliorati and Ongaro (2025). We compare the aforementioned approach with standard methods based on data transformation through a case study based on official data from the Italian National Institute of Statistics.

IS13: Data challenges in small area estimation

Date, time, room: Thursday, 18 June 2026, 14:00–15:30, SAE 2
Organisation: Organizer/chair: Silvia Pacei

A two-part small area model that accounts for heaping to map smoke habits

Speaker: Aldo Gardini
Authors: Aldo Gardini, Lorenzo Mori

Abstract

Estimating health indicators for restricted sub-populations is a recurring challenge in epidemiology and public health. When survey data are used, Small Area Estimation (SAE) methods can improve precision by borrowing strength across domains. In many applications, however, outcomes are self-reported and affected by coarsening mechanisms, such as rounding and digit preference, that reduce data resolution and may bias inference. This paper addresses both issues by developing a Bayesian unit-level SAE framework for semi-continuous, coarsened responses. Motivated by the 2019 Italian European Health Interview Survey, we estimate smoking indicators for domains defined by the cross-classification of Italian regions and age groups, capturing both smoking prevalence and intensity. The model adopts a two-part structure: a logistic component for smoking prevalence and a flexible mixture of lognormal distributions for positive cigarette counts, coupled with an explicit model for coarsening and top-coding. Simulation studies show that ignoring coarsening can yield biased and unstable domain estimates with poor interval coverage, whereas the proposed model improves accuracy and achieves near-nominal coverage. The empirical application provides a detailed picture of smoking patterns across region–age domains, helping to characterize the dynamics of the phenomenon and inform targeted public health policies.

An evaluation of data driven tuning constants in the Huber functions for robust small area estimation

Speaker: Chiara Bocci
Authors: Chiara Bocci, Paul A. Smith

Abstract

Small area estimation methods are generally based on mixed effects models, which have assumptions of normal errors, but many types of data have skewed distributions, which means that this assumption is violated. Several robust approaches have been proposed in the literature to handle such skewed data, using robust models to accommodate outlying tail observations. In this contribution we examine a range of approaches, investigating their performance empirically on a dataset with known outcomes. In particular, we consider (a) M-regression based synthetic estimators, (b) M-quantile small area estimators, and (c) robust EBLUP estimators. These estimators use

Huber functions and conditional bias approaches, which reduce the impact of outlying observations by relying on a tuning constant. It has become standard to use the value 1.345, based on Huber (1964, Table IV) and Holland and Welsch (1977) which suggest that this value ensures 95% efficiency for normal residuals. But since the assumption of normal errors is violated, it seems likely that a value fitted from the data will be more appropriate and offer scope for improving the performance of the estimators. We examine the impact of choosing different values of the tuning constants, presenting evidence from repeated sampling simulations. This enables us to evaluate which approaches work well, and to offer some guidance on the best approach(es) to consider when working with data with a skewed distribution.

Development of small area estimation methods for labour force indicators at LLMA Level in ISTAT

Speaker: Andrea Fasulo
Authors: Andrea Fasulo, Michele D'Alò, Davide Di Cecco, Danila Filipponi

Abstract

Since 2004, Istat has produced labour force estimates for Labour Market Areas (LLMAs) using Small Area Estimation (SAE) techniques. LLMAs are defined every ten years on the basis of commuting matrices derived from daily commuting flows between municipalities, traditionally collected through the Population Census. Over time, the estimation methodology has evolved in response to both methodological advances and production requirements. The approach currently adopted, in use since 2014, relies on a unit-level model with spatial and temporal correlation structures and uses a limited set of demographic covariates, namely sex and age class. This contribution reviews recent and ongoing developments in SAE methods for key labour force indicators at the LLMA level, in light of both the 2025 redefinition of LLMAs and the substantial enrichment of available auxiliary information. Particular attention is given to the opportunities offered by the Integrated System of Registers (SIR), available since 2018, which enables the use of population and labour register data, including income-related variables, as auxiliary information. In addition, the Permanent Population Census, introduced in 2018, provides further census-based information on professional status at the municipal level. We present alternative modelling strategies at both area and unit levels and illustrate how the integration of survey, administrative, and census sources can improve the coherence and temporal stability of official LLMA labour force estimates.

IS14: Overcoming data deficits: innovative inference under missingness and bias

Date, time, room: Thursday, 18 June 2026, 14:00–15:30, SAE 3
Organisation: Organizer: Zhengyuan Zhu; Chair: Caleb Leedy

Calibrated active learning

Speaker: Zhonglei Wang
Authors: Zhonglei Wang

Abstract

Sample selection bias fundamentally limits the validity of active learning when the objective is population-level inference rather than prediction. We study population-level mean estimation in a setting where labeling is costly and available data are subject to selection bias. We propose a calibrated active learning framework that integrates empirical likelihood with doubly robust estimation. Consistency and asymptotic efficiency are established under both known and estimated population moments, and simulation and real-data experiments demonstrate improved estimation accuracy and more reliable uncertainty quantification compared to standard active learning and uniform sampling baselines under fixed labeling budgets.

