ITACOSM2025-IASS Satellite Conference

Shaping the future of survey statistics
in the data-driven era

Rimini, 30 June & Bologna, 1-4 July 2025

BOOK OF ABSTRACTS

INDEX

Keynote Lectures

KEY1 – Prediction inference with no or limited probability survey data, 4
KEY2 – The role of probability samples in the 21st century, 5
KEY3 – Recent developments in sample coordination , 6
KEY4 – Indirect questioning in sensitive surveys, 7
KEY5 – Leveraging technology for data collection in social research and official statistics, 8
KEY6 – Small Area Estimation in the era of machine learning and new data, 9

Invited Sessions

INV1 – On the use of machine learning methods for survey data, 11
INV2 – Correcting for selection bias in nonprobability samples, 14
INV3 – Imputation strategies for social surveys, 17
INV4 – Multiple frame surveys in the context of new data sources, 20
INV5 – Advancing the AMELIA Data Platform: the role of small area estimation (GRINS SESSION), 23
INV6 – Spatial survey sampling with applications on environment, 27
INV7 – The Baltic-Nordic-Ukrainian network session, 30
INV8 – Modeling subpopulations with linked and self-selected data sources, 33
INV9 – Uncertainty quantification in multisource statistics, 36
INV10 – Development and implementation of statistical matching techniques (GRINS SESSION), 39
INV11 – y-SIS: Use and integration of innovative data sources, 42
INV12 – Bridging data integration and artificial intelligence: Innovations for statistical modelling, 45
INV13 – Differential privacy and synthetic survey data, 48
INV14 – Sampling schemes exploiting register data, 51
INV15 – Equality, diversity and inclusion in survey statistics and methodology, 54
INV16 – Inference from different data sources: Challenges in survey methodology, 57
INV17 – New forms of data collection, 60
INV18 – Advances in small area estimation, 63
INV19 – Combining diverse data sources for policy analysis: challenges and innovations, 66

Contributed Sessions

CON1 – Innovative methods and data integration for socio-economic measurement, 70
CON2 – Enhancing small area predictions and model diagnostics, 75
CON3 – The role of bootstrap in complex statistical inference, 78
CON4 – Privacy protection and disclosure control in statistical data and survey method, 81
CON5 – Innovations in small area estimation for social and economic indicators, 84
CON6 – New approaches to handling selection bias in survey research, 87
CON7 – Methodological innovations in sampling design and precision optimization, 90
CON8 – Advanced calibration and estimation methods in complex survey designs, 93
CON9 – Innovations in spatial and balanced sampling, 97
CON10 – Alternative sampling and data collection strategies in the digital age, 100
CON11 – Advanced weighting techniques using linked and administrative data in official statistics, 103
CON12 – Modern approaches to data collection, estimation, and dissemination, 106

Keynote Lectures

Prediction inference for finite population totals or means with no or limited probability survey data

***Beaumont, J. F.**, Statistics Canada*

For a number of years, National Statistical Offices in several countries have been seeking to reduce data collection costs and the burden on respondents, while increasing the use of alternative data sources and modern prediction techniques. At Statistics Canada, an idea that is currently being explored to address these challenges is to reduce the frequency of probability surveys and replace missing survey data with predictions. For example, a probability survey could be conducted every other year, instead of every year, and missing survey data in non-survey years could be replaced with predictions, provided that auxiliary data are available in the non-survey years along with training data. In this presentation, we will study the following two questions: 1) How, and what assumptions are needed, to obtain approximately unbiased predictors of finite population totals or means when no or limited probability survey data is available? and 2) How to estimate the quality (prediction variance) of these predictors?

We will consider two cases. In the first case, auxiliary data for a non-survey period are available in a probability sample, but no survey data is observed in that probability sample. This case leads to sample matching (Rivers, 2007) when the training sample is a subset of the population, and nearest-neighbour predictions are used to replace missing survey data in the probability sample. We will show that (approximately) unbiased predictors of finite population totals or means can be obtained using a linear working model, even under deviations from the linearity assumption. This surprising property requires each observation in the training sample to be suitably weighted so as to ensure a certain calibration constraint is satisfied. We will also present a simple estimator of the prediction variance. Then, we will discuss properties of nearest-neighbour predictions and, more generally, machine learning predictions, as alternatives to predictions from a linear working model. A few simulation results will be shown to illustrate the properties of predictors.

In the second case, auxiliary data are available for the entire population along with survey data in a small probability sample. The availability of survey data allows us to make valid inferences without relying on model assumptions. A new approach, called prediction-powered inference, has recently been proposed to handle this case when observations are independent and identically distributed. It consists in eliminating the bias of machine learning predictions by leveraging the observed survey data. Prediction-powered inference is essentially equivalent to the well-known model-assisted approach under simple random sampling.

The role of probability samples in the 21st century

Elliott, M., *University of Michigan*

In this talk I will review the rise, fall, and rise(?) of probability sampling for human populations in the 20th and 21st Centuries, focusing on the problems of quota sampling that lead to the rise of probability sampling as a "gold standard" for population inference, followed by the increasing difficulties that have plagued probability sampling, including increasing cost, declining response rates, and the collapse of convenient sampling frames. At the same time the rise of easily accessible administrative datasets and other forms of "big data" have opened opportunities for research outside the probability sampling paradigm. This had led to a situation where probability samples minimize selection bias but are expensive and may have limited data, and where non-probability sample provide rich data less expensively but may be subject to selection bias. Thus a logical approach is to develop methods that combine information from probability and non-probability samples in an attempt to leverage the strength of each of these approaches. These will include quasi-randomization approaches that use weights to adjust the distribution of common covariates in the non-probability sample to those in the probability sample, post-stratification methods that use prediction models to obtain population-level inference of variables present in only the non-probability sample, and double-robust approaches that combine use of model prediction and weighting and can provide correct inference as long as either the model to account for selection bias in developing the weights or the model used to predict the outcome of interest is correct (but not necessarily both). I will conclude with a discussion about the need for high-quality probability samples for calibration/integration purposes, and briefly touch on the prospects and perils of large language models in the survey setting. I will leave time for discussion about experiences outside the United States, as well as the recent impact of degradation of official statistics in the United States.

Sample coordination: Use and recent developments

***Matei, A.**, University of Neuchatel, Switzerland*

Depending on the application, it may be necessary to obtain or avoid overlap between samples. Sample coordination refers to methods that allow the creation of a probabilistic dependency between sample selections in order to optimize their overlap size. When the goal of a survey is to estimate change over time, or to reduce the cost of recruiting a new sample unit, the overlap size between samples should be maximized (positive coordination). Sometimes one wants to control the risk of the same sample unit being selected in different surveys, and therefore limit the response burden for that particular unit in a given period. In such cases, the goal is to minimize the size of the overlap between samples (negative coordination).

Several methods are used for sample coordination (for an overview, see Matei and Smith, 2023). Methods based on the use of permanent random numbers (PRNs) are particularly popular in the context of sample coordination systems used in official statistics. The use of PRNs has also been extended to coordinate spatially balanced samples in environmental studies.

Maximizing/minimizing the expected overlap size between samples in positive/negative coordination represents an overall standard to evaluate a coordination method. An efficient method (with respect to some theoretical bounds on the expected overlap size) is Poisson sampling with PRNs. While this sampling method is a very attractive scheme for sample coordination, it has an important drawback: the resulting samples have random sizes, which increases the variance of the estimates. For this reason, fixed-size sampling schemes are sometimes preferred, but with a loss of efficiency in sample coordination. An extension of Poisson sampling is Spatially Correlated Poisson Sampling (SCPS). This is an unequal probability sampling design with a fixed sample size that is used to draw samples spread over a space. SPCS with PRNs can be used in positive coordination to monitor changes in population totals over time in various fields, including environmental studies, and also in negative coordination, especially in official statistics (Matei, Smith, Smeets, and Klingwort, 2023; Matei, Pantalone and Smith, 2025).

In this talk, we recall the framework and the use of sample coordination and present some recent developments on this topic in official statistics as well as in environmental studies.

References

Matei, A., Pantalone, F., and Smith, P. A. (2025). Spatially balanced sampling and its applications in official statistics. Under revision.

Matei, A. and Smith, P. A. (2023). Sample coordination methods and systems for establishment surveys. In Snijkers, G., Bavdaz, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J., Thompson, K., and van Delden, A., editors, Advances in Business Statistics, Methods and Data Collection, chapter 27. Wiley, Hoboken.

Matei, A., Smith, P. A., Smeets, M. J. E., and Klingwort, J. (2023). Targetted double control of burden in multiple surveys. Survey Methodology, 49(2):363– 384.

Sensitive surveys with indirect questioning techniques: Designs and challenges

***Perri, P. F.**, University of Calabria*

In survey research, conventional direct questioning (DQ) formats are widely employed to gather data. However, when the issues under investigation are highly personal, sensitive, stigmatizing, or potentially incriminating — such as drug use, domestic violence, racial prejudice, sexual behaviors, illegal income, or noncompliance with laws and regulations — DQ surveys often result in high nonresponse rates and untruthful responding. Survey participants may in fact refuse to answer sensitive items or misreport behaviors and attitudes in an effort to maintain a positive self-image, influenced by social desirability bias.

These effects can seriously undermine the validity of survey analyses, potentially leading to inaccurate estimates of key population characteristics. Consequently, when surveying sensitive topics, it becomes important to adopt appropriate survey designs and techniques that improve data quality and provide meaningful estimates of population parameters of interest, particularly the prevalence of stigmatizing attributes.

Alongside traditional solutions, one approach that is gaining increasing attention involves indirect questioning techniques (IQTs), a class of data collection methods based on the randomization of individual responses. These techniques protect privacy and increase confidentiality by allowing respondents to answer sensitive questions without directly disclosing their true status to interviewers or researchers. This approach is expected to foster greater respondent cooperation, mitigate item nonresponse, and reduce untruthful reporting by minimizing the incidence of socially desirable responding.

The talk will address key issues related to privacy protection in surveys on sensitive topics, provide an overview of the IQT approach, and discuss current limitations and challenges for survey methodologists. Finally, it will illustrate how some techniques have been employed to collect data and obtain prevalence estimates in some recent real-world studies on Covid-19-related health behaviors, sexual orientation, and public attitudes toward the death penalty.

Leveraging technology for data collection in social research and official statistics: Opportunities and challenges

Struminskaya, B., *Utrecht University*

Technological advancements allow researchers to bridge the gap between traditional surveys and promising data collection methods, such as smart surveys that intelligently combine the use of self-report with smart features such as sensors of smartphones, wearables, and other devices, as well as survey augmentation with digital traces through data donation. These methods can help improve data quality, provide rich data, and reduce participant burden. However, researchers face many methodological and technical challenges when designing and implementing smart surveys and/or carrying out data donation studies. Participants have to be willing and able to use their devices to perform additional tasks such as recording geolocation in a smart travel survey, scanning receipts in a smart household budget survey, request data from digital platforms and be able to provide these to researchers in a data donation study. There are issues with measurement and missing data that stem from participants' behavior, the technological solution (e.g., an app) or both, as well as ethical and legal considerations. Participants can help improve data quality but this needs to be carefully considered against the usability of the data collection tool and participant burden. This talk examines challenges of implementing studies using new technologies in social research and official statistics, illustrating these with empirical findings from methodological studies and experiments. The talk concludes by providing practical recommendations and outlining the future research agenda.

Small area estimation in the era of machine learning and new data: Opportunities and challenges

***Tzavidis, N.**, University of Southampton*

Advances in machine (statistical) learning alongside the availability of large datasets from alternative sources have led to the production of small area-type estimates globally and at refined spatial scales. New and rediscovered algorithmic tools and data offer opportunities sufficient to support a period of exciting research but pose significant challenges too.

Private firms and other research organisations are publishing small area-type products using machine learning methods and data from several sources. Meta recently developed methodology to produce, among other estimates, global estimates of average wealth at 2.4km2 resolution (Chi et al., 2022). The methodology, however, fails to acknowledge important methodological and applied work in small area estimation over the past 30 years. It is also not clear if these estimates would pass the quality controls we use in survey and official statistics. And yet, Meta's estimates and methodology are attracting significant interest in applied research.

As Meng (2018) puts it, *"it is generally true that we, as statisticians, lament the increasing loss of principled statistical methodology. However, the more we lament how our nutritious recipes are increasingly being ignored, the more fast food is being produced, consumed and even celebrated as the cuisine of a coming age".* As survey statisticians, we should neither lament nor be complacent otherwise we run the risk that our research contributions become outdated. Instead, we must critically engage by scrutinizing newly proposed methods and data, compare these to industry standard methods, and use theoretical and empirical evidence to influence the direction of research and applications.

In this talk I will focus on recent and ongoing research on the following broad themes that underpin key differences between industry standard and new methods. The themes include (a) the types of models/algorithms used and their specification including the spatial scale at which estimates are produced, (b) measurement issues, (c) uncertainty measurement, and (d) the use of alternative sources of data (geospatial data in this talk) and how these are integrated with survey data. The presentation will include real data examples from ongoing work in Mozambique and the UK. I will conclude the talk by outlining areas where, in my view, more research is needed. The discussion cannot be exhaustive, but I hope it will generate an interesting debate.

References

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, E., J. (2022). Microestimates of wealth for all low and middle income countries. PNAS , 119 (3), 381–399.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. The Annals of Applied Statistics, 12 (2), 685–726.

Invited Sessions

Adjusting survey estimates with multi-accuracy post-processing

Fischer-Abaigar, U., *LMU, MCML; Kern, C., LMU, MCML; Kreuter, F., LMU, MCML, UMD*

With the rise of non-probability samples and new data sources, survey researchers face growing challenges related to selection bias. One emerging line of work adapts algorithmic tools from machine learning to improve robustness in such settings. This talk introduces multi-accuracy boosting (Kim et al., 2019), a post-processing method that reduces subgroup-level prediction error. Originally developed in the context of fairness, it has since been explored for use in survey adjustment tasks (Kim & Kern et al., 2022). I offer an accessible overview of the method and share reflections on its potential, and open questions for future research.

Generalized tree-based machine learning methods with applications to small area estimation

***Frink, N.**, Freie Universität Berlin; Schmid, T., Otto-Friedrich Universität Bamberg*

We propose small area estimation methods that utilize generalized tree-based machine learning techniques to improve the estimation of disaggregated means in small areas using discrete survey data. Our methods employ machine learning techniques to identify predictive, non-linear relationships from data, while also modeling hierarchical structures. Specifically, we present the Generalized Mixed Effects Random Forest (GMERF), tailored to address challenges associated with binary and count outcomes. In addition,

we present and evaluate three bootstrap approaches - parametric, semi-parametric and non-parametric - designed to assess the reliability of point estimators for area-level means. The effectiveness of these methodologies is tested through model-based (and design-based) simulations and applied to real-world datasets.

Statistical inference for a finite population mean with machine learning-based imputation for missing survey data

***Haziza, D.**, University of Ottawa; Dagdoug, M., Mc Gill University*

National statistical offices are increasingly using machine learning (ML) to improve survey estimates. ML methods help handle high-dimensional data and capture complex relationships, improving survey accuracy. In this presentation, we discuss a double/debiased ML framework for handling item nonresponse while ensuring valid statistical inference with ML-based imputation. We also present theoretical and simulation results that illustrate the framework's effectiveness across different scenarios.

Adjusting survey estimates with multi-accuracy post-processing

***Bianchi, A.**, University of Bergamo; Shlomo, N., University of Manchester*

Due to increasing non-response and budget cuts, survey organizations are researching methods for integrating data from multiple sources to improve accuracy and timeliness of population estimates. Particularly, on-line panels and administrative data are increasingly used for this purpose. Their low cost and fast availability make them very attractive to be used for statistical purposes. However, they are not based on probability sampling designs and hence they are often selective, and estimators based on such samples are usually biased.

Several approaches have been proposed in the literature to correct for such selection bias. Examples are the quasi-randomization approaches and the combination of estimates from the nonprobability sample with those from a probability sample (Elliott and Valliant, 2017). However, the theory for making inference with nonprobability samples and integrating them with probability samples is not fully developed yet (Lohr and Raghunathan, 2017; Rao, 2021).

In this framework, we focus on the situation where the variable of interest measured in the nonprobability source is also observed in a probability sample and we adopt a multiple frame (MF) approach for the production of the estimates. Traditionally, MF surveys are surveys in which two or more frames are used and probability-based independent samples are respectively taken from each frame. Inferences about the target population are based on the combined sample data. MF surveys have been used to improve coverage, reduce costs, and increase sample sizes for subpopulations of interest. Recently, the MF perspective has also been considered to include a nonprobability component. In the MF setting, the nonprobability sample is viewed as a census of part of the population and the bias is eliminated by the presence of a probability sample from the complete frame. In the estimation process, it is necessary to account for potential overlap among the samples (Lohr, 2021; Kim and Tam 2021). A thorough evaluation of such methods is still lacking in the literature.

We explore the feasibility of such an approach through simulations using the UK census data and mimicking administrative data and opt-in panel survey data. We compare several estimators and consider different scenarios, depending on available information on domain membership and presence of measurement error. Furthermore, we provide an application to the Publishers Audience Measurement Company Survey (PAMCo) conducted by Ipsos UK. PAMCo is the audience measurement currency for readership of newspapers and magazines in Great Britain and it is a probability-based survey. Recently, due to reduced budgets, the possibility of boosting the sample with a non-probability panel was explored and the same questionnaire administered to members of the panel.

Overall, the results show that the MF perspective is a promising approach for making inference from nonprobability sources. However, when the size of the nonprobability source is small compared to that of the population, the impact of the nonprobability source on the overall estimate is rather limited.

Variance estimation and robustness in integrated estimation from probability and non-probability samples

Čiginas, A., *Vilnius University; Krapavickaitė, D., Lithuanian Statistical Society; Nekrašaitė-Liegė, V., Vilnius Gediminas Technical University*

In this study, we address the integration of data from independent probability and non-probability samples for estimating finite population totals. The integration is performed through a linear combination of a design-based estimator from the probability sample and an inverse probability weighted (IPW) estimator from the non-probability sample. A key feature of this approach is the bias correction for the IPW estimator, ensuring robustness to misspecification of the propensity score model.

A particular focus is given to variance estimation, accounting for additional sources of randomness due to the uncertainty in the estimated propensity scores and the unknown selection mechanism of the non-probability sample. Two variance estimators are proposed: a Monte Carlo-based anticipated variance estimator and a bootstrap variance estimator. These are compared with an estimator derived from the asymptotic variance under a parametric propensity score model.

A simulation study, based on Lithuanian enterprise data, illustrates the performance of the estimators under different sample integration scenarios. The results demonstrate that the proposed composite estimator remains robust to shifts in the non-probability sample and that the alternative variance estimators lead to different estimates of variability, especially when the propensity score model is misspecified.

Some recent research at Statistics Netherlands on combining probability and non-probability samples

***Scholtus, S.**, Statistics Netherlands*

Many national statistical institutes (NSIs) and other producers of official statistics are looking for ways to supplement or even replace data from traditional sample surveys by data from other sources, such as register data and ‘big data’. One issue with the use of such data sources for official statistics is that they may cover only part of the target population, while the underlying selection mechanism is not controlled by the NSI and often unknown. Thus, in general, they are non-probability samples.