An imputation based approach to small area estimation for nonlinear models in the presence of partial auxiliary information

Speaker: Emily Berg
Authors: Emily Berg

Abstract

Small area estimation uses auxiliary information and model assumptions to obtain predictors for domains where standard survey estimators are judged to be unstable. Many unit-level nonlinear small area models require the unit-level values of covariates for every element of the entire population. This condition is difficult to satisfy in practical situations. In a more realistic situation, the population means of the covariates are available at the area level, and the individual covariate values are known for sampled elements. We develop a small area procedure for this more common data structure. The key idea of our approach is to define imputed covariate values such that the population means of the imputed covariates equal the known values. We then combine mean square error (MSE) estimators developed for small area models with standard procedures for multiple imputation to develop an MSE estimator that reflects the imputation variance. Our approach is applicable to a wide range of unit-level nonlinear models. We validate the proposed method through model-based simulations using two common nonlinear models – specifically, the logistic and lognormal mixed models. We then apply the proposed methods to conduct a unique analysis of a seminal data set on corn area in Iowa counties.

IS15: Producing model-based estimates for official statistics

Date, time, room: Thursday, 18 June 2026, 16:00–17:30, SAE 1
Organisation: Organizer/chair: Julie Gershunskaya

Efficient estimation of response propensity to nonprobability survey partially linked to a probability sample from the same population

Speaker: Vladislav Beresovsky
Authors: Vladislav Beresovsky, Julie Gershunskaya, Terrance Savitsky

Abstract

Lowering response rates and raising costs of traditional probability-based surveys motivate increased interest in using nonprobability data sources, such as web surveys and administrative records, to produce estimates of target population quantities. Methods have been developed to account for a selection bias associated with such "convenience" nonprobability-based data. We consider estimation of response propensity to nonprobability dataset by combining it with a probability "reference" sample obtained from the same target population and maximizing pseudo-Bernoulli likelihood for the observed sample indicator. We compare robustness and efficiency of our proposed method with a commonly used pseudo-likelihood approach by Chen, Li and Wu (2020) and establish criteria when maximizing a proposed pseudo-likelihood converges to efficient maximum likelihood estimation. We note that the proposed framework allows utilization of probabilistic data linkage methods for improved efficiency of the estimates.

Privacy amplification for synthetic data using range restriction

Speaker: Terrance Savitsky
Authors: Terrance Savitsky, Monika Hu, Matthew R. Williams

Abstract

We introduce a new, range restricted formal data privacy mechanism that conditions on data owner beliefs about sensitive data ranges. The range restricted standard protects only a subset (or ball) of data values that exclude ranges (or balls) already known to a putative intruder. The new privacy standard is designed for the risk-weighted pseudo posterior mechanism used to generate synthetic data under a formally private guarantee. The pseudo posterior mechanism (PPM) designed under an asymptotic differential privacy (aDP) standard is formulated by selectively downweighting each likelihood contribution proportionally to its disclosure risk. The PPM is adapted to the range restricted privacy guarantee by expressing intruder knowledge as a probability, λ, that a data value drawn from the underlying generating distribution lies outside the ball or subset of values that are not known by the intruder. The portion of each datum likelihood contribution deemed sensitive is $(1 - \lambda) \leq 1$, which reduces the sensitivity of the range restricted mechanism as compared to aDP. A datum-indexed risk-based weight, $\alpha \in [0, 1]$ is applied solely to

this portion of the datum deemed sensitive. We compare privacy and utility properties between the aDP and range restricted privacy under the PPM.

Echo state network forecast model for preliminary estimation of the chained CPI-U

Speaker: Kate Eckerle
Authors: Kate Eckerle, Erin Boon, Daniell Toth

Abstract

Calculation of the Chained CPI-U requires monthly item-area expenditure shares from the same time-period as item-area price index relatives. Yet, cell-level expenditure data only become available four quarters after the price data. In the interim, the BLS issues a preliminary estimate of the index using a Constant Elasticity of Substitution model. We propose an alternative method for preliminary estimation that instead retains the Tornqvist formula for aggregation and forecasts the missing item-area expenditure data using a set of hierarchical Echo State Networks (ESNs), a class of Recurrent Neural Networks in which reservoir and input couplings are randomized. ESNs are flexible, nonlinear, hidden variable models that can predict series with complex temporal dynamics after a relatively simple training process. We develop an iterative procedure to forecast a vector of item expenditures for a given area based on its past expenditure data as well as past and concurrent price data. Additionally, we include the option to supplement the ESN neuron states with discrete Fourier modes at the seasonal frequencies to improve prediction among items with strong seasonal components.

IS16: Poverty mapping through small area estimation: experiences from collaboration between national statistical offices and the World Bank

Date, time, room: Thursday, 18 June 2026, 16:00–17:30, SAE 2
Organisation: Organizer: Marcin Szymkowiak; Chair: Isabel Molina

Mapping poverty at the level of subregions in Poland using multivariate Fay-Herriot models

Speaker: Marcin Szymkowiak
Authors: Marcin Szymkowiak, Tomasz Józefowski, Kamil Wilak