Two different situations can be distinguished: (1) the target variable is observed only in a non-probability sample; (2) the target variable is observed both in a non-probability sample and a (usually much smaller) probability sample. In the first situation, there is an important risk of selection bias. This situation has therefore in the past mostly been avoided by NSIs. Recently in the literature, several methods have been proposed that try to address selection bias by modeling either the selection mechanism of the non-probability sample, or the population distribution of the target variable, or both. In the second situation, one option is to apply a traditional design-based survey method, by letting the non-probability sample represent only itself and using the probability sample to estimate the contribution of the rest of the population. In terms of selection bias this is a relatively ‘safe’ approach. In terms of variance, it has the drawback that relatively little information is used from the non-probability sample. Therefore, research is ongoing into more efficient methods for combining two samples.

In this presentation, a brief overview will be given of several recent research contributions at Statistics Netherlands about the use of non-probability samples in official statistics, specifically for the second situation mentioned above. These include:

- an approach to estimate a contingency table by combining a probability and non-probability sample, if only information on an aggregated level is available from both samples (Villalobos Aliste et al., 2025);

- an approach that uses regularized regression to estimate a linear regression model from a combination of a probability and non-probability sample (Westlund, 2024);

- an approach to design a probability sample to supplement a non-probability sample, so that the combined sample is representative of the target population (Boeschoten et al., 2023; Mensink, 2023).

References

L. Boeschoten, S. Scholtus & A. van Delden (2023), A Note on Efficient Audit Sample Selection. Discussion Paper, Statistics Netherlands, The Hague.

L. Mensink (2023), Efficiently Selecting Representative Audit Samples. Master Thesis, Utrecht University.

S. Villalobos Aliste, S. Scholtus & T. de Waal (2025), Combining Probability and Nonprobability Samples on an Aggregated Level. Journal of Official Statistics (advanced publication online).

J. Westlund (2024), Regularizing Probability Sample Estimates through an Angle-Based Similarity Approach. Master Thesis, Leiden University.

Imputing local income distributions with SSIT-GAMLSS

***Betti, G.**, University of Siena; Mori, L., University of Bologna*

This study addresses the methodological challenges involved in estimating inequality measures from survey data with imputed variables, particularly when dealing with limited and locally collected data sources. It proposes an advanced statistical framework, Survey-to-Survey Imputation Techniques based on Generalized Additive Models for Location, Scale, and Shape (SSIT-GAMLSS), designed to transform income data originally collected in broad categorical classes into a continuous variable. This transformation is guided by the distributional properties observed in a richer reference survey—EU-SILC—thus leveraging high-quality data to enhance the utility of smaller, local surveys that would otherwise offer limited analytical potential.

The SSIT-GAMLSS methodology provides significant flexibility by allowing covariates and random effects in all distributional parameters—location, scale, and 2 shapes—rather than only the location. This enables more accurate modeling of income distributions, even in the presence of heterogeneous or complex data structures. Unlike traditional imputation approaches, which often rely on rigid assumptions and struggle with non-normality or transformation-related biases, SSIT-GAMLSS adapts to the full shape of the empirical distribution.

The empirical application focuses on regional income data from Italy, particularly local surveys that capture income only in discrete brackets. Through the application of SSIT-GAMLSS, these class-based income variables are imputed as continuous, thereby allowing the estimation of inequality and poverty measures with far greater detail. Crucially, this method enables analysis at a much finer territorial scale than what is typically feasible using EU-SILC alone, unlocking insights at the sub-regional or even municipal level. The resulting estimates are validated against benchmark indicators and provide a robust foundation for understanding spatial disparities in income and living standards.

The research contributes to the literature on survey-to-survey imputation and income distribution modeling by introducing a practical tool for integrating auxiliary data, handling incomplete information, and producing reliable inequality indicators from otherwise limited surveys. Furthermore, it addresses common pitfalls in imputation such as back-transformation errors and unstable parameter estimates across regions or time. The SSIT-GAMLSS framework proves especially valuable in contexts where statistical infrastructure is limited or where high-resolution socio-economic data is unavailable through official channels.

This work has important implications for National Statistical Offices (NSOs), local administrations, and policy institutions seeking to strengthen evidence-based policy design. By enhancing the analytical power of small-scale surveys and enabling comparisons across space and time, the approach supports more targeted and equitable policy interventions. Future developments may include the extension of this framework to other dimensions of well-being, such as consumption or wealth, and its integration into routine statistical production processes.

Parametric and semi-parametric methods for estimating poverty transition with synthetic panels

***D'Alberto, R.**, University of Verona; De Nicolò, S., University of Bologna; Gardini, A., University of Bologna*

The study of poverty transition is hindered by the scarcity of longitudinal data. Hence, there is increasing interest in methodologies that infer transition probabilities from cross-sectional data sets. Both parametric methods, which employ econometric techniques on pseudo-panels by assuming a distributional income model, and semi-parametric ones, which rely on matching procedures to create synthetic panels are applied. Despite the widespread use, these methods have not been systematically reviewed or assessed statistically. This proposal critically examines existing approaches, highlighting their limitations, before introducing a novel scenario-based framework that enhances both methodological strands. In detail, the parametric extension integrates Bayesian models, incorporating scenario-based prior information to estimate income autocorrelation. The semi-parametric alternative refines matching techniques by introducing a tuning parameter that adjusts the number of neighbors, hence controlling the level of autocorrelation. The effectiveness of our proposals is assessed through a Monte Carlo simulation study based on Italian EU-SILC data. The performance of conventional methods is compared with our approaches regarding poverty transition probability estimates. The results offer valuable insights into the relative strengths and weaknesses of each method, providing a statistically rigorous foundation for future research in poverty dynamics.

Editing and imputation: An official statistics perspective

Rosati, S., *ISTAT; Filippini, R., ISTAT; Toti, S., ISTAT*

The National Statistical Institutes (NSIs), as official producers of statistical data, must guarantee high quality standards both at the micro and macro levels. This contribution aims to illustrate the essential points and the evolution of statistical data editing process, specifically the treatment of non-sampling errors, within the new context of statistical production, where multiple data sources are used.

The high availability of administrative data sources has led to create a new paradigm, placing registers at the centre of statistical production. This implied an innovation in the process of producing statistical data. More precisely, surveys as the primary source become a required auxiliary source for addressing the typical problems of registers, such as the completeness of information in terms of units and variables (coverage of units and estimation of variables not measured by administrative data sources). Despite the increasing availability of data from administrative sources, the necessity to collect information not provided by them persists, making it essential to continue to collect information through surveys that use representative samples of the entire target population.

Traditional statistical surveys, while continuing to be central to statistical production, require increasingly effective techniques and tools compared to the past that allow to improve efficiency and to reduce the errors (e.g. by adopting computer-assisted interviewing).

By exploiting administrative data and sample surveys, the Italian statistical system has evolved towards a multi-source statistics production model, built around an “Integrated System of Registers” (SIR). As a consequence, the adoption of new approaches has become necessary to preserve the efficiency and overall quality of statistical production.

In the new context of the registers, the usual procedures used for data editing and imputation have been adapted to ensure greater efficacy. In particular, it was necessary to intervene on several key aspects such as the evaluation of data consistency to check and correct any inconsistencies, the imputation of missing data and the quality assessment of the estimates, produced in terms of distribution and uncertainty evaluation (sampling variance and/or imputation variance).

For some of these aspects, innovative methods applicable to official statistics are currently under study, including the use of Machine Learning (ML) approaches for data imputation, which could improve the accuracy of estimates by optimizing the process.

The procedures adopted by ISTAT for the production of census estimates on Employment and Educational Attainment represent a significant example of these transformations. In the past, these statistics were based on exhaustive census surveys, whereas today they are produced by integrating administrative data with data from sample surveys. In this new approach, the statistical data editing process plays a crucial role in ensuring both the consistency of data collected through surveys and the quality of data from administrative registers.

In the specific case of Educational Attainment, to ensure the completeness of the Base Register of Individuals, a massive imputation was applied using all available information from administrative sources and surveys. This approach aimed to estimate missing data due to coverage problems and improve the overall representativeness of the entire system.

Roots, trends and emerging horizons of multiple frame surveys

***Mecatti, F.**, University of Milano-Bicocca*

This paper explores the development of Multiple Frame (MF) surveys, from their first appearance in the 1960s–1970s literature to their role in today's fast-evolving landscape of sampling statistics and emerging data needs.

Traditional survey methodology assumes a single and complete-coverage frame for sample selection. In contrast, an MF survey uses two or more frames, which may have partial coverage and usually overlap, but together adequately cover the study population. MF surveys are especially useful when no single complete frame exists, and there is not enough information or resources to build one through linkage. Still, MF surveys can be cost- effective even when a full frame is available, for example supplementing it with a partial list of email addresses to reduce data collection costs. This cost-saving motivation was behind their introduction in the pioneering work of Hartley (1962, 1974), who exemplified with a Dual Frame (DF) agricultural survey combining a complete area frame with a partial list frame. He showed that DF surveys can yield estimates as precise or more so than single frame surveys, and at lower cost.

Since then, MF surveys have undergone a significant evolution, both in objectives and application areas, and in estimation approaches. The initial cost-efficiency scope expanded to include difficult-to-sample populations, e.g., undocumented immigrants, drug users, or homeless people, usually lacking a unique conventional frame due to being "hidden and elusive" (e.g. Kalton & Anderson, JRSS-A 1986 for a review). Recent needs include MF setups to handle attrition in longitudinal studies (e.g., EU-SHARE www.share-project.org), DF agricultural surveys combining satellite and field data (Ferraz et al., SMA 2023), and using MF-based strategies to address long-standing challenges like rising costs and declining response rates. These strategies aim to leverage new digital data sources and big data (e.g., Ferraz's et al. 2025 pre-print), to integrate multiple data sources (Lohr, SurvMeth 2021), or to support inference from non-probability sam ples (Lohr, SurvStat 2025; Rao & Lohr, forthcoming SurvMeth 2025).

Turning to MF estimation, we see a parallel evolution since Hartley optimal approach, aimed at minimizing estimator variance but sub-optimal in practice due to the need to estimate variances from the sample itself. Alternative strategies emerged, initially focused on the simpler DF case and a limited alphabetical notation. Generalizations to full MF setups started in the late 2000s, introducing a more rigorous indexed notation (Lohr & Rao, JASA 2006; Mecatti, SurvMeth 2007). Recent methods are built on the multiplicity approach (Mecatti & Singh, JSFS 2014; Rao & Wu, JASA 2010), which offers a unified framework to MF existing estimators motivated by different considerations. This started with the introduction of the class of Generalized Multiplicity-adjusted Horvitz-Thompson (GMHT) estimators (Singh & Mecatti, JOS 2011) that includes all unbiased MF estimators previously proposed. The paper concludes with ongoing research on more general classes of multiplicity-adjusted GREG estimators, including the widely used PML estimator (Rao & Skinner; Lohr & Rao, 2006) and the forerunner calibration-based DF method by Ranalli et al. (2016).

Use of multiple frames in the integration of probability and non-probability surveys

Cobo, B., *University of Granada; Rueda, M.M., University of Granada; Rueda, J.L., University of Granada*

In recent years, there has been concern on the part of researchers about the lack of coverage and lack of response when carrying out a survey and, since the costs are increasing, they have considered whether the non-probability sampling could be a good option. Non-probability surveys have had great growth, since they are carried out quickly and economically, but they also have problems and that is why it is decided to integrate both methodologies, that is, the integration of probability and non-probability surveys, with the aim of taking advantage of each approach, the theoretical solidity of probability surveys and the accessibility of non-probability surveys.

In this context, we will consider the common practical situation in which one of the frames used for sample selection does not fully cover the target population. This situation is common in web surveys or social media surveys, where coverage is restricted to a population with certain sociodemographic characteristics different from the target population. To work in this context, we will use the multiple-frame methodology. This allows us to combine different sources of information to correct deficiencies such as lack of coverage, lack of up-to-date records, or difficulty accessing certain subpopulations. This approach is especially useful when some population groups are difficult to reach using a single frame, or when the cost and efficiency of sampling vary across sources.

Therefore, our objective is to efficiently combine the probability and non-probability sample to estimate a linear parameter in this context using multiple frame tools considered in the literature, in particular we will use the single frame technique and the dual frame technique considering different weighting values. In addition to carrying out the parameter estimation, we are also interested in obtaining its variance and for this we will use resampling methods, such as Jackknife and Bootstrap.

After analyzing the results, we see how the proposed estimators work quite well after analyzing the bias and the mean square error. If we focus on variance estimation, we see that the results obtained with the Bootstrap technique are better than those obtained by Jackknife in terms of studying their coverage and length of the confidence interval.

Multiple frame surveys in modern data integration

***Saegusa, T.**, University of Maryland*

Multiple frame surveys provide effective ways to integrate multiple data sets from heterogeneous sources. Well-motivated by traditional survey sampling, the scope of this design is unfortunately too limited to address modern applications in data integration. The first part of the talk develops methods for hypothesis testing often overlooked by survey sampling when data sets are obtained from independent surveys. Because parameter of interest is not a finite population parameter, we adopt the super population framework. This additional randomness introduces multitude of dependence within and across multiple data sets through potential duplication and finite population sampling so that quantifying uncertainty is much harder and challenging than in the finite population framework. With this complication, the distributions of the inverse probability weighted version of pivotal quantities in the i.i.d. setting becomes no longer parameter-free. Our proposed methodology first develops asymptotic theory for multiple frame surveys with a super population and then estimates complex parameter-dependent null distributions through simulation and/or bootstrap. Our methods are illustrated with data analysis of the Wilms tumor study. The second part is our attempt to apply the framework of multiple frame surveys to non-probability samples. In the non-probability sample research, one combines a reference survey and a non-probability sample whose missingness mechanism is unknown. We extend our asymptotic theory to combining data from sample surveys and missing data. After reviewing methodology in non-probability samples, we discuss limitations of this approach and potential solutions using techniques from surgery sampling such as record linkage.

Sanitary migration in Italy: A model-based approach

***Arima, S.**, University of Salento, Angelelli, M., University of Salento, Polettini, S., Sapienza University of Rome, Ciavolino, E., University of Salento*

Healthcare mobility represents one of the main criteria for evaluating Regional Healthcare systems, in terms of quality of services and patient satisfaction. It is widely recognized that most of the patients residing in the southern regions are more likely to be hospitalized in specialized centers in the northeastern regions. However, it is crucial to understand which factors influence such mobility. Some of these factors can be explored by examining the socio-economic demographic indices, such as income and education levels, of different areas at the province level. These characteristics may serve as informative proxies about patients' resources to support treatment outside their home territory, as well as the cultural background and the context that may influence trust in and satisfaction with the local healthcare system. At the same time, it is essential to consider the resources available in each territory, including both healthcare-related factors and logistical aspects. Additionally, since regional bodies play a central role in managing healthcare resources, it is important to account for their strategic initiatives, particularly their investments in health services. Therefore, studying health mobility and its underlying drivers requires a multiscale approach that considers both provincial and regional sources of variability to better understand the movement of residents seeking care outside their home regions. We collected data for all Italian provinces: we aim at modeling the time series of the migration rates in selected periods from 2000 to 2022 and evaluating the impact of social and economic covariates on such mobility. However, insightful covariates are rarely available at the provincial level, while they are available at the regional level. When data are aggregated from a fine (e.g. province) to a coarse (e.g. region) geographical level, there will be a loss of information. For each province, we collect variables for characterizing the efficiency of the sanitary structures, such as the number of beds in the hospitals, the number of medical doctors for 1000 patients, the availability of specialized centers. We also collect information about the social and economic context of each region (e.g. average age, percentage of individuals with higher education levels) as well as the distance from the closest airport. At regional level, BES indicators are also available, giving a more complete image of the economic and social background of each region. We propose a convolution model in which the province-level convolution model is connected to the standard regional level convolution model via shared spatial structured random effects representing the intra-regional and extra-regional mobility. This study was partially funded by the European Union - NextGenerationEU, in the framework of the "GRINS -Growing Resilient, INclusive and Sustainable project.

Poverty prevalence among people with disability: A small area estimation approach

***Fabrizi, E.**, Università Cattolica del Sacro Cuore*

People with disabilities often face significant barriers that contribute to higher poverty rates compared to those without disabilities. These barriers include limited access to education, employment, and social services, which can lead to lower income levels and increased dependency on social transfers. In 2023, EU-SILC data revealed that 28.8% of people with disabilities in the EU were at risk of poverty or social exclusion, compared to 18.0% of those without disabilities according to the latest EUROSTAT data. We focus on Italy and the EU-SILC survey as the sample source for poverty and disability measurement.

The exposure to the risk of poverty and social exclusion varies across different regions and social contexts. For this reason, we are interested in obtaining estimates for geographical regions classified by degree of urbanization. Geographical regions are very relevant when studying Italy, as health indicators such as life expectancy and prevalence of chronic diseases, efficacy of the health system shows significant differences between the North and South. Degree of urbanization is considered as a proxy of social context and access to services and job opportunities for people with disability. Specifically, urban areas often have better infrastructure, including accessible public transportation, healthcare facilities, and employment opportunities. However, they can also present challenges such as higher living costs and overcrowded services.

Given that people with disabilities are a minority, measuring their exposure to poverty for sub-populations cannot be conducted using standard survey weighted estimators because of insufficient sample sizes. This motivates our recourse to small area estimation methods.

Our approach relies on area-level direct estimates obtained from the EU-SILC survey complemented with auxiliary information from demographic registers along with social security and fiscal databases. Standard errors associated to direct estimates are involved in the estimation process to produce estimates with associated variability measures accounting for all the uncertainty sources. We adopt a Bayesian approach to estimation, to estimate and implement a mixed Beta regression model extended to accommodate for the possible presence of direct estimates equal to either 0 or 1, as well as out of sample areas. Posterior distributions of relevant parameters are obtained using Markov Chain Monte Carlo algorithms. The problem of benchmarking small area estimates of poverty rates to those obtained using standard survey weighted estimators for large areas with adequate samples is also considered. The recourse to Markov Chain Monte Carlo methods allows, given the choice of an appropriate loss function, to obtain benchmarked estimates within the (0,1) interval and endowed with an uncertainty measure accounting for the benchmarking too.

This research was conducted within the framework of the “GRINS -Growing Resilient, INclusive and Sustainable project funded by the European Union – NextGenerationEU.

A spatially geographically weighted Fay-Herriot model for small area estimation – An application to poverty indicators in Italy

***Schirripa Spagnolo, F.**, University of Pisa; Giusti, C., University of Pisa; Moretti, A., Utrecht University; Salvati, N., University of Pisa*

Local governments play a vital role in designing and implementing policies aimed at supporting vulnerable populations. However, accurately measuring poverty at the local level poses considerable challenges. National sample surveys, which are typically designed to generate reliable estimates at broader geographic levels, often lack the granularity needed for small-area analysis. These surveys may include only a limited number of observations for specific regions or subpopulations, resulting in unreliable direct estimates. In some cases, certain areas may not be represented at all, making it necessary to rely on indirect estimation methods to produce robust and policy-relevant evidence.