Abstract

The European Survey on Income and Living Conditions (EU-SILC) is the basic source of information published by Statistics Poland about the poverty indicators both for the country as a whole and at the regional level. This also applies to other countries facing a growing demand for good poverty maps. In order to follow an appropriate social strategy, which is consistent with the guidelines of the cohesion policy, one needs to measure poverty and provide information about this phenomenon at lower levels of spatial aggregation. In this context poverty maps are used to support decisions concerning important political issues, such as the allocation of development funds by governments, National Ministries of Infrastructure and Development or international organizations, such as the World Bank. Those decisions should be based on the most accurate poverty indicators, estimates or figures and should be delivered at the lowest level of spatial aggregation. However, given the small sample size in relevant cross classifications of the EU-SILC survey, it is necessary to use the latest techniques of indirect estimation and draw on alternative data sources to estimate the parameters of interest at low levels of spatial aggregation with acceptable precision. Since the EU-SILC survey does not cover adequately all the specific areas or population subgroups, the required information can only be obtained using small area estimation techniques based on the idea of ''borrowing strength". In Poland, for instance, EU-SILC data are only sufficient to publish poverty indicators at the level of the whole country and at the regional level (NUTS 1). Owing to small sample sizes and low precision of estimation, adequate estimates at lower level of spatial aggregation cannot be delivered. Given the growing demand for information about poverty indicators at lower levels of spatial aggregation (for instance NUTS 3), there is a pressing need to take advantage of appropriate small area estimation techniques and data from different statistical sources (EU-SILC, census or administrative registers). The main aim of the presentation is to present selected results of a study which was taken by Statistics Poland and the World Bank involving the use of multivariate Fay-Herriot model for the purpose of poverty mapping using data coming from different statistical sources at lower level of spatial aggregation – NUTS 3 i.e. at the level which has not been published officially in Poland to date.

Comparable small area estimates over time: Application with SILC data from Bulgaria

Speaker: Eva Romero Ramos
Authors: Eva Romero Ramos

Abstract

This study presents comparable small-area estimates of poverty rates for Local Administrative Units at level 1 (LAU1) in Bulgaria for the years 2021 and 2022. The estimates are derived using two distinct area-level models: a multivariate model that accounts for temporal correlation, and a spatio-temporal model that considers both temporal and spatial correlation. The results are compared with those obtained from cross-sectional area-level models applied independently to each year, as well as with direct estimates that do not rely on parametric assumptions. Using parametric bootstrap procedures specifically developed for each model, we estimate the mean squared errors of the small-area estimators for each year, along with the mean cross-product errors between the two time points. These error metrics enable us to rigorously assess whether the observed changes in poverty across the two years within a given area are statistically significant.

County-Level Maps of Income and Energy Poverty in Romania, 2020–2024: A Multivariate Fay-Herriot Approach

Speaker: Luciano Perfetti Villa
Authors: Luciano Perfetti Villa, Monica Robayo-Abril

Abstract

Reliable subnational poverty statistics are essential for targeting social policy, yet for Romania's 42 counties (NUTS-3 județe), direct survey estimates are too imprecise to guide county-level decisions. We present an updated set of small-area poverty maps covering two domains of deprivation, using area-level models. Income poverty is estimated from EU-SILC using the at-risk-of-poverty rate (AROP), an anchored AROP that fixes the real poverty line at its 2018 value, and mean equivalised disposable income; energy poverty is estimated from the Household Budget Survey using four affordability indicators (M/2, 2M, P10, and LIHC). For each domain, survey direct estimates are combined with census, administrative and geospatial covariates and fitted with univariate per-year (UFH) and multivariate (MFH) Fay-Herriot models that borrow strength across five survey waves. The model-based estimates reduce mean squared error relative to direct estimates and remain consistent with published NUTS-2 benchmarks. Across 2020-2024, the income maps reveal a persistent high-poverty cluster of counties (Vaslui, Olt, Botoșani, Mehedinți, Teleorman) along the country's eastern and southern margins. We discuss the strengths and limitations of multivariate and univariate specifications, and the value of these maps for targeting EU Structural Funds.

Contributed session abstracts

Contributed Session 1

Date, time, room: Tuesday, 16 June 2026, 11:00–12:30, SAE 3
Chair: Serena Arima

Small area estimation with covariate measurement error: unit-level empirical best prediction under a finite population framework

Speaker: Ika Yuni Wulansari
Authors: Ika Yuni Wulansari, Stephen Woodcock, James Brown

Abstract

Small Area Estimation (SAE) methods typically assume that auxiliary covariates are measured without error. In official statistics, however, auxiliary information is often obtained from surveys and is subject to measurement error across all areas, often with heterogeneous magnitudes. Ignoring such errors can lead to biased predictors and underestimation of uncertainty under complex sampling designs. This paper develops a unit-level Empirical Best Prediction (EBP) approach under functional measurement error within a fixed finite population framework. To reflect the context of official statistics, we construct pseudo-populations using a design-based generation mechanism that mimics realistic survey sampling structures. Measurement error arises naturally through the sampling process and affects all areas with varying magnitudes, consistent with practical survey settings. We consider a bivariate auxiliary structure consisting of one error-free and one error-prone covariate, allowing us to examine how incorporating an additional error-prone covariate of varying strength affects prediction accuracy and variance reduction. To quantify uncertainty, we adopt a design-consistent bootstrap procedure for estimating the Mean Squared Prediction Error (MSPE), explicitly accounting for both sampling variability and measurement error. An extensive simulation study evaluates the finite population performance of the proposed approach. We compare bias and MSPE across scenarios with and without measurement error correction, as well as against area-level EBLUP approaches. The results indicate that ignoring measurement error leads to bias and underestimation of uncertainty, whereas the proposed correction improves inferential validity and provides reliable uncertainty quantification for official statistics applications.

Bayesian estimation of income distributions and inequality across subpopulations via multilevel lognormal mixtures

Speaker: Yuki Kawakubo
Authors: Yuki Kawakubo, Kazuhiko Kakamu, Genya Kobayashi

Abstract

We propose a Bayesian method for estimating income distributions and inequality measures for arbitrary subpopulations defined by cross-classified categorical variables such as age group and region. The method is motivated by unit-level small area estimation of general parameters. When auxiliary variables are essentially categorical and population cell shares are known, the

usual finite-population prediction step can be viewed as constructing a mixture over cell-specific distributions. This perspective leads to a flexible framework for distributional inference that is closely related to poststratification while remaining naturally connected to small area estimation.

For each cell, we assume a lognormal distribution and reparameterize it in terms of a mean parameter and an inequality parameter. These parameters are modeled hierarchically through multilevel effects, allowing information to be borrowed across sparse cells, including cells with very small sample sizes. Because familiar inequality measures such as the Gini coefficient and Theil index are functions of the lognormal variance parameter, the proposed model directly supports estimation of both income distributions and inequality. Posterior inference is carried out by Gibbs sampling, and grouped income data are handled through latent log-income augmentation.

A key advantage of the mixture formulation is that it yields subpopulation estimates that are coherent with inequality decomposition. We illustrate the method using household income data in Japan cross-classified by prefecture and age group, where many cells are sparse and some are empty. The proposed approach provides a practical extension of small area estimation from scalar summaries to distributional and inequality-related parameters.

Contributed Session 2

Date, time, room: Tuesday, 16 June 2026, 14:00–15:30, SAE 2
Chair: María José Lombardía

Estimating poverty incidence, gap, and severity in South Africa using a unit-level GLMM approach

Speaker: Yegnanew A. Shiferaw
Authors: Yegnanew A. Shiferaw

Abstract

"Reliable subnational poverty statistics are essential for developing evidence-based policies and monitoring progress toward national and global development goals. However, direct survey estimates of poverty indicators are often unreliable at smaller geographic levels due to small sample sizes. This study employs unit-level small-area estimation (SAE) methods to produce district-level estimates of poverty measures defined by the Foster-Greer-Thorbecke (FGT) class of indices: poverty incidence (FGT0), poverty gap (FGT1), and poverty severity (FGT2) in South Africa. The analysis uses data from the 2022/23 Income and Expenditure Survey, along with auxiliary data from the 2022 Population Census. A unit-level generalized linear mixed model (GLMM) with a logit link function is used to assess individual poverty status. This model incorporates demographic and socioeconomic variables, along with district-specific random effects. Subsequently, the empirical best prediction (EBP) method is applied to produce model-based estimates of FGT0, FGT1, and FGT2 for all districts. To compute the log-likelihood function for GLMMs, we use adaptive Gauss-Hermite quadrature. The accuracy of the estimates is further evaluated using a parametric bootstrap method to derive mean-squared error estimates and their corresponding coefficients of variation. The results indicate that EBPs significantly enhance precision compared to direct survey estimates, particularly in districts with small sample sizes. Additionally, there is notable spatial variability in poverty levels across districts, revealing substantial differences in both the incidence and depth of poverty. The EBP estimates based on the upper-bound poverty line indicate varying levels of poverty across different districts. For instance, the FGT0 value for the DC27 district in KwaZulu-Natal is 0.588, whereas it is 0.271 in the DC5 district of the Western Cape. Specifically, the EBP estimates for the DC27 district are as follows: FGT0 is 0.588 (95% CI: 0.575 to 0.601), FGT1 is 0.242 (95% CI: 0.239 to 0.245), and FGT2 is 0.129 (95% CI: 0.127 to 0.131). Diagnostic assessments—including evaluations of district-level residuals, random-effects analyses, ROC curves, and simulation-based residual checks—confirm the robustness of the fitted models. Overall, the findings demonstrate that unit-level SAE provides a strong framework for producing reliable and policy-relevant poverty indicators at finer geographic scales, thereby supporting targeted poverty alleviation efforts and informed resource allocation in South Africa.

Small area estimation illustrated by its application to the cantonal poverty rate in Switzerland

Speaker: Jacques Saliba
Authors: Jacques Saliba, Anne Massiani, Daniel Kilchmann

Abstract

This illustration of the application of small area estimation methods focuses on a project from the Swiss Federal Statistical Office, which is still work in progress. Its goal is to obtain reliable cantonal absolute poverty rate estimations based on the Statistics on Income and Living Conditions (SILC). SILC is an annual survey that aims at studying poverty, social exclusion and living conditions of the Swiss population, using indicators that are comparable at the European level. In this illustration, we focus on the absolute poverty rate that corresponds to the proportion of people in the total population living in a household with a disposable income less than a certain threshold. This threshold is based on the social subsistence level that is defined in the guidelines of the Swiss Conference for Social Welfare (SKOS). Given that cantonal estimations obtained by classical methods can be very unstable, especially for cantons with a small sample size, we use in our case small area estimation methods in order to obtain more stable results. Since our variable of interest is a binary indicator of the absolute poverty, Generalized Linear Mixed Models (GLMM) are, in principle, better suited than Linear Mixed Models (LMM). However, our preliminary tests showed similar cantonal estimations for both models. This gives LMM models a slight edge, since they are lighter and faster in terms of computation time, which is an important advantage for the bootstrap method used for the variance estimation. Therefore, the tested models were based on a linear mixed model (LMM). After applying the corresponding EBLUP estimator, we eventually opted for the pseudo- EBLUP estimator that takes into account the extrapolation weights and has some interesting properties from the design-based perspective. Although SAE methods were initially developed under a model-based approach, we have chosen to evaluate the performance of SAE estimates under a design-based approach. In conclusion, we will show the potential offered by SAE methods in the context of the SILC survey, as well as the limitations that have been observed.