Small Area Estimation (SAE) methods offer an effective solution to this challenge by employing model-based approaches that combine survey data with auxiliary information available at the population level. Among these methods, the Fay-Herriot (FH) model is one of the most widely adopted. Based on a linear mixed model framework, the FH approach incorporates random area effects to capture between-area variability, thereby improving the precision and reliability of estimates for small geographic areas or population subgroups.

Spatial information is particularly relevant when estimating poverty indicators, as these measures often exhibit strong geographical patterns. Several spatial extensions of the FH model have been proposed, including approaches that incorporate spatially correlated random effects or penalized splines. An alternative way to integrate spatial information into small-area models is by allowing model coefficients to vary across geographical regions, reflecting local heterogeneity in the relationships between poverty indicators and auxiliary variables.

In this study, we propose an innovative extension of the Geographically Weighted Regression (GWR) approach within the Fay-Herriot framework for Small Area Estimation. The proposed model allows regression coefficients to vary spatially, offering a more flexible and nuanced representation of geographical disparities in poverty indicators. By explicitly incorporating spatial heterogeneity into the estimation process, our method enhances the precision and accuracy of small-area estimates - particularly in regions with sparse survey data. We apply this approach to estimate the spatial distribution of poverty indicators in Italy, demonstrating its effectiveness in capturing local variation and generating more reliable estimates to inform sub-national policy decisions. Our findings underscore the benefits of integrating geographically weighted techniques into the traditional Fay-Herriot model, providing valuable insights for policymakers aiming to design targeted interventions to alleviate poverty.

This study was partially funded by the European Union - NextGenerationEU, in the framework of the "GRINS -Growing Resilient, INclusive and Sustainable project.

Unit-level models for multivariate binary data: Small area estimation of social inclusion indicators

***Failli, D.**, University of Perugia; Marino, M. F., University of Florence; Ranalli, M. G., University of Perugia*

Promoting inclusion and cohesion is one of the six key missions of Italy's National Recovery and Resilience Plan (PNRR). Over the past decade, social exclusion in Italy has worsened, largely due to the economic crisis of 2008-2011. The COVID-19 pandemic further exacerbated these disparities, particularly highlighting differences in access to technology and digital skills, which have become significant factors in social exclusion. This decline in inclusion is unevenly spread across regions and various segments of society. To translate PNRR's vision into concrete policies, it is essential to measure indicators at the local level.

In this talk, we focus on estimating ISTAT BES indicators for the domain of Social Relations for subpopulations currently not covered by official statistics. In particular, we use data from the Multipurpose Survey on Households, "Aspects of Daily Life", conducted annually by ISTAT, which can provide reliable estimates only for domains planned in the survey design phase, such as Administrative Regions (NUTS 2). Consequently, at sub-regional level and for some subgroups of the population, the estimates may not be considered reliable. To overcome this limitation, we propose a multivariate unit-level SAE model for binary data to estimate the indicators of interest at a disaggregated level with significantly better efficiency compared to direct estimates that rely only on sample data related to the subpopulation of interest.

The proposed model assumes area-specific random effects to account for sources of unobserved heterogeneity that are not captured by the covariates and to describe correlation between units within the same small area. It also assumes the existence of a multidimensional, continuous, latent variable (trait) associated to each unit in a given area. This is assumed to capture the influence of unobserved (latent) characteristics on the multivariate binary variables.

In this context, the computation of the empirical best predictor and the analytic approximation to its mean squared error require the solution of multiple integrals that do not have a closed form. To solve the issue, we propose a semi-parametric multivariate empirical best predictor by leaving the distribution of the area-specific random effects unspecified and estimating it directly from the observed data. This approach is known to lead to a discrete mixing distribution that helps avoid (i) unverifiable parametric assumptions and (ii) heavy integral approximations. Furthermore, to avoid deviating from standard assumptions about the Gaussianity of the latent trait(s), we adopt a numerical approach based on Gaussian quadrature.

Overall, a multivariate model can lead to more efficient estimators of the small area proportions by taking advantage of the association among response variables, as opposed to a univariate model. Furthermore, given the scarcity of auxiliary information at the population level, the presence of a multidimensional continuous latent variable allows to capture the influence of unobserved (latent) characteristics on the multivariate binary outcomes.

The work has been developed under the support of European Union - Next Generation EU, Mission 4 Component 1 CUP J33C22002910001 --- Project INCLUDE-ALL: social INCLUsion and Digital indicators Estimation At Local Level.

Indirect sampling for the spatialization of maritime fishing activities: The case of Valpena

***Medous, E.**, IGN; El Ghaziri, A., Institut Agro Anger-Rennes; Rollo, N., LETG Nantes; Trouillet, B., LETG Nantes; Bellanger, L., Nantes University; Dieudonée, E., LETG Nantes*

In a context of growing maritime space sharing, spatializing fishing activities is an important scientific and planning challenge, especially for small vessels that are not equipped with satellite tracking systems.

The VALPENA project (éVALuation des activités de PÊche au regard des Nouvelles Activités, assessment of fishing activities in light of new activities) aims to develop tools that enable fishermen to spatialize their activities by creating maps. These tools use a grid system to divide maritime space.

The surveys conducted by VALPENA seek to observe the grids visited by fishing vessels. However, the visited grids cannot be directly observed, as only the geographic location of the grid cells is available. These grids are defined by the presence of fishing vessels, so VALPENA decided to use an indirect survey (Deville and Lavallée, 2006; Lavallée, 2007), utilizing the fishing vessel population to identify the visited areas.

Indirect surveys are often used to estimate totals within populations of interest (e.g., Deville and Maumy-Bertrand, 2006; De Vitiis et al., 2014). VALPENA's goal is to annually produce maps of the distribution of geographic variables of interest. The project seeks to recover the values of these variables for each visited grid, without focusing on the total values across all grids.

The variables of interest for indirectly sampled grids are obtained from the vessels that visit them. No data is available for unvisited grids, leading to incomplete maps. A census of fishing vessels could avoid this issue but would increase survey costs and non-response rates due to fishermen's weariness. Therefore, VALPENA decided to alternate between one year of census and two years of sampling.

VALPENA uses stratified sampling to select fishing vessels in order to maximize the number of observed visited grids while minimizing the number of vessels surveyed. If the links between the fishing vessels and the grids were known beforehand, stratifying the vessels based on these links would allow for good coverage of the grids with a small sample size of vessels. However, this information is unknown and variable from year to year, so it cannot be used for stratification. Nonetheless, the number of grids visited by a vessel during the census years is strongly related to the links between vessels and grids and remains relatively stable from year to year. Simulations were conducted to test stratification based on this information and compare it with stratifications based on annually available data.

This presentation aims to introduce the different stratifications tested by VALPENA and their impact on the sample size of the indirectly sampled areas.

References

Deville, J.-C. et Lavallee, P. (2006). Indirect sampling: the foundations of the generalized weight share method, Survey methodology, 32(2), 165-176.

Lavallée, P. (2007). Indirect sampling, Springer-Verlag New York.

Deville, J.-C. et Maumy-Bertrand, M. (2006), Extension of the indirect sampling method and its application to tourism, Survey methodology, 32(2), 177.

De Vitiis, C. et al. (2014), A methodological approach based on indirect sampling to survey the homeless population, Rivista di statisticaufficiale, 1(2), 9-30.

Variance estimation for spatially balanced sampling designs: A review and new results

***Pantalone, F.**, University of Southampton; Benedetti, R., University of Chieti-Pescara; Piersimoni, F., ISTAT; Ranalli, M.G., University of Perugia*

Environmental surveys are carried out to gain information on spatial units for inference purposes. Indeed, surveying environmental resources typically involves sampling units distributed over space. For example, a forest inventory might aim to estimate the number of trees in a given region; an agricultural survey could aim to estimate the crop yield of an area; the abundance of a rare or endangered bird species might be of interest. Spatially balanced sampling designs aim to randomly select samples well spread over the population of interest in order to capture spatial heterogeneity. It has been showed that this type of design leads to considerable gain in efficiency (in terms of variance) of the Horvitz-Thompson estimator when the target population has a spatial structure (compared to the use of sampling designs that do not take into account spatial information). Unfortunately, variance estimation becomes challenging as the second-order inclusion probabilities, which are used in the standard variance estimator, are usually intractable for these designs. We provide a review of estimators that can be used in this context (estimators either specifically introduced in the literature for this, or other estimators introduced for different purposes but that can be useful to apply here as well), and we present some new approaches based on variance estimators for non-measurable designs, and pseudo-population bootstrap based estimators. The performances of the estimators are investigated by means of Monte Carlo simulations over generated and real data.

Sampling artificial stands: Can spatial information play a role?

***Dickson, M.M.**, University of Padova; Bucci, G., CNR-Institute of Biosciences and BioResources; Ioveno, P., CNR-Institute of Biosciences and BioResources; Marchi, M., CNR-Institute of Biosciences and BioResources; Puletti, N., CREA; Serio, R.G., University of Trento; Toffoli, D., University of Trento*

Non-native tree species are often introduced from geographic and genetically differentiated source populations to ensure suitable genetic diversity and evolutionary potential to the introduced populations. The introduction of non-native plant species to new environments may cause a severe disequilibrium in the hosting ecosystems (Brus et al., 2019) altering the interspecific competition and affecting the community composition. In some cases, the geographic provenance and the genetic make-up of the original propagation material are unknown, so that the identification of source populations are crucial for managing the extant stands of exotic species, as their evolutionary potential and adaptability to the new environment depend on their genetic diversity. To achieve these aims, it is necessary to identify the most suitable sampling method for these artificial stands, as well as the desirable sampling size. Traditionally, such populations are sampled using conventional methodologies in forestry (e.g., plot sampling), aiming to preserve a certain degree of randomness while also adapting to practical constraints. However, in recent years, we have witnessed a flourishing of sampling techniques that consider the geographic location of the plants, to obtain spatial samples which are spread over the study area (see, among others, Grafström, 2012; Grafström et al., 2012; Grafström and Tillé, 2013). These methods are very efficient when it comes to natural forests, but in the case of artificial woodlands, this is not necessarily the case. In this work, we conduct a comparison study on Douglas-fir [Pseudotsuga menziesii (Mirb.) Franco] forests in Italy, to evaluate performances of different sampling methods and trying to establish a establish a protocol that can be used in similar research contexts.

References

Brus, R., Pötzelsberger, E., Lapin, K., Brundu, G., Orazio, C., Straigyte, L., & Hasenauer, H. (2019). Extent, distribution and origin of non-native forest tree species in Europe. Scandinavian journal of forest research, 34(7), 533-544.

Grafström, A. (2012). Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference, 142(1), 139-147.

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

Grafström, A., & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics, 24(2), 120-131.

Predictor selection under misspecified models

***Żądło, T.**, University of Economics in Katowice; Chwila, A., University of Economics in Katowice; Hadaś-Dyduch, M., University of Economics in Katowice; Krzciuk, M., University of Economics in Katowice; Stachurski, T., University of Economics in Katowice; Wolny-Dominiak, A., University of Economics in Katowice*

In the literature, many methods of model selection are presented. These methods can generally be classified into three categories based on the goals of the selection criteria: those focused on the sample performance of the model (which includes goodness-of-fit measures and hypothesis testing), ex-post prediction accuracy (including cross-validation), and ex-ante prediction accuracy. The methods in the last category are not as widely discussed as those in the first two groups, and many address ex-ante prediction accuracy only indirectly or are limited to selecting the best combination of independent variables within a single model. We introduce a novel simulation-based approach that generalises existing model selection methods based on ex-ante prediction accuracy in several significant ways. First, rather than concentrating solely on model selection, we broaden the focus to select a predictor. Second, our analysis extends beyond predicting just the dependent variable; we investigate the more comprehensive issue of predicting a vector of any functions of this variable. Third, our procedure is not restricted to sample-based models; it can also be applied to misspecified models, aligning with the objectives of establishing a robust predictive strategy. Fourth, the proposed method is highly flexible, making it applicable under any model, whether parametric or nonparametric, including machine learning techniques and utilising any type of data - cross-sectional, longitudinal, or time-series. Finally, our approach can take into account several measures of prediction accuracy simultaneously, rather than relying on a single measure with an observed minimum value leading to the selected predictor. Theoretical considerations are supported by a real data application in the US real estate market. We consider the joint prediction of several price dispersion measures in the whole population and within specific subpopulations, under misspecified models that account for four market shock scenarios not observed in the sample data.

Modelling workplaces and commuting flows using survey and administrative data

Pavasare, R., *Central Statistical Bureau of Latvia*

The greatest challenge addressed in this project is the lack of a direct assignment of individuals to specific LKAUs. This arises because administrative data typically link individuals only to enterprise-level identifiers, while survey data, although available at the LKAU level, do not contain individual-level detail needed for precise spatial analysis. To close this gap, the project introduces a model that links individuals to the most plausible workplace unit, drawing on geographic proximity and occupational fit.

The modelling process involves several activities. For enterprises with only one LKAU, all individuals are directly assigned to that unit. For enterprises with multiple LKAUs, residence of each individual is geocoded, and road distances to all potential LKAUs within the enterprise are calculated. At the same time, the model assesses compatibility between the occupation (based on administrative tax data) and economic activity of each LKAU. To reflect this in the outcome, the distances are weighted – shorter distances are given greater weight when the occupation-sector match is stronger – increasing the likelihood of realistic assignments.

The employment counts available in the Statistical Business Register and surveys are used as constraints. If the total number of individuals linked to the enterprise does not match the survey-based LKAU counts, proportional redistribution with rounding is applied to ensure consistency. The final assignment is formulated as a Mixed Integer Linear Programming (MILP) problem, with the objective of minimising the total commuting distance while meeting constraints on employee allocation and unit-level capacities.

There are two key scenarios where special treatment is applied. One involves educational institutions, matched using names from the State Education Information System and LKAU registers that are sufficiently similar to allow direct linking. The other concerns the units affected by the 2021 administrative-territorial reform – the reorganised municipal authorities in particular – which are treated separately to ensure that structural changes are accurately reflected.

The outlined methodology was applied to 2017–2022 administrative data, resulting in a harmonised individual-level dataset in which each person is linked to a workplace, with additional indicators such as commuting distance, occupation, and economic activity also available. This modelling approach enables new forms of analysis, including small-area labour force estimation, sub-municipal economic profiling, and evaluation of policy impacts in regional development and mobility planning. In addition, the model provides a reliable framework to support microsimulation and experimental statistics initiatives.

Adapting a small area estimation model for non-probability sample data integration

***Burakauskaitė, I.**, Vilnius University & Statistics Lithuania; Čiginas, A., Vilnius University & Statistics Lithuania; Šlevinskas, D., Statistics Lithuania*

The increasing demand for more detailed, timely, and accurate statistics – along with efforts to reduce respondent burden and address declining response rates in statistical surveys – has led to growing interest in data integration research. Often, the sample size of a statistical survey is too small to obtain reliable estimates for parameters in population domains. To improve the accuracy of these estimates, small area estimation models can be employed. Moreover, with expanding access to a wider range of alternative data sources – such as administrative records, social media, and web-scraped data – these sources can be used to improve estimation. However, due to the non-probabilistic nature of such data, which may lead to biased estimates, proper integration is essential. This presentation introduces a modification of the Fay-Herriot small area estimation model when non-probability sample data is incorporated as auxiliary information, accompanied by real data examples to illustrate its application.

When non-response makes estimates from a census a small area estimation problem: The case of the survey on graduates' employment status in Italy

***Ranalli, M.G.**, University of Perugia; Pennoni, F., University of Milan Bicocca; Bartolucci, F., University of Perugia; Mira, A., Università della Svizzera Italiana & University of Insubria*

Since 1998, AlmaLaurea — a consortium of 80 Italian universities and a member of the Italian National Statistical System — has conducted an annual census on graduates' employment status. The survey provides estimates of descriptive indicators at both the population level and for specific subpopulations (domains) of interest, such as degree programmes. Some domains have very few observations due to a small population size and non-response. In this paper, we address this estimation problem within a Small Area Estimation framework. Specifically, we propose using generalized linear mixed models that incorporate two variables as proxies for graduates' response propensity, making the assumption of non-informative non-response more plausible. Degree programme estimates of employment rates are derived as (semi-parametric) empirical best predictions using a finite mixture of logistic regression models, with their mean squared error estimated via a second-order, bias-corrected, analytical estimator. Sensitivity analysis is conducted to assess the explanatory power of variables modelling response propensity and to evaluate potential correlations between area-specific random effects and observed heterogeneity. To a broader extent, the problem addressed in the talk can also be interpreted as a big data integration problem, specifically related to a large non-probability self-selected sample where the variable of interest is surveyed, whereas covariates are available on both the non-probability sample and the population. Even though the observations in the non-probability sample are not necessarily representative of the target population, the association among variables in the non-probability sample is employed to develop a predictive model for mass imputation. Thus, the non-probability sample is used as training data for developing a model for mass imputation. In this regard, the proposed approach based on Small Area Estimation can be considered as an imputation method under a finite mixture of logistic regression models.

Integrating datasets from finite populations: A focus on registries and self-selected surveys

***Nicolussi, F.**, Politecnico di Milano; Masci, C., University of Milan; Bertarelli, G., University of Venice Ca' Foscari; Terzera, L., University of Milan-Bicocca; Mecatti, F., University of Milan-Bicocca*

The integration of datasets collected from finite populations remains a significant challenge in statistical and data science research. While many studies have explored methods for data integration, relatively little attention has been paid to the specific case of combining information from a complete registry—where the entire target population is covered—with data from a self-selected survey. This issue is increasingly relevant due to the growing availability of data derived from self-selection mechanisms, where individuals choose whether or not to participate.

Merging data from a self-selected survey with a comprehensive registry can offer substantial benefits. First and foremost, it allows for the enrichment of registry data with additional information collected in the survey. This can be particularly useful when the survey contains variables or details not recorded in the registry. Secondly, the integrated data can be used to assess and validate assumptions regarding the unknown sampling design of the self-selected survey. For instance, through the construction and analysis of a probabilistic decision tree, one can model the sample design, address biases, and test whether the sampling mechanism is informative.

In this work, we start from a complete registry and focus on a self-selected survey that targets an unknown subpopulation. The main objective is to identify and characterize this subpopulation in terms of its size and composition. Conceptually, we treat an individual's membership in this subpopulation as a partially observed binary variable. One natural approach is to adopt a supervised learning framework, using the observed subpopulation data from the survey to train a model. This model can then be applied to the rest of the registry to predict subpopulation membership for those with missing values. However, the effectiveness of this approach is often limited by the typically small size of self-selected samples compared to the entire population.

To address this limitation, we propose an unsupervised method. Instead of relying entirely on labeled data, we aim to impute the missing binary variable by leveraging the partial information provided by the survey. Our approach is based on the assumption that membership in the subpopulation induces a dependency among observations. When modeled using a regression framework, this dependency results in heteroskedasticity of the errors—meaning that the variance of the error terms varies across subgroups.