Small area estimation of employment indicators under area-level Dirichlet mixed models

Speaker: Domingo Morales
Authors: Esteban Cabello, María Dolores Esteban, Tomas Hobza, Domingo Morales, Agustín Pérez

Abstract

We propose an area-level Dirichlet mixed model for predicting compositional indicators in small areas. The model focuses on the direct estimators of domain-specific category proportions from a classification variable as the primary target parameters. After selecting and fitting the model to the data, predictions for small-area proportions, totals, and rates are derived, and their mean squared errors are estimated using a parametric bootstrap approach. A series of simulation studies is conducted to evaluate the performance of the fitting algorithm, the small-area predictors,

and the bootstrap method. The methodology is illustrated with an application to data from the Spanish Labour Force Survey for the fourth quarter of 2022, aiming to estimate provincial proportions of employed, unemployed, and inactive individuals, as well as unemployment rates, disaggregated by sex and age group.

Contributed Session 3

Date, time, room: Tuesday, 16 June 2026, 14:00–15:30, SAE 3
Chair: Christopher K. Wikle

Incorporating industry dependence into small domain estimation modeling for employment

Speaker: Julie Gershunskaya
Authors: Julie Gershunskaya, Terrance Savitsky

Abstract

Small domain estimation models for economic data, such as price changes, express spatially-indexed correlation that is included in small domain estimation models to reduce estimation error for small sample domains. By contrast, the correlation between employment trends is not well-predicted by spatial indexing because underlying economic bases are often quite different for relatively adjacent spatial areas; for example, a rural or suburban county may lie next to an urban county that expresses large differences in industry concentrations. The dependence structure among employment trends may best be captured by industry. Our proposed method uses the North American Industry Code system (NAICS) to build a network graph based on code nesting and then, in turn, converts this graph to an adjacency matrix. The adjacency matrix is used to formulate a conditional autoregressive (CA) prior distribution for industry-indexed random effects in a Bayesian hierarchical model. The CA prior is a typical framework used for spatial dependence encoded through a spatial adjacency matrix that we have repurposed to model industry dependence. We demonstrate the improved performance of a model that accounts for industry dependence on industry and area indexed total employment from the Current Employment Statistics survey administered by the U.S. Bureau of Labor Statistics.

Calibrating routine HIV testing data for subnational surveillance: a non-probability sampling framework applied across four African countries

Speaker: Li-Chun Zhang
Authors: Adrien Allorant, Li-Chun Zhang

Abstract

Background. Subnational HIV prevalence estimation in sub-Saharan Africa relies on population-based household surveys such as PHIA and DHS. Routine health information system (RHIS) indicators may serve as covariates in small area estimation models, but the survey remains the primary data source. This dependence is problematic: the surveys are expensive, infrequent, and vulnerable to funding disruptions. At the same time, every country operates an RHIS that records every HIV test at every health facility, at the district level, every month; a large nonprobability sample whose potential for direct prevalence estimation is limited by self-selection into testing.

Objective. We develop a framework that reverses the roles of these two data sources: the RHIS serves as the primary signal for subnational surveillance, and a probability survey calibrates the selection mechanism. The approach can be understood as an instance of statistical calibration, in analogy to scientific calibration where a precise reference instrument is used to adjust readings from an imprecise one. Here, the reference (survey) is not precise but unbiased, and the instrument being calibrated (RHIS) is not imprecise but systematically biased; this extends the classical calibration problem from a precision adjustment to a bias correction. Estimand. Under the assumption that previously diagnosed individuals rarely re-enter routine testing at the facility level, supported by the near-complete overlap between diagnosis and treatment initiation observed in survey data, RHIS test positivity estimates the prevalence of HIV among the undiagnosed population, scaled by a relative test propensity $\lambda = \pi_1/\pi_0$. The selection bias is multiplicative in the odds and additive on the log-odds scale. The target, undiagnosed prevalence, is directly estimable from surveys that collect both biomarker HIV status and awareness of diagnosis. Model. We specify a measurement error model in which district-level undiagnosed prevalence is a latent variable observed with noise by both data sources. Province-level fixed effects parameterise the selection bias. The latent prevalence surface is estimated via joint likelihood with Laplace-approximated integration, yielding district-level small area estimates as precision-weighted combinations of the bias-corrected RHIS signal and the survey signal. Application. We apply the framework to four countries in eastern and southern Africa: Mozambique, Lesotho, Zimbabwe, and Malawi. In each country, we pair two data sources. The first is the national RHIS, which records every HIV test conducted at health facilities, aggregated to the district level. The second is the most recent Population-based HIV Impact Assessment (PHIA) survey, a nationally representative household survey in which a random sample of adults are tested for HIV by blood draw and asked whether they already knew their status. These surveys, conducted between 2020 and 2021 depending on the country, provide the biomarker-confirmed estimates of undiagnosed prevalence that serve as our calibration anchor. The four countries span 10 to 159 districts, 5 to 10 provinces, and a combined total of approximately 8 million facility-based HIV tests in the matched survey years. Results. Selection bias is substantial in all four countries, with national-level odds ratios ranging from 2.0 (Lesotho) to 3.5 (Zimbabwe), and province-level odds ratios varying by a factor of 1.5 to 3 within countries. We test whether province-level calibration is sufficient by adding district-level random effects to the selection model, using geo-referenced survey clusters joined to district boundaries. In Mozambique and Malawi, district effects collapse to near-zero; in Zimbabwe, district-level heterogeneity is statistically detectable but calibrated prevalence estimates agree within one percentage point in all but one of 64 districts. A variance decomposition shows that 96–99% of district-level prediction uncertainty is attributable to survey calibration precision, not RHIS sampling noise. Implications. The RHIS provides sufficient volume for district-level resolution; the constraint is survey precision at the province level. These results point toward an operational cycle in which the RHIS provides continuous district-level estimates and a small audit survey, targeted at provinces where calibration uncertainty is greatest, periodically refreshes the correction. The temporal stability of the selection parameter between survey rounds remains untested and is a key assumption for this model. Subject to that assumption, the approach could reduce the current dependence on large, infrequent household surveys for subnational HIV monitoring.

Bayesian small area estimation of continuous survey outcomes: methodology and application

Speaker: Ioannis B. Nikolaidis
Authors: Ioannis B. Nikolaidis, Stefanos Giakoumatos

Abstract

Reliable socio-economic indicators at fine geographical levels are essential for evidence-based regional policy, yet traditional design-based survey estimators often perform poorly in small domains due to limited sample sizes and high sampling variability (Rao, Molina, 2015). This study develops a Bayesian small area estimation (SAE) framework for continuous and highly skewed survey outcomes and applies it to the estimation of household expenditure across Greek prefectures using microdata from the 2021 Household Budget Survey conducted by the Hellenic Statistical Authority (ELSTAT, 2021).

The proposed approach is based on a Bayesian unit-level hierarchical model within the generalized linear mixed modeling framework, extending the classical Battese–Harter–Fuller model (Battese, Harter, Fuller, 1988). A Gamma distribution with a log-link function is employed to accommodate strictly positive and right-skewed expenditure data (McCullagh, Nelder, 1989). The conditional mean is specified as a log-linear function of household-level socio-economic and demographic covariates, including household size, income group, employment status, education level, age structure, and housing characteristics, together with prefecture-level random intercepts capturing unobserved regional heterogeneity.

To increase model flexibility, the dispersion parameter of the Gamma distribution is also modeled as a function of selected covariates, allowing for covariate-dependent heteroskedasticity through a distributional regression specification. Weakly informative priors are adopted following modern Bayesian hierarchical modeling principles (Gelman et al., 2013). Regression coefficients are assigned normal priors centered at zero with relatively large variance, while area-level random effects follow a normal distribution with an exponential prior on their standard deviation to encourage shrinkage across prefectures.

Posterior inference is obtained via Markov Chain Monte Carlo using Stan through the brms framework (Bürkner, 2017). Empirical results show substantial reductions in coefficients of variation relative to direct survey estimators, particularly in prefectures with small samples. The hierarchical Bayesian model stabilizes estimates through partial pooling while preserving consistency with direct estimates in well-sampled areas, demonstrating the value of flexible Bayesian SAE methods for producing reliable regional expenditure statistics.

Contributed Session 4

Date, time, room: Wednesday, 17 June 2026, 14:00–15:30, SAE 3
Chair: Jan van den Brakel

Small area estimation from a large nonprobability sample with varying domain coverage

Speaker: Andrius Čiginas
Authors: Andrius Čiginas

Abstract

We consider small area estimation when the study variable is observed only in a large nonprobability sample, and the amount of information contributed by that sample differs markedly across areas. Coverage is high in some domains but limited in others, so the available local evidence is not equally strong everywhere. This motivates a cautious model-based strategy that combines information from the nonprobability sample with auxiliary data while allowing domains to contribute differently to the resulting estimates. An empirical illustration is drawn from the accommodation sector. The presentation focuses on the problem formulation, the main modeling ideas, and diagnostic considerations for assessing when the resulting estimates may be practically useful.

Meeting the needs of Members of Parliament for small area statistics on welfare benefits

Speaker: Isabel Buchanan
Authors: Rachael Harker, Isabel Buchanan

Abstract

Members of Parliament in the UK are increasingly interested in data relating to their specific parliamentary constituencies and smaller areas within these. Although administrative data on welfare benefits is available for small areas, the raw data counts are not meaningful in considering whether certain areas contain high or low levels of claimants.

We have developed a range of constituency and small area data dashboards, presented in Power BI or R shiny, which detail benefit claims as a proportion of the working age or other population groups.

The small area data used is for Middle layer Super Output Areas (MSOAs) in England and Wales and Intermediate Data Zones (IDZ) in Scotland data. These areas comprise between 2,000 and 6,000 households. The areas are based on data from the 2021 census for England and Wales and the 2022 census for Scotland.

Our talk will outline the methods we used to match population estimates to benefits data at

these levels. Matching is carried out using R scripts and in most cases is a straightforward task. However, we will also discuss strategies used to overcome difficulties of outdated geographical boundaries being used in some benefit statistics publications.