To exploit this property, we implement an algorithm that seeks to maximize the variance between groups, given an initial allocation of individuals. In this way, we attempt to uncover the hidden structure of the subpopulation without requiring full supervision.

Validation in unsupervised settings presents its own set of challenges. In our case, we propose a validation strategy that involves withholding portions of the available data and analyzing how the algorithm assigns these units to different groups. This provides a practical way to assess the model's effectiveness in detecting the latent subpopulation, even in the absence of complete ground truth.

Small area estimation using area level linked data

***Chambers, R.**, Australian National University; Salvati, N., University of Pisa; Fabrizi, E., Università Cattolica del Sacro Cuore*

Small area estimation methods overcome the limitation of small sample sizes for population subgroup inference, by using models to integrate survey data and auxiliary information from alternative sources such as administrative registers. Many models have been proposed for this integration, but here we focus on one of the oldest, and arguably the most widely used. This is the Fay-Herriot model, which is based on integration of the survey and the auxiliary information at subgroup level. In particular, it assumes that the only data available for inference about the subgroup averages of a target variable are their corresponding sample estimates, together with contextual subgroup information derived from the auxiliary data sources.

In an increasing number of cases, however, sample estimates for the target variable are not available. Instead, what is available are corresponding estimates for a linked version of the target variable - i.e., estimates based on sample data from a population created by linking the units in a register containing the variable of interest with those in another register containing the auxiliary variables that underpin the contextual information. Furthermore, this linkage is error-prone, so that the available subgroup estimates contain a mix of sample error and measurement error due to linkage errors. Standard Fay-Herriot methods allow for the former but not the latter, so linkage errors impact on predictions based on these methods. In particular, since units from different areas can be incorrectly linked, subgroup estimates based on these data are biased, leading to biased estimation of the Fay-Herriot model parameters, typically observed as an attenuation effect.

In this presentation we use the missing information principle to derive a modified Fay-Herriot predictor that corrects for these effects. We also derive the mean squared error of this modified predictor and develop an estimator for it. Moreover, we apply these ideas to a simulation study that compares the proposed modified Fay-Herriot approach with application of the original Fay-Herriot approach that ignores linkage errors. Finally, we illustrate the practical value of this approach with an application to the estimation of equivalised income averages for Labour Market Areas in central Italy. The estimates are obtained by linking the 2016 EU-SILC survey data for Italy with the Italian Integrated Archive of Economic and Demographic Microdata. Our results show that the modified Fay-Herriot approach improves estimation accuracy by mitigating bias arising from linkage errors.

Conformalized bayesian prediction in official statistics

Deliu, N., Sapienza University of Rome; ***Liseo, B.****, Sapienza University of Rome*

Official Statistics are experiencing an important renovation time. There is a need to exploit the potential that the digital revolution has made available in terms of data. However, this process occurred together with a progressive deterioration of the quality of classical sample surveys, due to an increasing rate of missing responses. The switch from survey-based inference to a hybrid system involving register-based information has made more stringent the debate and the possible resolution of the design-based versus model-based controversy. In this new framework, statistical techniques that provide exact coverage guarantees to model-based procedures are essential. Here we explore using the relatively new technique of conformal prediction. Although already popular in statistical learning, the potential of conformal prediction is still underexplored in official statistics. We argue that conformal prediction can be beneficial both in design-based and model-based approaches to survey sampling, and it is particularly suited for small area estimation (SAE).

Assessing uncertainty when integrating Mobile Network Operator data into official statistics with some transfer learning methods

***Tuoto, T.**, ISTAT; Zhang, L.-C., Southampton University & Statistics Norway*

The integration of new data sources into official statistics poses new challenges for statistical agencies, from the data access, with legal, privacy and partnership issues in case of privately held data, to methodological challenges, in identifying integration methods to combine new and traditional data sources to exploit the potential of both and in assessing the quality of the results.

In this paper, we consider mobile network operator (MNO) data and the contribution they can provide to official statistics as an example, exploring transfer learning methods to integrate the new data source with traditional surveys and statistical registers.

MNO data indeed seem very promising as source of information on human presence and mobility. In this paper we propose applications related to commuting and trips statistics. Despite the widespread use of mobile devices among the population and the large land coverage of mobile phone networks, the production of official statistics from MNO data must necessarily take into account the fact that the observation units, the devices, are only a proxy for the units of interest, people, and the correspondence between them is not 1:1. It is therefore unlikely that MNO data alone suffice to produce official statistics, hence it is necessary to supplement them with other relevant data (such as surveys) with respect to which, however, MNO data allow for increased spatial granularity and timeliness.

Since MNO data can be considered a good proxy for the target information when it comes to mobility, we propose to use transfer learning methods to enable combining sample surveys with MNO data, noting that an advantage of transfer learning compared to other integration methods is that it does not require mobile device usage data at the person level. Pan and Yang (2009) classify different settings of transfer learning. We explore first the use of inductive learning. In this setting, some labelled data in the target domain (the survey) are used as training data to 'induce' the target predictive function, given there are a lot more labelled data in the source domain (the MNO data). In addition, to make prediction for population domains not observed in the sample, transductive learning can be developed where labelled data do not exist in the target domain.

To assess the uncertainty, we propose design-based MSE estimation for inductive learning. Meanwhile, assumptions beyond the sampling design are necessary for transductive learning where, by definition, we do not have any target observations. In the same spirit as conformal inference, we shall only admit exchangeability assumptions in addition to the sampling design.

Accuracy measures for text mining methods applied to public administration PIAOs

***Righi, P.**, ISTAT; Bianchi, G, ISTAT; Giubilei, R., Food and Agriculture Organization of the United Nations; Naccarato, A., Roma Tre University*

The Italian National Statistics Institute (ISTAT) has been working with the Department of Public Administration (DFP) on a National Recovery and Resilience Plan (PNRR) project aimed at assessing the impact of reforms on Italian Public Administrations (PAs), administrative procedures and staff-related activities.

The Italian National Statistics Institute (ISTAT) has been working with the Department of Public Administration (DFP) on a National Recovery and Resilience Plan (PNRR) project aimed at assessing the impact of reforms on Italian public administrations (PAs), administrative procedures and staff-related activities.

Among its contributions, ISTAT has developed advanced tools for automated text analysis of the Integrated Plan of Activities and Organisation (PIAO) documents, using machine learning (ML) techniques. The resulting procedure identifies the sections and subsections of each PIAO and analyses their text to check whether the PAs have planned or implemented a set of strategic objectives, denoted as target variables. The output of the ML procedure consists of predicted values for these targets. Specifically, from the population of 4953 PIAOs, a stratified random sample of 927 PIAOs has been drawn, with strata defined by PA size class, geographical area and type of PA (municipality, non-municipality).

Each sampled PIAO underwent expert validation, where domain experts verified the presence of strategic goals. The sample evidence served three purposes: to train the supervised ML model, to observe the prediction errors of the ML model, and to adjust the estimator based on the predictions over the 4953 PIAOs, taking into account the prediction errors. Hereinafter, we refer to the adjusted estimator as the pseudo-calibration (PC) estimator. The PC estimator is particularly appropriate when the target variable is observed in the probability sample and predicted or observed with error in the non-probability sample.

The paper focuses on evaluating the accuracy of the ML predictions for the specific sample and the accuracy of the PC estimator with respect to the 4953 PIAOs. The PC estimator is compared with a standard design-based estimator that uses the PIAO sample from which the PA expert extracts the value of the target variables.

Resampling methods were used for variance estimation, allowing the efficiency and robustness of the estimators to be assessed.

The accuracy results of this innovative estimator encourage the use of PIAOs for the production of public administration statistics. Based on these results, the project will carry out two further steps: i) extend the results of the PC estimator to the population of 12,000 public administrations: ii) define a common framework for PIAOs that is suitable for administrative purposes and for having as small a prediction error as possible when using automatic text analysis.

An interactive tool for statistical matching

Guastadisegni, L.*, University of Bologna; Trivisano, C., University of Bologna*

Statistical matching is a technique used to integrate multiple data sources that have different units but share a set of common variables. However, the application of statistical matching methods poses several challenges for practitioners, including the complexities of data preprocessing and variable harmonization, as well as the application of appropriate matching methods, which can be at the micro or macro level. To streamline this process, we have developed an R-Shiny application designed to facilitate data integration from different surveys. Through a demo example, we provide a comprehensive overview of the application's features, demonstrating how it simplifies the workflow and enhances usability for practitioners. Indeed, it offers a user-friendly interface with various functionalities, such as selecting relevant information from existing surveys, performing data wrangling, choosing matching variables, applying different statistical matching methods, and downloading the matched dataset.

Integrating social survey data: A comparative evaluation of statistical matching techniques

Fontanella, L., *University of Chieti-Pescara "G. D'Annunzio"; Cucco, A., University of Chieti-Pescara "G. D'Annunzio"; Aretusi, G., University of Chieti-Pescara "G. D'Annunzio"; del Gobbo, E., University of Foggia; Sarra, A., University of Chieti-Pescara "G. D'Annunzio"*

Social scientists today have access to extensive data collected through large-scale and cross-national surveys. These datasets offer the opportunity to conduct in-depth analyses of the complex interplay between social and psychological factors influencing individual and societal changes.

However, as these data are often collected from different statistical units, integrating insights from multiple datasets is an essential step to enhance analytical potential. Among the methods designed to achieve this, Statistical Matching (SM) stands out as a valuable tool for combining data from distinct, independent surveys targeting the same population. By connecting different domains of inquiry, SM enhances the complementary use of existing datasets, expands variable coverage, and mitigates the limitations of single-source data.

Despite its advantages, integrating social surveys poses several challenges. These include data heterogeneity, the complexity of survey designs, the selection of common matching variables, and the handling of composite scores. This study explores SM as a micro-level data integration technique and conducts a comparative evaluation of various SM approaches. Our primary objective is to assess and compare parametric, non-parametric, mixed, Bayesian, and machine learning-based SM techniques in terms of matching accuracy and the preservation of joint distributions across integrated datasets.

As a case study, we focus on integrating two prominent European surveys: the European Social Survey (ESS) Round 10 (2020) and Eurobarometer 93.1 (2020). These surveys were selected due to their rigorous methodologies, robust sampling designs, and relevance in exploring key social issues such as immigration, political engagement, and public opinion. Both datasets provide free access to microdata for scientific research, with user-friendly formats (SPSS) and comprehensive documentation that facilitate harmonization and advanced analyses using statistical software like R. Our integration process involves several key steps. First, we harmonize shared variables to ensure consistency across datasets. We then apply and compare the performance of different SM methodologies. Parametric approaches, such as regression-based matching, rely on specific distributional assumptions. Non-parametric methods, like nearest-neighbor matching, provide flexibility by avoiding predefined functional forms. Mixed methods combine parametric and non-parametric elements, while Bayesian approaches integrate prior information to model joint distributions more flexibly. Finally, we evaluate machine learning matching techniques.

Our findings offer insights into the strengths, limitations, and practical trade-offs of each approach. By comparing SM techniques, this investigation offers helpful advice to researchers seeking to enhance their analytical capacity through data integration.

Funding: This research has received financial support from the ENCLOSE project, selected as part of the Research and Innovation Program "Extended Partnerships involving universities, research centers, and companies for the funding of basic research projects" (MUR notice no. 341/2022), under the "GRINS – Growing Resilient, Inclusive, and Sustainable" project [PE0000018]. The project is financed within the framework of the PNRR, with funding from the European Union–NextGenerationEU.

On the sampling distribution of the ordinary least square estimator in parametric statistical matching

***Narcisi, M.**, University of Bologna*

Effective decision-making increasingly depends on timely and detailed information, often at fine spatial and temporal scales. However, traditional survey data often fall short of these requirements due to high costs, lengthy implementation times, and limitations in questionnaire design that increase response burden and compromise data quality.

To overcome these constraints, integrating multiple data sources has become a cornerstone of modern empirical research, especially when key variables are not jointly observed. Statistical matching techniques are widely used to merge separate datasets using shared variables. While conceptually appealing and frequently applied in practice, these methods pose serious methodological challenges, particularly when the merged data are used for regression analysis.

Hirukawa and Prokhorov (2018) showed that under such conditions, the ordinary least squares estimator is generally inconsistent. The extent of the bias depends primarily on the relationship between the matching variables and the relative sample sizes of the datasets. In response, the authors proposed two bias-corrected semiparametric estimators that achieve consistency under certain conditions.

Building on this foundation, the present study investigates regression-based imputation to address missing covariates. Matching variables are observed in both datasets; the covariate to be imputed is observed only in the donor, while the outcome is observed only in the recipient, which also includes additional covariates relevant for the regression. We consider two approaches based on parametric matching methods: conditional mean matching, which uses the expected value given the matching variables, and prediction matching, which adds stochastic noise to capture conditional variance. The resulting completed dataset is used to estimate the target regression of the outcome on both observed and imputed covariate.

In this work, we derive closed-form expressions for the conditional and marginal expectations of the ordinary least squares estimator under regression-based imputation, leveraging on the theory concerning quadratic forms in Gaussian random variables.

Bayesian integration of different data sources: Balancing cost, bias and measurement

***Salvatore, C.**, Utrecht University; Sakshaug, J., IAB & LMU-Munich; Wisniowski, A., Manchester University; Struminskaya, B., Utrecht University; Biffignandi, S., University of Bergamo*

In survey research there is an ongoing transformation, with a growing range of data collection strategies available. Traditional probability sample surveys, long considered the gold standard for inference, are facing declining response rates and rising costs. In response, researchers are increasingly turning to less expensive, but potentially biased, non-probability sample surveys. At the same time, interest is growing in sensor-based measurements (e.g., wearables, biological specimens) to overcome measurement issues related to self-reports (e.g. recall, social desirability).

These developments raise a central question: how can we integrate data sources with differing quality characteristics (e.g. selection bias, measurement error, timeliness, and relevance) to draw valid conclusions about a target population?

In this talk, we present two case studies that demonstrate how combining data sources can lead to improved inference for regression coefficients and more cost-effective study designs. We introduce a Bayesian framework that enables dynamic borrowing of information across data sources, based on the degree of similarity between them.

We apply the method to two scenarios:

1. Integrating parallel web-based probability and non-probability samples, with a focus on adjusting for selection bias; and

2. Combining two probability samples, a small sample collecting objective health measurements to contain costs, and one larger sample, relying on self-reports of the same variable.

In both applications, we show that, under certain conditions, the Bayesian integration approach can yield gains in efficiency (e.g., lower mean squared error of regression coefficients) and substantial cost savings, in some cases up to 80%. For survey practitioners, this method provides a systematic, data-driven strategy for integrating diverse sources in a way that complements probability sampling, rather than replacing it. It can be particularly beneficial in studies with constrained budgets or small sample sizes, where significant gains in efficiency can be achieved.

We conclude by discussing future extensions of this approach to different goals (e.g., alternative regression models, estimation of population parameters) and to cases where both selection bias and measurement error are present.

Make valid and reliable inferences about technological innovation and Industry 4.0 of firms in Tuscany using data from web scraping

Braito, L., *University of Florence; Rocco, E., University of Florence; Lodetti, P., Baloon srl*

This study investigates firms' behaviour regarding technological innovation and Industry 4.0 adoption in Tuscany, using a combination of traditional and innovative data sources. While the role of technological innovation in automation and digitalisation has received increasing attention, detailed and representative data on firm-level behaviours are often lacking, particularly at regional or sectoral levels. This study gives particular attention to the distinction between innovation propensity — defined as a firm's intention, planning, or declared openness toward innovation — and actual innovation performance, which is reflected in the effective use of enabling technologies, digital tools, and automation systems. Traditional business surveys based on probability sampling provide reliable firm-level data but may often miss direct indicators of innovation-related activities. Conversely, emerging sources like web scraping enable large-scale data collection on indirect but tangible signals of technological adoption — such as references to advanced machinery, digital services, or Industry 4.0 terminology. However, estimates derived solely from such data may suffer from significant selection bias.

To address this, we explore how non-probability web scraping data can be integrated with traditional survey data, under the assumption that both sources share auxiliary variables and refer to the same target population. In our study, the reference population is defined as a subset of firms included in the ORBIS database. Non-probability data are collected from firms' websites and social media, and indicators based on keywords related to innovation, digitalisation, and automation are constructed. A smaller, stratified probability sample is drawn from the same population.

When working with non-probability data, producing unbiased estimates of key parameters — such as the population mean — requires satisfying two key conditions: first, that all relevant auxiliary variables are observed (i.e., the Missing at Random assumption holds), and second, that the statistical models used to adjust for these variables are correctly specified. Three common adjustment approaches involve modelling the outcome, modelling the inclusion probabilities or combining both. Doubly robust estimation methods, in particular, have gained attention in this field due to their resilience to partial model misspecification. The combination of these models has been explored through various approaches, and we focus on four main methods, prominently discussed in the literature on causal inference and missing data. In this context, we review and adapt the following methods: mass imputation with residual bias correction, also known as augmented inverse propensity weighting (AIPW), calibration and augmented calibration estimators, mass imputation with inverse-propensity weighted coefficients, and mass imputation with propensity-based covariates. We adopt one of these approaches to estimate key innovation-related outcomes, selecting the most suitable one also in light of a previous study based on a simulation. Our findings highlight the value of integrating traditional and non-traditional data sources to obtain more comprehensive and externally valid insights into firm behaviours. The results underscore the crucial role of auxiliary information — particularly from probability-based surveys — in correcting for biases and improving inference quality when working with non-probability big data sources such as web scraping.

Improving the sampling strategy for the Community Innovation Survey using machine learning algorithms

***Klingwort, J.**, Statistics Netherlands; van Berkel, K., Statistics Netherlands; van den Brakel, J., Statistics Netherlands*

National statistical institutes (NSI's) are increasingly interested in using non-probability data to produce official statistics. Examples are information on the internet, social media messages, sensor data, and web-scraped data. Relying on this kind of data sources implies an increased impairment risk for producing official statistics. Important risk factors are selection bias, lack of control over the minimum required precision of the statistical output, data availability, incomparability over time, and less optimal operationalization of concepts to be measured.

This paper proposes using information extracted from these kinds of data sources to improve the sampling strategy of a probability sample. In this way, this kind of new data sources are used in the widely applied design-based inference framework, which is predominantly used by NSI's to produce official statistics. Compared to those that use these new data sources as primary sources for compiling official statistics, the key advantage of this approach is the minimization of impairment risk. This concept is illustrated with an application to the Community Innovation Survey (CIS).

The CIS collects data on business innovation, relying on efficient sampling and weighting to ensure accurate population estimates. The currently used Horvitz-Thompson estimator (HT) may not fully capture the complexity of the Dutch business landscape. With the availability of auxiliary data, there is an opportunity to refine the CIS weighting strategy to improve the accuracy of survey estimates.

This study analyzes how incorporating auxiliary information from administrative data sources and data obtained with web-scraping can improve the CIS weighting strategy, aiming to improve the precision of survey estimates. Nine weighting models were specified next to the HT. Each weighting model is used to estimate the five target variables.

Three auxiliary datasets are utilized for the weighting models: (1) probabilities on innovation derived from web-scraped business data, (2) administrative records of businesses receiving R\&D subsidies, and (3) business patent data. The probabilities for source (1) are derived using a combination of TF-IDF, a bag of words, and logistic regression. In addition, a large language model has also been used to derive probabilities from the web-scraped data.

The generalized regression estimator (GREG) is applied to assess the impact of auxiliary data on survey estimate accuracy. The performance of the GREG is compared against the existing HT estimator.

The key finding is that machine-learning-based probabilities, intended to indicate a company's innovativeness, do not correlate with the survey target variables. These language models may learn patterns but do these patterns not necessarily indicate true innovation. This misalignment could stem from the models being unsuitable for the task or the textual data lacking useful signals. The scraped content often contains single words and no complete sentences, which may make it difficult to extract useful signals. Whether scraped website content contains useful signals and information should be subject to further research. In contrast, administrative data shows the greatest potential to improve the sampling strategy.

This study contributes to the discussion on integrating new data sources in official statistics, providing insights for statisticians and policymakers seeking to modernize survey methodologies.

Statistical frameworks for data integration and beyond

***Wu, C.**, University of Waterloo*

We provide some general discussions on statistical frameworks for data integration involving probability and non-probability survey samples and data from other sources. Two main principles, namely, validity and efficiency, for methodological developments are discussed under different scenarios for data integration. The usefulness of artificial intelligence as a tool for survey data collection is briefly discussed, and the recent research topic on prediction-powered inference using machine learning techniques is reviewed under the model-assisted framework for survey sampling.

Beyond string matching: Embeddings and artificial intelligence for statistical integration

Bruno, M., *ISTAT*

In official statistics, the ability to identify and connect related information across different sources is a long-standing challenge. Traditionally, matching techniques relied on exact string comparisons or rule-based probabilistic models. Today, recent advances in semantic methods — based on vector representations of words and sentences — offers powerful alternatives that can capture meaning, not just surface similarity. This presentation traces the evolution of matching approaches in statistical institutes, from deterministic and probabilistic string matching to modern sentence embeddings used in semantic search and the classification of non-traditional data sources. Real use cases from ISTAT illustrate how these techniques are being integrated into statistical workflows, offering new opportunities for automation, quality improvement, and innovation.

Integrating big data and alternative sources into small area estimation

Pratesi, M., *University of Pisa*

As data needs become more granular and policy demands more timely and localized statistics, Small Area Estimation (SAE) is evolving to incorporate big data and alternative data sources. Traditional SAE approaches have relied heavily on survey and census data. However, the growing availability of non-traditional sources — including administrative records, satellite imagery, mobile phone data, and citizen-generated data — offers new opportunities to enhance estimation accuracy and relevance at small geographic or demographic scales.

This presentation explores the integration of big and alternative data into SAE models, focusing on methodological innovations, practical challenges, and real-world applications. We discuss how these data sources can be used as auxiliary variables, how to address issues of representativeness and bias, and how to balance model complexity with interpretability.

Future possibilities open to AI-powered Small Area Estimation (SAE) that integrates survey data with big data sources (e.g., satellite images, mobile phone data). Would Real-Time Poverty Mapping be possible? imagining AI that combines geospatial and survey data to model income levels and social indicators in small regions.

Emphasis is placed on ensuring data quality, ethical use, and transparency, with the goal of producing reliable, actionable statistics for unplanned domains and underserved populations.

Confidentiality and differential privacy in the dissemination of frequency tables

Shlomo, N., *University of Manchester, Rinott, Y., Hebrew University, O'Keefe, C., CSIRO Australia, Skinner, C., The London School of Economics and Political Science*

For decades, National Statistical Institutes have been publishing frequency tables containing whole population counts from censuses. To protect the confidentiality of individuals, census tables are modified before release. In response to user demand for more flexible and responsive table publication services, frequency table publication schemes have been augmented with on-line table generating servers, such as the Australian Bureau of Statistics (ABS) TableBuilder. These systems allow users to build their own custom tables and make use of automated perturbation routines to protect confidentiality.

Motivated by the growing popularity of table generating servers, we assess the confidentiality protection for perturbed frequency tables with respect to the differential privacy standard. We focus on a version of the ABS TableBuilder as a concrete example of a data release mechanism and examine its properties. We show that the 'same participants-same perturbation' scheme of the ABS TableBuilder results in a non-interactive differential privacy mechanism that can be useful for data dissemination at an NSI. We compare alternative perturbation mechanisms and compare the utility of the resulting tables for a given level of confidentiality protection. Findings show that the differentially private discrete Laplace perturbation for census counts have clear advantages in terms of the utility.

We note the advantages of differential privacy with respect to making the perturbation mechanism and its parameters available to users and the possibility that users could take account of this knowledge when analysing the data. Thus, in principle, given a specified model for the data and a perturbation mechanism, it may be feasible to determine a likelihood function

for the perturbed data, and make inference on the parameters of the data model. We demonstrate this procedure in a simple example.

This is joint work with: Yosi Rinott, Christine O'keefe and Chris Skinner, and was funded by the Simons Foundation and the Isaac Newton Institute for Mathematical Sciences, University of Cambridge during the Data Linkage and Anonymisation programme (EPSRC grant no EP/K032208/1).

Emphasis is placed on ensuring data quality, ethical use, and transparency, with the goal of producing reliable, actionable statistics for unplanned domains and underserved populations.

On the formal privacy guarantees of synthetic data (generated without formal privacy guarantees)

Neunhoeffer, M., *Institute for Employment Research & Ludwig-Maximilians-Universität München; Seeman, J., Urban Institute & University of Michigan; Drechsler, J., Institute for Employment Research & Ludwig-Maximilians-Universität München & University of Maryland*

What privacy guarantees can synthetic data satisfy even without formal guarantees during the training of the synthesizer? In this paper, we explore this question using a synthesizer under simplified settings to show that the privacy guarantees offered by this synthesizer and potential others can be directly translated into a formal privacy guarantee, specifically a ρ-zero Concentrated Differential Privacy (zCDP) guarantee.

We further explore the conditions under which this equivalence holds and show that getting formal privacy guarantees for more realistic synthetic data models is significantly harder.

PrivGen: A human-in-the-loop approach to generate private synthetic data

***Nobani, N.**, University of Milan-Bicocca; Officioso, G., University of Milan-Bicocca; Mezzanzanica, M,. University of Milan-Bicocca; Sperlì, G., University of Naples Federico II; Mercorio, F,.University of Milan-Bicocca*

Synthetic data is a powerful tool for sharing sensitive information while protecting privacy and enabling research. However, even with the latest generative models, synthetic data can still reveal patterns from the original datasets—especially when there are outliers or the data is sparse. Techniques like differential privacy help, but they often fall short of what domain experts expect, particularly in real-world, messy datasets.

In this talk, I'll introduce PrivGen, a human-in-the-loop approach to making synthetic data safer. PrivGen brings expert knowledge into the preprocessing phase, focusing on smarter outlier detection and cleaning. By addressing potential privacy leaks before the generation phase even starts, we can produce synthetic data that is less tied to the original sensitive records without sacrificing too much quality.

I'll walk you through how we tested PrivGen: we applied it across nine top generative models on three public datasets, and measured the results using seventeen different metrics—covering privacy, data quality, and sanity checks. The findings are encouraging: for sparse datasets especially, PrivGen helped boost privacy significantly, while keeping any negative impacts on data quality relatively low.

Stratifying the second stage sample units via marginal allocation: an innovative two-stage sampling design for ISTAT social surveys

***Asti, L.**, ISTAT; De Vitiis, C., ISTAT; Inglese, F., ISTAT; Terribili, M., ISTAT; Trambusti, D., ISTAT*

In ISTAT social surveys utilizing two-stage sample designs (municipalities and households or individuals), the second-stage units are usually not stratified. However, to enhance the quality of final estimates for target domains defined by individual characteristics, the stratification of secondary sampling units becomes necessary and complex methods (apparently not explored in literature) must be established to integrate the stratification of both primary and secondary sampling units (PSUs and SSUs).

The planning of the last editions of the ISTAT sampling surveys on "Families, social subjects and life cycle" and on "Discriminations" required the study of a two-stage sampling design with stratification of both stages: municipalities (PSUs) are stratified according to the geographical variables, while individuals (SSUs) are stratified also by "profiles" (marital status, citizenship and age classes).

In this context, the total planned sample size is first distributed across strata defined simultaneously on geographical and individual variables, according to an optimal allocation. Next, the PSUs are selected together with the related sample sizes in terms of SSUs, according to a self-weighting selection scheme starting from the optimal allocation aggregated at geographical level. Finally the municipal sample sizes are distributed on an allocation matrix where rows represent municipalities and columns represent individual profiles; the number of final units to select from each cell of the matrix is computed constraining, for each geographical domain, row and column sums of the matrix on the municipality sample sizes and on profile marginal distribution of the optimal allocation, thus avoiding to select SSUs from each profile in each PSU.

To define the entire allocation matrix from its marginal sums, we explored two strategies: distributing the sample following Cochran algorithm (used to reduce strata when fine cross-classification produces too many strata), distributing the PSU sample sizes according to profile optimal allocation proportions. Depending on the strategy, the required rounding and oversampling steps are performed at different stages of the design pipeline, dealing also with lack of population units or privacy constraints while preserving at the same time marginal allocations.

In addition to the allocation strategies implemented for the two surveys, as a study for future similar survey designs a balanced sampling design has been tested constraining the selection to respect the marginal allocations and based on probabilities derived from profile optimal allocation. This scheme could allow avoiding the definition of rounded sample sizes for random selection at the municipality-profile level. The comparison between the different strategies is carried out in terms of weight variability and possibly with sample replication methods.

The current applications of two-stage stratified sampling offer a beneficial approach to social surveys on individuals and are increasingly becoming a common need in National Statistical Institutes. Such designs, although they may involve an increase in variability of sampling weights, which must be taken into account in the estimation phase, are based on efficient sample allocation schemes, which facilitate the data collection phase by preventing the sample from being dispersed across an excessively large number of municipalities, while ensuring effectiveness in representing the diversity of the subpopulations.

Evidence supporting the design of the italian integrated system of social surveys

***Guandalini, A.**, ISTAT; Loriga, S., ISTAT; Terribili, M.D., ISTAT*

In recent years, the integration of data from multiple sources for statistical production has garnered increasing international attention. Rising nonresponse rates have made sample surveys more expensive and potentially less representative, prompting greater use on other sources of data. Firstly, administrative data, which are the focus of significant investment by NSIs in terms of data integration and processing, aimed at their use for statistical purposes. Such data have long been used to improve direct survey estimates by using statistical techniques that incorporate external sources – particularly demographic administrative data – as auxiliary information. These include well-known methods such as calibration of survey weights and small area estimation (SAE), which enhance efficiency and provide more granular statistics without increasing respondent burden.

As the demand for more detailed and timely information grew, new data sources – such as big data and other low-cost, non-probabilistic sources – gained relevance. This marked the beginning of a new phase focused on developing data production models capable of meeting the five challenges identified: Wider, Deeper, Better, Quicker, and Cheaper. Although probability-based surveys remain essential, interest in using alternative sources as primary data inputs or even auxiliary information is growing. Also a "holistic" approach, selecting among various techniques – probability sampling, record linkage, imputation, multi-frame sampling, SAE, and Bayesian models – depending on the context and data quality, has been proposed. The innovation lies not in the techniques themselves – many are long established – but in how they are selected, combined, and adapted.

Applying multi-source methodologies to entire statistical ecosystems has led to the development of complex infrastructures of integrated microdata and indicators, supporting the modernization of official statistics. In Italy, ISTAT's modernization strategy includes the Integrated Register System (IRS) – which uses administrative data as a primary source – and the Permanent Censuses, which combine administrative records with survey data.

The SICIS (Integrated Census and Social Surveys System) project fits within this context – a long-running initiative, focused on designing a harmonized and integrated system of social surveys. SICIS aims to improve consistency, accuracy, and granularity through coordinated sampling strategies and the use of administrative registers.

Currently, two sampling strategies have been identified as central in the SICIS project: the two-phase sampling design and spatially balanced sampling. Two-phase sampling is a strategy to integrate the samples of traditional household surveys with the large-scale sample selected for the Population Census. In practice, the survey units are selected as sub-samples of the units drawn for the Census survey. This can be applied to select the municipalities or both municipalities and households. Spatially balanced sampling addresses inefficiencies from traditional stratification and enables more effective municipality selection.

This paper presents evidence from a simulation study on real data to guide the optimal implementation of the two-phase sampling design within this context. It evaluates the combined effects of two-phase and spatially balanced sampling and explores outcomes under different nonresponse mechanisms.

Some applications of optimality principles to sample selection from a register

***Tillé, Y.**, University of Neuchatel*

Modern survey methodology increasingly leverages auxiliary information from population registers to enhance both estimation and sample design. This presentation explores how principles of optimality — particularly anticipated variance minimization — can guide the selection of samples that are both efficient and theoretically grounded. Building upon classical results such as Neyman optimal stratification, we examine model-based frameworks that justify the use of unequal probability sampling designs and balanced sampling designs. We present general results linking statistical modeling assumptions with the structure of optimal sampling strategies, and highlight the importance of spreading and balancing techniques when autocorrelation or heteroscedasticity is present. Real-world examples, including medical record audits, biodiversity monitoring, rolling censuses, and register-based national surveys, illustrate how these principles are applied in practice. Emphasis is placed on translating model assumptions into implementable algorithms that respect operational constraints while optimizing inference quality.

Defining ad-hoc sampling designs for small area estimation

Falorsi, P. D., Sapienza University of Rome; Falorsi, S., ISTAT; ***Nardelli, V.****, Università Cattolica del Sacro Cuore; Righi, P., ISTAT*

The paper sketches a proper statistical setting necessary to define the sampling design for Small Area Estimation (SAE). Since SAE techniques are commonly used in official statistics, relying on appropriate sampling designs to improve the quality of estimates becomes crucial. The sampling design is based on both allocation and sampling selection. The allocation step solves a non-standard problem necessary for finding the minimum-cost solution that controls the accuracy of the model-based small area estimator. The sampling selection ensures the planned sample sizes for each level of random effects affecting the variables of interest.

Weighting survey data for respondent embeddedness in a recruitment environment: Rationale, experiences and challenges

Albert, E., *University of Vienna*

Research about and for hard-to-survey (H2S) populations frequently relies on nonprobability data. Among the options to improve the quality of such data, the weighting of respondents for their embeddedness in a recruitment environment – or e-weighting – has received growing attention. Experiences with the method have become possible following applications by the EU's Fundamental Rights Agency (FRA), in the form of affiliation weights for their LGBTI(Q) surveys since 2019. The presentation offers an introduction to the rationale, the assumptions and the requirements underlying e-weighting approaches. The central role of a chosen set of outreach activities to a target population is highlighted, as it is the basis of the method's two essential auxiliary variables: the indicator of embeddedness and the participation proxy. Among the reportable experiences with the method so far, examples of good performance can be provided regarding (a) the transferability to different country settings, (b) the strength of the required auxiliary variable associations, (c) the impact of weights on key outcomes, and (d) the ability to keep weights moderate, given an offered amount of bias adjustment. However, important challenges must also be discussed. While it is tempting to improve the indicator of embeddedness by extracting it as a metric (with a known measurement error) from several or even many auxiliary variables, each additional variable in the questionnaire will increase the burden on respondents. Also, even though it is predictable how different link functions supported by the method, different trimming decisions and possible corrections for measurement error will affect final weight distributions – it is currently less clear which criteria should be the most important ones when deciding on a final e-weighting scheme, in a particular research situation.

Strategies for dealing with digital inequality in self-completion survey designs – Evidence from Germany

Cornesse, C., *GESIS*

Like many countries across the world, Germany can increasingly be described as a digital society, in which people regularly use the internet in their everyday lives, for example for shopping, working, communicating, and entertainment. However, internet usage skills as well as the availability and quality of internet access are unequally distributed across the country, thus systematically restricting some people's possibilities of engagement. This digital inequality, which is often referred to as the "digital divide", affects the whole country, including its survey landscape, which is shifting more and more away from interviewer-administered survey modes towards the online mode of data collection. This trend promises a number of benefits, such as cost-efficient data collection and speedy availability of large amounts of survey data. Some groups of the population may even have a higher response propensity online, for example because they can nowadays conveniently fill out a survey on their train commute or while waiting in line at the supermarket. However, other subgroups may be left behind by this development. How can we balance the promises and pitfalls of the online survey data collection mode for social research? Which consequences can it have if we fail to successfully account for the digital divide? And will the problem solve itself over time with increasing levels of digitalization? In my presentation, I will provide a survey methodological overview of experimental and observational evidence on the impact of different offline population inclusion strategies across three panel survey studies and over ten years of survey data collection in Germany. I will emphasize the importance of contextualizing methodological evidence with a country's past, present, and future state of digitalization and derive suggestions for the future of survey data collection.

Statistical inference when data come from different sources: Perspectives, limitations, and pitfalls

***Conti, P. L.**, Sapienza University of Rome*

The Big Data era is characterized by large amounts of freely (or cheaply) available data, typically coming from different sources. The use of such data, a sort of golden mine, is tempting for statisticians. This is the multi-source statistical world, where each data source provides a part of the needed statistical information, in terms of variables and/or units. A major point, in this new multi-source world, is to combine together data come from different sources into a unique synthetic data file, as if it were produced by a single sample survey, even if it is not.

As a consequence, the traditional paradigm of statistical inference, based on an ad hoc sample survey with a well-defined, probabilistic sampling design, and providing all necessary data for statistical analysis, is challenged by the new multi-source approach. In this new era, data integration is the challenge. Although the concept of data integration takes on different shades depending on the adopted methodological perspective, roughly speaking it can be considered as falling into two main categories.

- Construction of new databases by combining together single ones, in order to have a broader domain of application.

- Making statistical inference from several databases simultaneously considered.

Among the main examples of data integration (or integrated use of statistical data), we may quote:

- Statistical matching, where two samples A, B of the same population are observed. In sample A, variables X, Y are observed, while in sample B variables X, Y are observed. The goal is to construct a statistical data base containing variables X, Y, Z.

- Record linkage, where records in different samples and pertaining to the same population unit have to be matched together.

- Non-probability samples, where the major technique to overcome the uncontrolled selection mechanism from the population consists in the integrated use with a parallel probability sample.

Although the multi-source paradigm is promising, it contains several potential troubles, that have to be fixed.

- Different data sources could not refer to the same population.

- Different data sources could refer to different units of the same population. For instance, this occurs in case of different samples from the same population.

- Even if two sources refer to the same units, it could be difficult to exactly identify records pertaining to the same unit.

- Variables from different sources could be observed with different quality levels, with different incidence of measurement errors.

- The data collection mechanism (sampling design, in many cases) could be controlled for some sources (probability samples) and uncontrolled for some other (non-probability samples).

- Identifiability of statistical models could be lost, with dramatic consequences in terms of properties of estimates and hypotheses tests. On the other hand, identifiability can be frequently recovered only through non-testable, very restrictive assumptions.