We will also discuss how we renamed the standard MSOA codes to make them more meaningful to users of our data dashboards. MSOAs were not given recognisable names when designed – instead they have standard codes (e.g. 'E02006827') and schematic names relating to the local authorities where they are located (e.g. 'Ashfield 004').

To address this, we have designed a set of recognisable names for MSOAs based on the towns, villages and neighbourhoods that they cover. We published draft names on an interactive map website to invite feedback and suggestions on proposed names. This feedback was used to finalise a set of recognisable names to make MSOA data easier to interpret and present. A demonstration of some of our main welfare benefit dashboards will also be included.

Publishing data on named people

Speaker: Elliot Bridges
Authors: Elliot Bridges

Abstract

The Library of the UK House of Commons performs impartial research on behalf of politicians. While our research largely focuses on policy development, it often involves looking at the activities of specific politicians. For example, we may be asked to investigate how much they have spoken about specific issues, or to explore their earnings from private companies.

By necessity, individuals can be identified within the outputs of such research. There are also particular difficulties around impartiality. In this presentation I will discuss different examples of how releasing data on specific individuals is approached in the Commons Library.

I will describe how we approach data on politicians' gender, ethnic background, earnings from other private companies, and other similar information. Some aspects of this information we choose to release publicly, while other aspects we intentionally do not release. Decisions around what to release are typically based on a combination of past experience, views of data subjects, the sensitivity of the data, and how much information is already in the public domain.

This talk will be of interest to researchers working with particularly small area data, where the risk of identifying individuals is high. We will discuss how we decide what information to release, what to withhold, and how this has evolved over time.

Contributed Session 5

Date, time, room: Wednesday, 17 June 2026, 16:00–17:30, SAE 3
Chair: María Bugallo

Estimation of the average number of trips per household at the neighborhood level in Bogotá using data integration methods and small area estimation

Speaker: Nicolás Ramírez Vargas
Authors: Nicolás Ramírez Vargas, Felipe Ortiz, Cristian Fernando Tellez Piñerez

Abstract

The estimation of mobility indicators at disaggregated spatial levels poses a significant statistical challenge due to limited sample sizes and the high variability in the population dynamics of individuals who reside in, work in, or visit the city of Bogotá.

This study addresses the estimation of the average number of trips per household at the neighborhood level in Bogotá using Small Area Estimation (SAE) techniques, which improve the precision of estimates in domains with limited information.

Within this framework, area-level mixed models are implemented, particularly the Fay-Herriot model, incorporating relevant covariates related to urban environment characteristics, accessibility, and socioeconomic composition. Additionally, model extensions that account for spatial correlation structures are evaluated in order to capture dependencies across neighboring areas.

The performance of the estimators is assessed in terms of bias, variance, and mean squared error, comparing the results with direct estimators and secondary data reported by official sources. The findings show substantial improvements in the stability and precision of neighborhood-level estimates, enabling the identification of differentiated spatial patterns in trip generation per household across the city.

Overall, this study highlights the potential of integrating data sources and combining them with SAE methods to produce highly disaggregated and precise estimates, providing robust inputs for the design of public policies aimed at optimizing and adapting public transportation systems in response to the city's demographic and population dynamics.

Estimation of design mean squared error under a unit level model in small area estimation

Speaker: Felipe Ortiz
Authors: Felipe Ortiz, Isabel Molina

Abstract

An estimator of the design mean squared error (dMSE) for a general small-area parameter is proposed under a unit-level model. Its performance is compared with that of other approaches

available in the literature through simulation studies. These approaches include resampling-based estimators such as the parametric bootstrap, nonparametric bootstrap, mixed bootstrap, and parametric design bootstrap, as well as analytical estimators in the case of area means. Among the latter, we consider composite estimators and the conditional model mean squared error estimator (cmmse), which relies on the conditional distribution given the observed values. In our simulation experiments, the proposed method outperforms the existing bootstrap approaches and shows performance comparable to that of cmmse for area means. In the context of dMSE estimation for poverty indicators, the proposed method performs well and represents a novel alternative for measuring uncertainty. The approach is straightforward to implement and can be readily adopted by national statistical institutes and researchers to assess the quality of estimates produced in small area estimation settings under unit-level models and complex sampling designs. An application based on data from the Survey on Essential Characteristics of Population and Housing (ECEPOV) for Spain illustrates the proposed procedures.

Pseudo empirical best prediction of multiple characteristics in small areas

Speaker: William Acero
Authors: William Acero, Isabel Molina, Domingo Morales

Abstract

Small area estimators that ignore the sampling design lack design consistency when the sampling mechanism is complex and may be severely biased under informative designs. Existing procedures that account for the survey weights under unit-level models typically focus on a single response variable. This paper addresses the estimation of area means for several dependent target variables under a multivariate nested error regression (MNER) model. We propose a multivariate pseudo–empirical best linear unbiased predictor that accounts for the sampling mechanism. Moreover, by aggregating the MNER model, we derive a unified predictor that can be obtained from either unit-level or area-level data. Bootstrap procedures are proposed to estimate the mean squared errors (MSEs) of the proposed predictors. Simulation experiments are conducted to examine the properties of the proposed small area estimators and the MSE estimators. Finally, an application with housing data illustrates the proposed methods.