Data integration for mapping vegetation biodiversity

***Golini, N.**, University of Turin; Zoppi G., University of Turin; Lanteri, A., University of Turin; Cameletti, M., University of Bergamo; Ignaccolo, R., University of Turin; Lo Presti, A., University of Turin*

The SVeBio (Statistics for Vegetation Biodiversity: Estimation and Mapping)* project aims to provide innovative statistical methodologies for biodiversity assessment and monitoring. Vegetation is vital to biodiversity, forming a key element of ecosystem composition, function, and structure. As a result, biodiversity protection is closely linked to conserving vegetation diversity. Then, mapping the spatial distribution of habitats, biodiversity indices, and other relevant attributes is crucial for developing effective conservation strategies and addressing biodiversity loss. One of the goals of the project is to develop model-based methods to produce spatial maps for forest data from probability sampling surveys, purposive field campaigns (non-probability samples), and remote sensing information. This work considers forest data in the Tuscany region (Italy). In particular, we have data on the presence and absence of forest in this region, partitioned by 22,976 grid cells of 1km × 1km. Additionally, we have data on several forest attributes relevant to biodiversity, e.g., basal area, observed from a probabilistic sample surveyed on the ground in Tuscany in 2015 and purposive samples recorded between 2009 and 2021. Moreover, remote sensing data, such as slope, altitude, exposure and multi-band Landsat images, are extrapolated and harmonized into each dataset. In our framework, data integration refers to multiple data sources related to the response variable measured using two different sampling strategies. In particular, we consider a hierarchical spatial Bayesian model for integrating data from probability and non-probability samples accounting for preferential sampling (i.e., bias in sampling locations visited for presence/absence data). We assume that a common latent spatial field is shared across all hierarchies of the model. We perform a fast Bayesian inference using Integrated Nested Laplace Approximation (INLA) and the stochastic partial differential equation approach. By using a test set, we evaluate the estimation and predictive performance of our model.

* All the authors acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Statistics for vegetation biodiversity: estimation and mapping (SVeBio) - P2022AW4LX - CUP B53D23029510001 - Grant Assignment Decree No. 1378 adopted on 01.09.2023 by the Italian Ministry of Ministry of University and Research (MUR).

On admissibility in bipartite incidence graph sampling

***Zhang, L.C.**, University of Southampton & Statistisk Sentralbyrå*

The estimator of Horvitz and Thompson (1952, HTE) in finite population sampling is admissible in the class of unbiased estimators (Godambe and Joshi, 1965). Inbipartite incidence graph sampling (Zhang, 2021, 2022), short-handed as BIGS, the sampling units are distinguished from the target study units and there can exist more than one way by which a given study unit maybe observed via sampling units. This multiplicity of access (Birnbaum and Sirken, 1965) to a given study unit is a fundamental distinction between BIGS and finite-population element or multistage sampling, where each study unit can be sampled via one and only one primary sampling unit.

For example, sampling 'influencers' on a social media platform via an initial sample of users is a case of BIGS, where each user in the initial sample may lead to other users (possibly outside the initial sample) which she 'follows' and an influencer can be sampled via more than one user. Similarly, BIGS is the case if one samples webpages by following the links from an initial sample of webpages in multiple steps, where the study units are cliques of webpages with mutual links between any two webpages inside the same clique, while the initial sampling units are all the webpages.

BIGS operates formally in a bipartite simple digraph which have two sets of nodes representing sampling and study units, respectively, where directed edges only exist from the first set of nodes to the other,and each edge represents the incidence observation relationship between a sample unit and a study unit under the specified sampling method, regardless the actual operations and data structure involved or if the incidence relationships may be unknown in advance. To enable design-unbiased estimation of any study units total, we assume an initial probability sample from all the sampling units, and there is available a minimum level of knowledge of observation relationships from sampling units to the observed study units on the given occasion of BIGS, referred to as the ancestry knowledge (Zhang and Patone, 2017; Zhang, 2022). BIGS extends the scope of application by allowing study units such as net-works of individuals or clusters of contiguous habitats. It enables a unified treatment (Zhang, 2022) of the so-called 'non-standard' population sampling techniques as special cases of BIGS, such as multiplicity or indirect sampling, network sampling and adaptive cluster sampling, as well as various breadth-first or depth-first techniques explicitly devised for graph-structured data, such as snowball sampling and random-walk type sampling.

Provided ancestry knowledge in BIGS, a large family of unbiased incidence weighting estimators (IWEs) have been proposed (Zhang, 2022; Patone and Zhang, 2022), which includes the HTE as a special case as well as the many other estimators known in the sampling literature. There is thus a need to study admissibility in BIGS, since there may be other admissible estimators than the HTE which may or may not be IWEs.

We present our admissibility results in BIGS and explain how our results encompass the existing finite-population sampling situations mentioned above.

The role of participant understanding in data donation studies

***Boeschoten, L.**, Utrecht University; McCool, D., Utrecht University, Struminskaya, B., Utrecht University*

Problem

Recently, a workflow has been introduced that allows academic researchers to partner with individuals interested in donating their digital trace data for academic research purposes. In this workflow, the digital traces of participants are processed locally on their own devices in such a way that only the subset of participants' digital trace data that is of legitimate interest to a research project are shared with the researcher, which can only occur after the participant has provided their informed consent.

In the last years, multiple studies have been conducted that make use of research infrastructure that facilitates the use of this workflow. An important challenge in these data donation studies is that this is a relatively unknown and complicated procedure that needs to be clearly explained to participants. In this project, we aim to gain a better understanding of to what extent participants understand the data donation procedure. In addition, we investigate if a better understanding of the data donation procedure is related to higher participation rates.

Data

We use a combined dataset from 8 different data donation studies that have been conducted in 2023 and 2024 among different populations and focusing on different platforms. In each of these studies, immediately after explaining the data donation procedure, participants conducted a short test to see to what extent they remembered what had just been explained to them.

Approach and model

We make use of a multilevel logistic regression model where we include participant level and study level characteristics to find out how understanding of the data donation procedure is related to willingness to donate data, and actual data donation. In this model, we incorporate participant level characteristics such as demographic characteristics, privacy concerns and technological fluentness. We incorporate study level characteristics such as the platform under investigation, the recruitment strategy and the incentive level.

Main results

As preliminary results, we investigate the answers to test questions in one of the 8 studies under investigation. This was study conducted in the Centerpanel in 2023, where participants were asked to donate parts of their Google Semantic Location History data. Here, 7 different test questions were asked to the participants, and we see that consistently about 30% answers "don't know". In addition, we see that participants score the best on a question testing if they understand which data will be collected (62.3% correct) and the worst in a question testing if they understand on who Google collects data (24.8% correct). Furthermore, in this study only 5.3% of the participants had all questions correct. Most interesting, we saw that participants with more correct answers were both more willing to donate (4.54 vs 2.56 correct, OR=1.572, p<.001) , and more likely to donate (5.33 vs. 3.94 correct, OR=1.572, p<.001).

We will investigate to what extent these first results generalize over different studies, and how they potentially differ for different study characteristics and participant characteristics.

Collecting expenditure data with a household budget app in Germany. Participation behavior, non-participation bias, and receipt scanning

***Keusch, F.**, University of Mannheim; Fritz, M., University of Mannheim; Volk, J., Destatis; Häufglöckner, L., Destatis*

Diary-based household budget surveys (HBS) are often burdensome, as respondents must manually log details (e.g., product type, quantity, price) for each purchase over time. A smartphone app that allows participants to upload photos of shopping receipts could reduce this burden and increase survey participation. However, requiring an app that accesses the camera may raise privacy concerns, potentially affecting willingness to participate.

As part of the Smart Survey Implementation (SSI) Project, funded by EUROSTAT, we conducted a large-scale field test in Germany to answer three research questions: (1) How willing Germans are to download a HBS app and complete a 14-day diary?; (2) What bias does arise due to non-participation in the study?; and (3) Does highlighting the app's camera feature in the invitation influence participation rates and camera usage in the diary? In November 2024, we sampled 7,049 German residents aged 18 to 71 years from the official municipality registries and invited them via a postal letter to download the AusgabenAtlas app to their smartphone and complete the 14-day HBS. The invitation letters included a randomized experiment with three groups varying whether and how the camera function was presented.

We find that 5.7% of invited sample members activated the app, 4.1% provided at least one entry to the diary, and 1.8% completed the full 14 days. In terms of non-participation bias, we find evidence in line with the digital divide: younger people and thos with German citizenship were overrepresented among people who completed the diary compared to the invited gross sample. Mentioning the camera function in the invitation letter did not influence participation behavior, but increased the likelihood and frequency of using the smartphone camera to scan receipts.

Innovative approach to real-time data collection: Insights from an in-the-moment survey of beachgoers in Spain

Rivilla, K. S.*, RECSM-Universitat Pompeu Fabra; Ochoa, C., RECSM-Universitat Pompeu Fabra; Revilla, M., RECSM-Universitat Pompeu Fabra*

Sun protection plays a vital role in preventing skin cancer and other health issues. Thus, it is crucial to design better health interventions to understand how variables such as knowledge and attitudes towards sun exposure, alongside contextual variables such as the weather, impact sun protection behaviours.

Nonetheless, studying sun protection behaviours through surveys presents measurement challenges. When asked general questions such as "Do you use sunscreen when you go to the beach?", social desirability can bias responses, whereas when asked about specific occasions, respondents are more likely to admit non-compliance. Conventional surveys can inquire about the most recent beach visit, but psychological research highlights the limitations of recall, compromising data quality. Moreover, participants might not be aware of relevant information, such as the temperature during their visit or ultraviolet (UV) levels.

This presentation introduces an innovative approach that has the potential to enhance data quality and engagement in research: using GPS data to trigger surveys at the moment an event of interest (in this case, a beach visit) is occurring. In addition, it highlights the feasibility and implications of integrating geolocation-triggered, in-the-moment surveys with visual data collection tools and to link the survey data with meteorological data such as temperature, cloudiness and UV radiation levels.

During the summer of 2024, an in-the-moment survey was conducted using the Netquest opt-in online panel among participants already sharing their geolocation data (i.e., around 4,795 panellists) one hour after arrival at one of the 2,480 mainland beaches in Spain (N= 454). To effectively compare methodologies, a conventional retrospective survey was also fielded in the Netquest panel, with 465 separate valid participants reporting on their most recent beach visit within the last two months (covering the same period as the in-the-moment group). Uniquely, respondents were asked to upload up to five photos of the beach and if applicable, up to five photos of their sunscreen.

A comparative analysis was then conducted to evaluate participation rates, data quality, and response behaviours across three groups: on-site participants (those who responded whilst at the beach), post-visit participants (those who responded after leaving), and conventional survey respondents. Findings revealed that participation rates and post-survey evaluations were comparable between in-the-moment and retrospective surveys, but on-site participants had the highest engagement levels for event-specific tasks, such as uploading photos. The immediate temporal proximity facilitated the collection of higher-quality data and reduced discrepancies, emphasising the benefits of in-the-moment data.

Beyond the study on sun protection, this methodological innovation can be leveraged for other public health applications and in other fields, including environmental and urban studies, and consumer research focused on in-store experiences or purchasing decisions.

However, despite the advantages, challenges remain, including participant comprehension issues with task instructions and privacy concerns related to inadvertent sharing of personal information. Thus, user-friendly interfaces and vigorous privacy safeguards are essential when moving away from conventional survey methods.

Gradient boosting for hierarchical data in small area estimation

***Messer, P.**, University of Bamberg; Schmid, T., University of Bamberg*

Small Area Estimation (SAE) combines survey data with auxiliary sources such as administrative records, census data, or alternative datasets that typically offer broader coverage. By integrating these sources, SAE enhances the accuracy of (direct) survey estimates. To account for the hierarchical structure of survey data, model-based SAE methods often rely on Linear Mixed Models (LMMs). However, the distributional (e.g., normality) and structural (e.g., linearity) assumptions of LMMs may not always hold in practice, and the accuracy of model-based SAE depends on the validity of these assumptions. To address these limitations, we propose a Mixed Effect Gradient Boosting (MEGB) approach, which combines the flexibility of gradient boosting machines with the ability of mixed models to account for hierarchical data structures. MEGB extends standard gradient boosting by incorporating random effects, allowing it to capture unobserved heterogeneity across domains while retaining a nonparametric framework that models non-linearities and interactions in the data. MEGB supports the derivation of area-level means from unit-level data and uses a nonparametric bootstrap to estimate the Mean Squared Error (MSE). Its performance is assessed through a model-based simulation study, comparing MEGB to established estimators, and further demonstrated using real-world data. The results suggest that MEGB offers promising area mean estimates and may outperform existing SAE methods in various scenarios.

A support-vector mixed effects regression model for small area estimation

***Berg, E.**, Iowa State University; Pham, H., Mount Holyoke College*

Small area estimation (SAE) plays a crucial role in providing reliable estimates of population parameters at granular geographic or demographic levels where direct survey data is sparse. Traditional unit-level SAE predictors, such as the best linear unbiased predictor (EBLUP), rely on the assumption of linear associations between covariates and response variables, along with normally distributed random effects and errors. However, in practical applications, these assumptions may not hold. Therefore, the use of more flexible models is desirable. Support Vector Regression (SVR) is one way to estimate a nonlinear expectation function. We propose a new approach, the Support Vector Mixed Model (SVMM), that integrates SVR with a linear mixed effects model to obtain small area predictors. The proposed procedure can capture nonlinear associations to covariates, while maintaining a hierarchical structure and within-area correlation. We evaluate our procedure through model-based and design-based simulation studies. We compare the proposed estimator against the Unit-level EBLUP and Mixed Effect Random Forest (MERF). In the model-based simulation, we compare the three estimators under linear and nonlinear data generating processes, with 50 areas and sampling fractions ranging from 1/1000 to 50/1000. When the true model is linear, SVMM and EBLUP have comparable MSEs. Under nonlinearity, SVMM has better efficiency than the EBLUP. The proposed method is uniformly more efficient than MERF. We also propose a bootstrap MSE estimator and compare it to a linearization-based MSE estimator. The bootstrap MSE estimator exhibits a bias of around 1% for a sampling fraction of 10/100. The analytical MSE estimator has a downward bias of about 10%. In general, the simulation studies support the proposed SVMM predictor and corresponding MSE estimators.

The use of web data in small area estimation of attitudes towards climate change

***Moretti, A.**, Utrecht University; Salvatore, C., Utrecht University*

Climate change is a global problem that has a significant impact on the world’s economy and society. To effectively address climate change, policymakers require reliable estimates of relevant indicators measuring attitudes towards climate change at a sub-national level, given that these vary at a geographical level. Measuring public attitudes towards climate change is crucial in order to investigate the collective action towards sustainable practices. However, nationally representative sample surveys collecting variables around these phenomena, e.g., the European Social Survey (ESS), are not usually designed for producing accurate and precise estimates at sub-national level. In this work, we propose to use small area estimation techniques to obtain reliable estimates of attitudes towards climate change at regional level based on the ESS. The key idea of small area estimation models is to “borrow strength” from the other areas and auxiliary information based on administrative data or the Census, to improve the survey-based estimates. In recent years, the integration of digital trace data (e.g. from websites, social media, google trends) with survey data has gained importance. A novel aspect of our approach is that we include non-traditional auxiliary information, specifically web data, into our model. Our results demonstrate that incorporating web data, in some cases, yields more reliable estimates than the model without them. The results are assessed and discussed via model selection and diagnostics. Finally, we also acknowledge and address certain limitations associated with the use of web data in small area estimation.

Improving estimates on housing wealth with survey and administrative data

Porreca, E., *Bank of Italy; Benedetti, R., University of Chieti-Pescara; Neri, A., Bank of Italy; Ranalli, M. G., University of Perugia*

This study estimates housing wealth by combining survey data from the Survey on Household Income and Wealth (SHIW) with administrative data from the Italian Real Estate Market Observatory (OMI). Both sources have limitations: survey data are subject to reporting bias, while administrative data may not capture property-specific characteristics. To address these issues, we employ a finite mixture model combined with factor score prediction to account for measurement errors and exploit the complementarity of the two sources. The results reveal systematic differences between survey and administrative estimates, and show that their integration leads to more reliable and accurate estimates of housing wealth, improving our understanding of its distribution across households.

Distributional wealth accounts

***Blatnik, N.**, European Central Bank; Celislami, E., European Central Bank*

This paper presents the euro area Distributional Wealth Accounts (DWA), a dataset developed by the European System of Central Banks (ESCB) which provides experimental statistics on household wealth. In addition, the paper examines the value and challenges of using survey data for computing distributional accounts.

The DWA data complement traditional macroeconomic national accounts and household surveys by providing timely information on household wealth distribution that is consistent with the macroeconomic national accounts. The DWA respond to the growing demand for insights into wealth distribution dynamics within and across euro area countries and support the G20 Data Gaps Initiative's recommendation to improve the availability of distributional data.

The DWA reconcile two critical datasets: the Quarterly Sector Accounts (QSA) and the Household Finance and Consumption Survey (HFCS). The QSA provides comprehensive statistics on financial and non-financial transactions and positions for all euro area/EU countries, adhering to the European System of Accounts (ESA 2010) methodology, with data quarterly data starting from the Q1 1999. Complementing this, the HFCS offers detailed insights into the distribution of wealth among households, with survey waves released in 2010, 2013, 2017, and 2021.

The methodology for linking household surveys and sector accounts involves multiple steps. Initially, a wealth concept specific to the DWA is defined to maximize the overlap between the HFCS and the QSA, with necessary adjustments to individual items from both datasets. At the euro area level, this shared framework encompasses approximately 90% of household assets and liabilities as recorded in the national accounts' wealth concept, though some items like cash and pension entitlements remain excluded due to data availability constraints. Subsequently, for each HFCS release, the QSA data nearest in time are aligned, and the HFCS population scope is scaled to match that of the QSA. To address discrepancies, particularly in deposits which are often lower in HFCS compared to QSA figures, adjustments are made to account for survey outliers. Additionally, a vital step of the DWA process is estimating the wealth of rich households, who are typically underrepresented in surveys. Any remaining gaps between the adjusted HFCS and the QSA are proportionately distributed across households for each wealth concept item. Finally, quarterly DWA data are compiled by interpolating and extrapolating HFCS wave information, combined with aggregate quarterly changes in wealth components as reported in the QSA.

The results of the DWA provide valuable insights into wealth inequality within the euro area. They reveal a highly uneven distribution of household wealth, with a substantial share concentrated among the wealthiest households. This inequality varies significantly across countries, influenced by structural factors such as homeownership rates. Furthermore, wealth concentration shifts over time and across different financial instruments, shaped by diverse preferences and constraints related to credit and liquidity.

Statistical matching approaches to investigate the relationship between household income, consumption and wealth in Italy

D'Orazio, M., *ISTAT; Donatiello, G., ISTAT; D'Orazio, M., ISTAT; Loschiavo, D., Bank of Italy; Neri, A., Bank of Italy, Tullio, F., Bank of Italy*

The availability of data on the joint distribution of household (HH) income, consumption and wealth (ICW) in Italy is essential for understanding economic well-being and designing policies to fight poverty. The lack of a single survey that collects data on income, consumption and wealth led ISTAT and the Bank of Italy to join forces and study the integration of existing surveys using statistical matching (SM) methods. The surveys in consideration are the "EU Statistics on Income and Living Conditions" (EU-SILC) and the "Household Budget Survey" (HBS), both conducted annually by ISTAT, and the "Survey on Household Income and Wealth" (SHIW), carried out every two years by the Bank of Italy. The main challenges related to this SM exercise are: (i) the number of surveys to be integrated, three instead of the two usually found in SM applications; and (ii) the need to carry out an integration at the micro level.