Contributed Session 6

Date, time, room: Thursday, 18 June 2026, 9:00–10:30, SAE 3
Chair: Emily Berg

Algebraic dimensional reduction for latent-class models applied to record linkage

Speaker: Yves Thibaudeau
Authors: Yves Thibaudeau, Daniel Weinberg

Abstract

The Fellegi-Sunter methodology for record-linkage has stood the test of time and remains popular for scoring and classifying links between records. Its simplicity and high-scalability makes it the choice methodology in large-production open-source implementations of record linkage, such as fastLink, Splink and BigMatch. At the same time the estimation of the parameters of the associated latent-class model can be problematic. Bayesian methods provide solutions but can be computationally prohibitive. Another possible avenue is exploiting the likelihood-based approach of Fienberg and Rinaldo (Maximum likelihood estimation in log-linear models. Ann. Stats, 40: 996–1023, 2012). The paper recasts the Extended MLE of these authors in the context of record linkage and expands on another of their proposal to posit a dimensionally reduced likelihood parameterized through a toric variety. This method restores strict local concavity of the likelihood and ensures that a local maximum can be identified. This approach "borrows strength" across dimensions and leads to the estimation of a submodel. As such, we avoid having to recourse to specifying a prior distribution. The final estimates can readily serve as input to available computationally scalable record-linkage engines.

Domain estimation from weighted nonprobability samples

Speaker: An-Chiao Liu
Authors: An-Chiao Liu, Sander Scholtus, Katrijn Van Deun, Ton de Waal

Abstract

When inferring population characteristics from a nonprobability sample, it is crucial to correct the possible selection bias therein by, for example, pseudo-weighting. Many selection bias correction methods focus on estimating the population means or totals of the target variable. However, often the quantities of subpopulation, or domain, are also of interest. It is unclear whether pseudo-weights are suitable for domain estimation or may cause implications. Since the weights unavoidably introduce variation and possibly even bias in the downstream estimation. To address this issue, modeling on the domain level may be an option to increase the resolution of the population. In this paper, we evaluate two promising domain estimation methods on weighted nonprobability samples. The first one is iterative proportional fitting (IPF), where the known or estimated margins are considered in the domain estimation, so that the marginal values may be fixed when improving the domain estimates. The other is a hierarchical Bayesian model, in which

the pseudo-weights are included in the modeling process, and it enjoys the flexibility of modeling when different types of information are available. We evaluate a range of modeling options for the two methods, and compare them in a simulation study that varies the variable generation and sample sizes of the nonprobability sample. We also evaluate the methods with resampled real data sets to mimic the scenario where the relation between variables and the inclusion mechanism of the nonprobability samples are unknown to the researchers.

From the evaluation, we found that while no single method is the best across all scenarios and estimands, applying IPF to the unweighted table and the hierarchical Bayesian model improves the domain estimation in most cases. If both marginal and domain estimates are of interest, the estimated population total or mean should be considered in the domain modeling process. The result may manifest the future development of the domain estimation method from weighted nonprobability samples.

On the use of geospatial data in small area estimation: data integration, model specification and intercensal updating

Speaker: Luciano Perfetti Villa
Authors: Luciano Perfetti Villa, Nikos Tzavidis, Ángela Luna Hernández, Vasilis Chasiotis, Adrien Allorant

Abstract

Advances in machine (statistical) learning, alongside the availability of large datasets from alternative sources, have enabled the production of small-area-type estimates globally and at refined spatial scales. Private firms and research organisations are publishing small area-type products using machine learning methods and a range of data sources, including remote sensing data. For example, Meta has developed methodologies to produce global estimates of average wealth at a 2.4 km^2 resolution. Assessing the quality of these granular estimates, particularly their uncertainty, using approaches comparable to those employed in survey and official statistics, is essential for their reliable use, yet remains limited. Greater engagement with the extensive literature on small area estimation developed over the past three decades could further support methodological integration and cross-fertilisation.

New and rediscovered algorithmic tools and data offer opportunities sufficient to support an exciting period of research, but also pose significant challenges. In this talk, we focus on the theoretical and applied aspects of using geospatial data in small-area estimation. These include (a) the integration of survey and geospatial data and the choice of spatial scale for processing geospatial data, (b) the use of geospatial data for intercensal updating, and (c) the adoption of alternative measures of wealth and poverty, such as the relative-wealth index.

The methods are illustrated using real data from ongoing collaborative work with the World Bank in Mozambique and the National Office for Statistics in the UK, as well as the EUSILC data for Greece. The findings highlight the potential of geospatial data for poverty mapping and small-area estimation, but also illustrate situations in which it can fail. Nonetheless, leveraging these data requires additional quality assessments and underscores the importance of a comprehensive modelling workflow that includes quantifiable uncertainty. Additionally, alternative poverty indicators, such as the relative wealth index, can aid informal poverty analyses; however, their integration into official poverty mapping remains limited.

Speaker Index

Practical conference information

Venue:	Faculty of Sociology and Social Work, University of Bucharest, building situated behind the Rectorate of University of Bucharest
Address:	*90 Panduri Road, Sector 5, 050663, Bucharest*
Website:	https://sae2026.faa.ro/
Contact:	smallareaestimation2026@gmail.com

Conference website

LinkedIn page

SAE 2026 Conference

Small Area Estimation, Survey and Data Science 2026

15–19 June 2026, Bucharest, Romania

https://sae2026.faa.ro/