In order to meet these challenges, it was decided to test two parallel and independent approaches to SM. ISTAT's approach takes stock of past experience in matching SILC and HBS, but constraints (i) and (ii) led it to abandon the "sophisticated" approaches in favour of simpler nearest-neighbour donor (NND) hotdeck SM methods, modified accordingly to use auxiliary information when integrating SILC and HBS surveys. The Bank of Italy's approach takes stock of past experience and consists of a single SM step that uses NND to impute in SHIW the expenditure observed in HBS, which is used as the donor. This simpler approach can be considered because it is not designed to produce microdata for Eurostat, but rather to provide a second ICW distribution to be used for validation purposes, i.e. to assess the coherence with the results obtained with ISTAT's SM approach.

The two parallel SM exercises provide, in addition to the well-known SM diagnostics, a tool to assess "indirectly" the accuracy of the SM results. However, given the importance of this issue, in the ISTAT SM exercise we have added an additional validation step based on a rough assessment of the uncertainty attributable to the matching framework.

These two SM applications were developed with survey data referring to the year 2016 (see Donatiello et al, 2025) and then replicated with surveys referring to the year 2020. The SM exercises with 2020 data are not just a simple replication of the exercises carried out in 2016, because in order to produce national experimental statistics comparable to the data produced by Eurostat, we applied common concepts and consistent classifications. As a result, we had to consider slightly different target variables, such as income and consumption excluding imputed rents and other non-monetary components, and we also had to face some difficulties related to a change in some marginal/joint distributions of these variables due to the COVID-19 pandemic period.

Contributed Sessions

Population size estimation and consensus ethnicity classification for domains using multiple sources

***Smith, P.A.**, University of Southampton; van der Heijden, P.G.M., Utrecht University & University of Southampton; Cruyff, M., Utrecht University; Pantalone, F., University of Southampton; Diener, H., Statistics New Zealand; Dunstan, K., Statistics New Zealand*

The availability of multiple administrative sources in New Zealand, each collecting information on ethnicity, supports multiple system estimation for population size estimation, and also allows a consensus view of ethnicity based on the responses by use of a latent class model. These two approaches were combined into a latent class multiple system estimation (LCMSE) model by van der Heijden et al. (2021), who used it as the basis for estimating the size of the Māori population in New Zealand. In Smith et al. (2025) this was extended to four ethnic groupings covering combinations of Māori and Pacific, and used for multiple years which showed that the estimates were very consistent over time.

In this paper we extend the models with an additional classification variable, aiming to make population size estimates for ethnicity groupings within more detailed domains. There are many potential domain classifiers of policy interest, and we experiment initially with sex. Using an additional classifier opens up a wide range of models (Clogg & Goodman 1984), where the latent classes can be different in different domains or common across domains, and a range of intermediates. We explore in particular three central cases, a complete homogeneity model, a partial heterogeneity model where each source is an equally reliable indicator of each latent class, and a complete heterogeneity model.

Using sex we find that the complete homogeneity model is satisfactory (there are no differences in the model by sex), and that the results are stable. We explore the quality and interpretation of the latent classes in the best fitting models, and the impact of the model choices on the population size estimates.

References

Clogg, C.C. & Goodman, L.A. (1984) Latent structure analysis of a set of multidimensional contingency tables. Journal of the American Statistical Association 79 762–771.

Smith, P.A., van der Heijden, P.G.M., Cruyff, M., Pantalone, F., Diener, H. & Dunstan, K. (2025, in press) Population size estimation using covariates having missing values and measurement error: estimating ethnic group sizes in New Zealand. Australia and New Zealand Journal of Statistics.

Van der Heijden, P.G.M., Cruyff, M., Smith, P.A., Bycroft, C., Graham, P. & Matheson-Dunning, N. (2022) Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Māori population in New Zealand. Journal of the Royal Statistical Society, Series A 185 156-177. doi: 10.1111/rssa.12731

Exploring M2M transmission for short-term business official surveys in the industry 5.0 era

Papa, P., ISTAT; Bosso, P., ISTAT; Di Paolo, G.G., ISTAT; ***Distefano, D.****, ISTAT*

The increasing digitalization of industrial processes, driven by the Industry 5.0 paradigm, has enhanced system interoperability, promoted standardization, and enabled more efficient and scalable operations across supply chains. This transformation has resulted in vast data availability, positioning data management as a strategic asset for businesses. Within this context, the study explores the feasibility and implications of applying a generalized Machine-to-Machine (M2M) data transmission approach, built on advanced ERP platforms integrated with Industry 5.0 technologies, for official short-term business statistics in Italy.

The paper outlines the main phases of an experimental trial developed in the framework of a project under definition at ISTAT. The study addresses key aspects including the statistical burden of short-term surveys, the digital maturity of Italian enterprises (especially SMEs), and the design of an experimental trial. ISTAT's 2023 business surveys involved over 450,000 companies, with short-term surveys alone engaging around 127,000 businesses. Despite a lower number, large enterprises contribute disproportionately to employment and added value. Business surveys rely heavily on CAWI methods, which, while cost-effective, still require significant internal resources. In 2024, 11 short-term surveys involving over 72,000 firms resulted in nearly one million data submissions.

The trial adopts AI usage as a proxy for digital maturity. ICT 2024 data, produced by ISTAT, reveals growing AI adoption, particularly in knowledge extraction, generative AI, and speech-to-text technologies, with substantial differences between SMEs and large enterprises. While 97.8% of large firms have achieved basic digitalization, only 70.2% of SMEs have done so, and fewer reach advanced levels. The metallurgy sector, exhibiting high digital maturity, was selected for the experimental phase.

The experimental design includes selecting a reference variable (industrial production volume), identifying suitable ERP/MES platforms, engaging ERP vendors covering 60% of the market, and working with a sample of firms to assess the compatibility of platform data with official needs. Key trial components include evaluating acceptability among businesses, data alignment with EU Regulation 2019/2152, and comparing data collected via M2M with traditional survey methods.

Initial findings highlight clear opportunities: reducing response burden, enhancing timeliness, and lowering medium-term costs. However, challenges remain, particularly around the heterogeneity of platforms, resistance to automated data sharing, and the need for harmonization between business data and official classifications. Moreover, implementing M2M systems requires initial investments and maintaining parallel traditional and administrative data channels during transition phases.

The study contributes to a broader movement among National Statistical Institutes (NSIs) toward a multi-source data ecosystem, increasingly leveraging automation, AI, and digital innovation. While short-term benefits are limited, the long-term potential for efficiency and improved data quality is substantial. The findings emphasize the importance of sustained investment and stakeholder engagement to realize these gains and underline that embracing innovation inevitably involves managing transitional complexity and uncertainty.

Exploring patterns and determinants of agricultural carbon footprint in Northern Italy: small area models integrating survey, census and remote sensing data

Borgoni, R., *University of Milan-Bicocca; Carillo, F., CREA; Maranzano, P., University of Milan-Bicocca; Pajno, R., University of Milan-Bicocca*

In this paper, we examine the spatial dynamics of the carbon footprint (CF) associated with farms in the Po Valley (northern Italy), one of the most polluted regions in the world. Intensive livestock farming plays a crucial role in driving high levels of ammonia and particulate matter in the atmosphere. Specifically, we analyze the relationship between agricultural CF, the economic and techno-productive characteristics of farms, and landscape features using small-area spatial models based on geocoded data.

The data used in this paper come from several institutional sources, namely the Farms Accountancy Data Network (FADN) survey, the 2020 Italian National Census of Agriculture, and the National Veterinary Registry Office Database. We also consider some cartographic and orographic information related to the area of interest for this study and incorporate it into our analysis. At the Italian level, the FADN has been conducted by the Italian Council for Agricultural Research and Economics (CREA). The data include information from a random sample of farms selected from the population of Italian farms with a Standard Output of no less than 8,000 euro per year. The Italian FADN sample considered in our study includes 2,806 farms spread across the region under study.

To generate accurate estimates of CF at a highly granular spatial scale — specifically, the 254 Agrarian Sub-regions covering the Po Valley — we incorporated total NH_3 emissions into our model. Data on NH_3 emissions were obtained from the Copernicus Atmosphere Monitoring Service (CAMS). CAMS is one of the most recent global databases of emissions and concentrations from anthropogenic sources. The CAMS datasets provide annual emission inventories for many atmospheric compounds from 2000 to 2020 and are recorded on a point grid with a resolution of 0.5°×0.5°. These emissions were spatially realigned with the other georeferenced data using spatial interpolation and included among the potentially relevant predictors.

The inclusion of satellite information led to an improvement in the fit of our models. Nevertheless, the main determinant of CF was revealed to be the number of livestock head, being the most significant variable in our model.

Empirical results have clearly shown that the small area model-based approach significantly outperforms direct estimators that do not use auxiliary information. The CF was found to be very heterogeneous among the agricultural sub-regions, while remaining coherent with the economic and environmental configuration of Northern Italy. In particular, the CF is very high in the plain areas, where intensive livestock farming is widespread.

Mismatched measures: Comparing asset-based wealth index and income/consumption poverty in Mozambique

***Allorant, A.**, University of Southampton; Perfetti Villa, L, University of Southampton; Luna Hernandez, A, University of Southampton; Tzavidis, N, University of Southampton*

Micro-targeting poverty alleviation programs, which identify and reach the most economically vulnerable populations, are widely regarded as an effective strategy for reducing poverty. However, many low- to middle-income countries (LMICs) lack the granular income or consumption data needed to accurately locate and support the poorest households. Recent large-scale mapping initiatives- including satellite-based approaches such as Meta's Data for Good, rely on the Wealth Index (WI) as a "ground-truth" measure of socio-economic status (SES) for training machine learning models. The WI, derived from readily available household asset data using principal components analysis (PCA), has long been used as a proxy for SES in LMICs. Yet, recent evidence suggests notable discrepancies between WI-based classifications and direct measures of economic well-being (income or consumption), raising concerns about the validity of using WI as to guide policy and interventions.

Using three nationally representative household surveys from Mozambique (2008, 2014, 2019), each containing both asset and income/consumption data, we constructed the WI following established methods. We performed PCA on a set of standardized household assets to derive factor loadings, which were then used to generate a continuous wealth score for each household. Quintiles and centiles of the WI distribution were computed using survey weights. Households were classified as poor or non-poor by comparing their consumption with the national poverty line.

To generate policy-relevant district-level estimates across 134/161 districts (because boundaries changed during this period), we employed small area estimation (SAE) techniques. Specifically, we applied Fay-Herriot models to estimate both mean income and head-count ratios at the district level. These models transform direct survey estimates and incorporate district-specific random effects to improve precision.

At the household level, rank correlations between WI-based and consumption-based classifications were modest (46–56%), revealing substantial discrepancies in who is categorized as poor. Notably, an urban bias emerged when using the WI, wherein rural poverty rates appeared higher and urban poverty rates appeared lower compared to those derived from income or consumption data. For instance, in 2019, the WI suggested that 81% of rural households were poor versus 71% by expenditures; meanwhile, 22% of urban households were poor by the WI versus 45% using expenditures. District-level comparisons via SAE showed somewhat stronger correlations between mean WI and mean consumption, with the WI performing better at identifying the richest districts. Nonetheless, there were systematic differences in poverty estimates for certain districts, indicating that WI-driven measures may still misrepresent the spatial distribution of poverty.

Our findings underscore that, despite its widespread use, the asset-based Wealth Index can yield significantly different poverty classifications than direct income or consumption measures. Given that major big-data initiatives, such as Meta's global poverty mapping, rely on WI-based training data, these mismatches raise concerns about the potential misallocation of resources and under-coverage of vulnerable populations.

Are we reaching the poor? Evaluating Austria’s social assistance schemes via small area estimation

Broka, A., *University of Bologna; De Nicolò, S., University of Bologna; Angel, S., Austrian Institute of Economic Research*

In 2011, Austria implemented a national-level Minimum Income Scheme, which remained in effect until 2016 and sought to harmonize key elements of social assistance across the nine federal provinces. Following its expiration, the regulation of social assistance returned to the federal states, and national minimum standards were no longer ensured. A second attempt at harmonizing regional social assistance schemes was made in 2019 through the introduction of the Basic Social Assistance Act, which, however, as of January 2025, has been implemented only in seven out of nine regions. This highly fragmentated context prompt us to scrutinize the effectiveness of Austria’s basic social assistance schemes from 2011 to 2022 at the local level, considering regional and household-type heterogeneities that aggregated analyses may overlook. Effectiveness is measured through three key indicators: the Coverage Rate (the share of poor individuals receiving benefits), the Eligibility Rate (the share of poor households that meet the eligibility criteria), and the Take-up Rate (the share of eligible households that receive benefits). The indicators are estimated at the domain level using small area estimation models and exploiting the relationship among them. Specifically, we apply area-level Hierarchical Bayes Beta and Extended Beta regression models, estimated via Markov Chain Monte Carlo (MCMC) methods. Our approach combines data from the EU-SILC, administrative records on beneficiaries, domain-level covariates, and the HFCS. We rely on the assumption that only eligible households can take up benefits. Our findings reveal significant heterogeneity across years, regions, and household types, highlighting how policy fragmentation and instability have led to uneven effectiveness. While the Austrian social assistance system tends to be most effective for single-parent families, it systematically underperforms in addressing the needs of individuals living alone and childless couples.

Area-level small area estimation with random forests

***Harmening, S.**, Otto-Friedrich-Universität Bamberg; Lee, Y., Otto-Friedrich-Universität Bamberg; Runge, M., Freie Universität Berlin; Schmid, T., Otto-Friedrich-Universität Bamberg*

Interactions among explanatory variables and nonlinear relationships between them and the dependent variable are present in many data applications. An approach that combines a small area estimation model with tree-based methods to provide a solution when only area-level data are available is presented, namely the area-level mixed-effects random forest. In particular, the linear regression synthetic part of the Fay-Herriot model is replaced by a random forest to link survey data with related administrative information or data from other sources. By using a random forest, possible interactions and nonlinear relationships are accounted for, and automatic variable selection and robustness to outliers are indirectly provided as a property of the random forest. To obtain point estimates for an indicator of interest, the familiar structure of the Fay-Herriot estimator is retained. The estimation is done by implementing an expectation maximization algorithm. To determine the uncertainty of the point estimator, a nonparametric bootstrap method for estimating the mean squared error is presented. The use of data transformations like the log transformation to achieve a more symmetric distribution and to move extreme observations toward the center of the distribution is investigated in the context of machine learning methods. In particular, a log transformation is applied to the direct estimates and due to the nonlinearity of the logarithm, the final point mixed-effects random forest and mean squared error estimates on the original scale are back-transformed by taking into account a bias-correction. To evaluate the accuracy and precision of the proposed estimator and its uncertainty measure, model-based simulations are carried out. The simulation results highlight that the proposed point estimator leads to unbiased estimates and that the presented approach can improve the efficiency with a large number of domains in the presence of interactions and additional noise variables compared to the standard and semiparametric Fay-Herriot estimators. The proposed bootstrap scheme for estimating the mean squared error of the point estimator is also proven through simulations and leads to reliable uncertainty measures. The presented methodology is illustrated by using household survey and remote sensing data from Mozambique to estimate average per capita consumption at a km grid-level. Preliminary analyses of the data indicate interactions between the explanatory variables and nonlinear relationships between the dependent variable and the explanatory variables. The application to real world data shows that the approach is able to produce reasonable results and leads to an improvement in efficiency compared to the direct estimator and also to the well-known log-transformed Fay-Herriot estimator.

Improving prediction accuracy in small area models via clusterwise regression

***Maranzano, P.**, University of Milano-Bicocca; Mattera, R., Luigi Vanvitelli University of Campania; Sugasawa, S., Keio University*

In the context of small-area estimation (SAE) models on data from sample surveys, a typical approach is to use auxiliary information to enhance the precision of estimates obtained from direct total or mean estimators. Direct estimates and auxiliary information are integrated through regression models that can include both fixed and random effects, thus allowing for the development of mixed-effects models with potentially spatially and temporally structured components (Morales et al., 2021). These models typically assume that the regression coefficients remain constant over time or space. However, as observed by Wang et al. (2023), this strategy may prove to be inadequate in light of the possibility that the relationships between variables may vary in space, thereby giving rise to the presence of spatial heterogeneity. The objective of this paper is to present an innovative approach to simultaneously address the issues of heterogeneity and spatial dependence in small-area estimation (SAE) models. This strategy aims to enhance the predictive capabilities of SAE models and provide more accurate estimates of the sample variables of interest. The proposed methodology integrates three previously suggested methodologies: (1) Sugasawa and Murakami (2021) proposal of spatially-clustered regression models, in which regression coefficients can vary according to a spatial cluster structure determined endogenously through penalized likelihood; (2) Wang et al. (2023) proposal, in which in a context of spatial penalized least squares, location-specific weights are employed to estimate local regression coefficients and clustering membership; (3) Cerqueti et al. (2024) proposal which extended the spatially-clustered linear regression model to encompass the leading spatial econometric models (e.g., SAR and Durbin model). In particular, the proposal entails the estimation of linear mixed effects models belonging to the Fay-Herriot family with clusterwise spatially-varying coefficients, wherein areas are merged through a spatially-penalized likelihood. The proposed methodology is applied to data on Italian farms provided by the Farm Accountancy Data Network (FADN) survey of the European Union (Baldoni et al., 2017). The dataset consists of a sample of thousands of farms across the country, the economic, production, technological, energy, and environmental impact information of which is collected annually. In particular, the application involves estimating the carbon footprint of farms in the Po Valley in recent years supported by auxiliary information from the 2020 national agricultural census.

References

Baldoni, E., Coderoni, S., & Esposti, R. (2017). The productivity and environment nexus with farm-level data. The Case of Carbon Footprint in Lombardy FADN farms. Bio-based and Applied Economics, 6(2), 119-137.

Cerqueti, R., Maranzano, P., & Mattera, R. (2024). Spatially-clustered spatial autoregressive models with application to agricultural market concentration in Europe. arXiv preprint: 2407.15874. https://doi.org/https://doi.org/10.48550/arXiv.2407.15874

Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). A course on small area estimation and mixed models. Methods, theory and applications in R.

Sugasawa, S., & Murakami, D. (2021). Spatially clustered regression. Spatial Statistics, 44, 100525. https://doi.org/https://doi.org/10.1016/j.spasta.2021.100525

Wang, X., Zhu, Z., & Zhang, H. H. (2023). Spatial heterogeneity automatic detection and estimation. Computational Statistics & Data Analysis, 180, 107667. https://doi.org/https://doi.org/10.1016/j.csda.2022.107667

On some new diagnostic measures for area-level models in small area estimation

***Marcis, L.**, University of Cassino and Southern Lazio; Pagliarella, M.C., University of Cassino and Southern Lazio; Salvatore, R., University of Cassino and Southern Lazio*

In small area estimation models, standard diagnostic measures are given by common tools used in linear regression and adapted to the linear mixed effects models. These measures are important for the data analysis, when the main issue is to understand the impact of particular observations on the estimation of a linear (or a linear mixed) model. The measures proposed for a linear mixed model are different from those used for the standard linear regression model due to the presence of a multiple set of model covariance parameters. In general, in the area-level small area estimation (Rao and Molina, 2015), the model check consists of the residual analyses, the influence measures, and, more importantly, the evaluation of the efficiency of the empirical best linear unbiased predictor (Eblup). Traditionally, the last issue is evaluated by averaging the bias and the prediction mean square error (mse) of the different models under investigation, given by studying the behavior of simulated and/or actual data applications.

Even if the small area models can be reconducted to a linear mixed model, the focus needs to be on their differences concerning the latter. Influence measures designed for general mixed models are not always well exploitable when applying the area-level models (Morales et al., 2021), i.e., the Fay-Herriot model and its extensions, because sometimes those tools do not take into account the peculiarities of the small area models. For example, in the residual analysis, the circumstance given by the presence of the sampling variances of direct estimators in the sampling model sometimes leaves practitioners in doubt about the actual extent of the prediction of the corresponding model error. An analysis of the correlation between residuals and model error was introduced by Marcis et al. (2023), and some other measures were proposed to analyze the efficiency of the linear predictor. Moreover, diagnostic measures like Cook's distance may be useful to indicate the impact of influential observations (i.e., the small areas) on the model estimation. Some statistical properties related to the given values of the sampling variances in the area-level models lead to special consideration of this index.

The present work addresses a comprehensive discussion on the evaluation of the issue of the efficiency of the linear predictor and introduces some mathematical properties of the Cook's distance when adapted to the Fay-Herriot model and its extensions. In particular, we derive the corresponding distribution of the Cook's distance, together with the assessment of the model in case of deletion diagnostics.

References

Marcis, L., Morales, D., Pagliarella, M.C., Salvatore, R. (2023) Three-fold Fay-Herriot model for small area estimation and its diagnostics. Statistical Methods & Applications, 32: 1563-1609.

Morales, D., Esteban, M.D., Perez, A., Hobza, T. (2021). A course on small area estimation and mixed models. Springer.

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

A comparative study of bootstrap confidence intervals in small area estimation

Ferrante, M. R., University of Bologna; ***Mori, L.****, University of Bologna; Tizianel, E., University of Bologna*

Small Area Estimation (SAE) encompasses statistical techniques designed to provide reliable estimates for small geographical regions or specific demographic sub-populations. Among SAE models, unit-level and area-level approaches are widely used. The unit-level approach typically relies on the Nested Error Regression (NER) model, which incorporates individual-level covariates. In contrast, the area-level approach, represented by the Fay-Herriot (FH) model, utilizes aggregated survey estimates and area-level auxiliary data. A critical aspect of SAE is the estimation of the Mean Squared Error (MSE) to assess the variability of the estimates. The parametric bootstrap is a widely used technique for MSE estimation (Liu et al., 2023). However, in the SAE context, less attention has been given to the construction of Confidence Intervals (CIs).

The traditional approach of constructing CIs based on the normality assumption, using ±1.96√MSE, may yield unsatisfactory results when MSE is estimated via bootstrap (see, among others, Hall (1988) and Jung et al. (2019)). This is primarily due to deviations from normality, biased estimations, and the effects of small sample sizes. Several alternative methods have been proposed to address these limitations, yet no comprehensive comparison has been made in the SAE framework.

This work focuses on evaluating and comparing different bootstrap-based CI methods in the context of SAE. The study considers both unit- and area-level models, with a specific focus on estimating the area mean, Headcount Ratio (HCR), and Gini index under different distributional hypotheses. The analysis is conducted using R software and employs Monte Carlo simulations to assess the empirical performance of the different CI methods. The CI methods evaluated include the Standard Normal CI, Percentile Bootstrap CI, Reverse Percentile Bootstrap CI, and two versions of the Bias-Corrected and Accelerated (BCa) Bootstrap CI, one using jackknife correction at the area level and the other at the unit level.

Preliminary results show, as expected, that the traditional normal CI performs well for linear parameters such as the mean but struggles with skewed distributions and small sample sizes. Furthermore, the percentile method, while simple, tends to underestimate coverage due to its lack of bias correction. The reverse percentile method provides better coverage but generates wider intervals. The BCa methods offer improved performance by accounting for bias and skewness, yet produce excessively narrow intervals, leading to under-coverage.

References

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. The Annals of Statistics, 16,927–953.

Jung, K., Lee, J., Gupta, V., and Cho, G. (2019). Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation. Frontiers in psychology, 10, 2215.

Liu, Y., Liu, X., Pan, Y., Jiang, J., and Xiao, P. (2023). An empirical comparison of various MSPE estimators and associated prediction intervals for small area means. Journal of Statistical Computation and Simulation, 93, 1532–1558.

Estimating the variance of basic summary statistics in the absence of complete design information. A parametric simulation study

***Stacchini, A.**, University of Bologna*

External analysts of data collected through complex designs often do not have access to microdata about all design variables, due to security, privacy, industrial or commercial issues. Even if a design is negligible when known fully, it can become informative when only partial information about it is available. This makes it a puzzling problem to estimate the variability of even basic summary statistics, in the absence of complete design information, especially because it is still an under-researched issue.

When the sample values of all design variables are known, the sampling variance is computed on non-parametric bootstrap replicas, obtained by mimicking the original design. This avoids overestimation, when stratification induces efficiency gains, and underestimation, when clustering increases within correlation, occurring by resampling observations as if they were independent. The latter estimation strategy (naïve) is commonly applied, within target domains, when no design variable is available. An alternative consists in resampling observations with inclusion probabilities equal to the inverse of the design weights (WB). So far, no comparison of the empirical performance of these two methods has been carried out, particularly in case target domains are not (sub-)elements of the design. When microdata about some design variables are accessible, it is unclear whether mimicking the original design as faithfully as possible with the information available (MM) is worth its cost. This study is a first attempt to start to clarify this issue, through a parametric simulation, that also compares the accuracy of the weighted and naïve bootstrap-based estimates.

As parameters of the data generating process, we plug-in statistics from the 2022 Survey on International Tourism, conducted to estimate the total inbound tourist spending. Eight design variables are used, half of which are accessible. As target domains, we consider the Italian local administrative units, because their highly variable sample sizes allow us to assess the sensitivity of the variance estimates to the observations number. We estimate the variance of the total, the mean and the 0.1, 0.2, 0.5, 0.6, 0.8, 0.9 quantiles and pseudo-quantiles of tourist expenditure. We check if the performance of the three estimation methodologies vary across degrees of centrality, particularly for extreme quantiles in a skewed distribution, as we specify a skew normal model.

Results suggest that MM is not helpful, while WB slightly improves the estimates’ accuracy compared to the naïve and does not imply a higher computational and time cost. WB yields larger variances than the naïve, incorporating an approximation of the design effects. Conversely, MM tends to widely underestimate variances, excessively replicating stratification-induced efficiency gains. The least uncertain estimates are obtained through the naïve. The bias and MSE are greater for quantiles and pseudo-quantiles in the left tail, smaller for the mean compared to the median and for pseudo-quantiles compared to quantiles, confirming that the more non-linear the estimator, the more complex to estimate its variance. As expected, the larger the sample size, the smaller the MSE becomes. The uncertainty is particularly high when the number of observations is fewer than 10.

On the use of bootstrap to enhance variance estimation for more than one-frame surveys

Cocchi, D., University of Bologna; ***Ievoli, R.****, University of Ferrara*

Dealing with more than one-frame case in complex survey sampling still represents a relevant topic. Hartley introduced the first estimation method in the sixties and several developments have been proposed in the last two decades, including the multiplicity estimator of Singh and Mecatti (2014). Moreover, variance estimation in dual or multiple-frame settings still represents one of the main challenges for scholars and practitioners. In this sense, two main issues should be taken into account: a) the overlapping of the frames, and b) the possibly complicated articulation of the sampling design within the frames. Linearization, which is not always straightforward, has been an initial solution was an initial solution but has since been replaced by resampling methods. Indeed, some bootstrap techniques have been proposed to estimate the variance in this case. Advantages are a) the flexibility of the bootstrap versus the jackknife, b) the bypassing of the calculation of higher-order inclusion probabilities, and c) the suitability for smooth and non-smooth statistics.

The first proposal was developed by Lohr (2007) for dual frame surveys, adapting the rescaling bootstrap. More recently, some authors have resumed bootstrap algorithms in the case of dual/multiple frames. Among them, Aidara (2024) deals with a frequentist bootstrap approach in estimating the variance of the multiplicity estimator under multiple frames. In addition, Kumar et al. (2024) introduced a post-stratified rescaling bootstrap suited for the case of dual frames, comparing the performance of the proposal to a conventional naive bootstrap. Finally, Cocchi et al. (2022) provided the first Bayesian bootstrap algorithm suited for estimating the variance of the multiplicity estimator in multiple frames.

Unfortunately, these approaches remain quite unrelated, since they are discussed in specific settings and evaluated via non-comparable simulation setups. Then, a relevant gap can be identified in the systematization and comparison of these proposals. The present work contributes to the literature concerning variance estimation via resampling in more than one-frame. We aim to compare different proposals to understand the advantages and disadvantages of such approaches via a small-scale simulation study which considers different simulation setups. Performance indicators such as the relative bias and the coefficient of variation are used for comparison purposes. The applicability of these methods is also explored through a case study.

References

Aidara, C. A. T. (2024). Enhancing Multiple Frame Surveys: Improved Calibration and Efficient Bootstrap Techniques. European Journal of Statistics, 4.

Cocchi, D., Marchi, L., & Ievoli, R. (2022). Bayesian bootstrap in multiple frames. Stats, 5(2), 561-571.

Kumar, R., Rai, A., Ahmad, T., Biswas, A., Misra Sahoo, P., & Moury, P. K. (2024). Rescaling bootstrap variance estimation technique under dual frame surveys with unknown domain sizes. Communications in Statistics-Simulation and Computation, 1-14.

Lohr, S. (2007). Recent developments in multiple frame surveys.

Mecatti, F., & Singh, A. C. (2014). Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. Journal de la Société Française de Statistique, 155(4), 51-69.

Evaluation of anonymised georeferenced data

***Gril, L.**, Freie Universität Berlin; Rendtel, U., Freie Universität Berlin*

Georeferenced data is often anonymised for data protection reasons. This is done either by aggregation to larger spatial units (such as higher-order administrative units or grids with larger edge lengths) or by the use of stochastic methods that specifically overlay the original coordinates. The aim of such an analysis is to visualise the spatial distribution of a feature of interest, for example in the form of maps. Conventional analysis methods often do not take the anonymisation process into account and treat anonymised coordinates as actual coordinates. However, a statistical measurement error model enables much more efficient analyses by explicitly considering the influence of anonymisation. The presentation aims to show results from an ongoing research cluster in co-operation with the Federal Statistical Office. In addition to the presentation of the methods developed, empirical results from a demographic and health survey as well as the regional distribution of income taxpayers in Berlin will also be presented.

Survey sampling and differential privacy

***Bonnéry, D.**, Insee; Jamme, J., Insee*

Epsilon-differential privacy is a mathematical criterion controlling how much the output distribution of an estimator, after adding random perturbation, can change when a single individual's data is modified, while epsilon-delta differential privacy allows for a small failure probability; both are widely adopted as industry standards for data protection. Results obtained for the protection offered by surveying data in terms of epsilon and epsilon-delta differential privacy will be presented. More precisely, we derive the relationship between the level of protection offered by a particular survey, its single and double inclusion probabilities and the dispersion of the variable of interest, when the published statistic is the Horvitz Thompson estimator of a total. We provide a tool to compute the standard deviation of the additional privacy mechanism to apply, if necessary, to reach a desired level of protection.

Using perturbative methods for magnitude tables in statistical disclosure control

Adriansson, N., Statistics Sweden; ***Sabolová, R.****, Statistics Sweden; Tepe, Ö., Statistics Sweden*

Data in tables published for the Swedish R&D survey in the business enterprise sector (BERD) were previously protected by cell suppression to prevent disclosure of sensitive information. In order to avoid cell suppression, key respondents were asked to sign waivers allowing the publication of their data. However, consent was rarely given to disseminate cells where an enterprise's data potentially could be disclosed. The steps undertaken to ensure confidentiality were extensive, involving multiple staff and requiring a significant amount of time during production. For users, this resulted in the withholding of statistical information, particularly in tables presenting industry-specific data and its combination with other domains of interest. Consequently, the usefulness and relevance of the published statistics on a granular level were reduced despite the efforts made.

To address these challenges, the BERD survey became the first survey at Statistics Sweden to use a perturbative method for disclosure limitation in magnitude tables. The EZS-method, introduced by Evans, Zayatz, and Slanta (1998: Journal of Official Statistics, 14, 537-551), adds noise to microdata to ensure table additivity and preserve links among tables. Each enterprise is assigned random values for the direction and noise factor, which are kept confidential. Perturbed values are calculated as perturbed value = original value * (1 + direction * noise factor/100), where both the direction and noise factor are applied to all values reported by the object. The distribution of directions of perturbation is chosen so that it is symmetric around 0 and thus does not introduce any consistent bias.

Cells with one object or dominant contributions receive more noise to protect individual data, while noise in cells with many smaller contributions cancels out. The balancing procedure proposed by Massell and Funk (2007: Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III), Montreal, Canada) is used to reduce overall noise, applied only to cells without risk of disclosure.

The method was tested on 2021 data from the BERD survey and successfully used for 2023 data published in autumn 2024. It allows for disseminating tables with approximate values without suppressing any cells and is simple to implement without specialized software. The method works well for high-dimensional tables or tables with hierarchies and is more effective than cell suppression.

Small area estimation under several spatial specifications

***Billé, A.G.**, University of Bologna; Ferrante M.R. University of Bologna; Salvati N. University of Pisa*

The recent literature on Small Area Estimation has begun to introduce several types of spatial effects into the standard Fay-Herriot approach of estimating parameters for small areas. To mention a few, Singh et al. [2005], Pratesi et al. [2009], You and Zhou [2011], Kubacki and Jedrzejczak [2016], Chandra and Salvati [2018], and Chung and Datta [2020] focused the attention on a spatial autoregressive process among the (areal-specific) random effects. The aim of this paper is to overcome the above-mentioned literature by extending the FH model specification through the inclusion of different sources of spatial dependency. The first specification takes into account the usual correlated random effects and the spatial effects directly in the direct estimates of interest, whereas the second specification considers again spatial random effects and lagged exogenous covariates. We properly analytically approximate the MSE by extending the work of Molina et al. [2009]. Monte Carlo simulations of the proposed ML-based estimators are included. Finally, an empirical application is included.

Measuring educational poverty in Italy: A MIRT-SAE approach

Ranalli, M.G., University of Perugia; Bertarelli, G., University of Venice; ***Del Sarto, S.****, University of Perugia; Guandalini, A., ISTAT; Pratesi, M., University of Pisa*

Educational Poverty (EP) is an increasingly relevant concept, particularly in the context of policy enforcement and monitoring at the local level. Its significance for well-being and sustainable development is widely recognized in the broader discussion surrounding the 2030 Agenda for Sustainable Development Goals, which emphasizes the need to ensure inclusive, equitable, high-quality education and lifelong learning opportunities for all. Measuring EP is challenging due to its latent and potentially multidimensional nature, and there is still no consensus among researchers on a definitive set of indicators. In this study, we focus on a set of binary indicators derived from the ISTAT Activities of Daily Living survey, conducted on individuals aged 15–29. We propose a multidimensional Item Response Theory (MIRT) model to extract the latent dimensions of EP, combined with a small area estimation (SAE) approach to obtain subregional estimates. Specifically, we employ an exploratory MIRT model as the measurement component to investigate the latent structure of EP, identifying subsets of indicators that contribute to measuring its distinct dimensions. The structural component consists of a unit-level SAE model, where the latent traits are modeled using covariates and area-specific random effects. This estimation process allows us to produce local-level model-based predictions of EP’s latent factors, as identified through the measurement model. While a one-step estimation approach is possible, it is not desirable that covariates influence the definition of the latent variables, which should conceptually remain distinct from them. Therefore, a more reliable two-step approach is adopted, where the structural model is estimated by maintaining the underlying latent structure fixed and equal to that determined in the exploratory measurement model. Finally, a measure of precision of the estimates for small areas can be obtained using block bootstrap.

Municipal-level estimation of tourism perception: A machine learning-based approach to small area estimation

Niederhametner, N., *Statistics Austria; Daul, R., Statistics Austria*

Understanding the perception and acceptance of tourism among local residents is crucial for sustainable tourism management. In this talk, we want to present Austria's new "Tourism Acceptance" survey and its methodological framework, which includes an innovative model for estimating the level of tourism acceptance in Austria's municipalities. Drawing on a sample of approximately 12,000 respondents per year, the Tourism Acceptance survey asks participants about their opinions on tourism's impact on and significance for the economy, labor market and leisure activities in their place of residence, as well as whether they think the number of tourists is too low, excessive or acceptable. Tourism intensity and its impact on the local population vary in different regions of Austria. Consequently, there is a need to monitor tourism acceptance and perception on a smaller regional level than the federal province. However, interviewing a representative sample in each region or municipality in Austria is neither affordable nor feasible. Therefore, we have developed a Small Area Estimation Model, which builds on possible tourism indicators that can influence tourism acceptance in the municipalities.

To estimate the perception of the entire population, we propose to employ machine learning models, leveraging auxiliary data to impute the response variable for non-surveyed citizens at the unit level. We do this by linking each respondent's answer to auxiliary administrative data, including demographic information (age, sex, place of residence, income), employment sector (NACE classification), and municipal-level data (tourism-related profits, number of overnight stays per capita). These auxiliary variables are available for the entire country's population. An XGBoost model is then trained on this data to predict a person's answer for the question "How do you perceive the number of tourists in your place of residence?". We predict the answers of residents not included in the survey. This allows us to aggregate both the predicted estimates and, if applicable, the actual survey responses of all residents within the same municipality, providing a single acceptance estimate per municipality. Additionally, we use a bootstrap approach to estimate the errors, and construct confidence intervals. This is done by calculating the standard errors from 1000 model predictions per person from 1000 bootstrap weights. The confidence intervals per municipality are then constructed through a bootstrap-based simulation: for each individual, a simulated probability is drawn from a truncated normal distribution with the estimated probability as the mean and the bootstrap-derived standard deviation. This process is repeated 1000 times, generating a distribution of regional estimates, and the 2.5th and 97.5th percentiles of these estimates define the 95% confidence interval.

This approach parallels methods used in Small Area Estimation, extending the analysis beyond the sample to provide comprehensive estimates.

The survey results offer valuable insights into tourism acceptance, highlighting spatial variations and underlying socio-economic factors influencing residents' perspectives on tourism. This model-based approach not only enhances our understanding of local tourism acceptance but also demonstrates the potential of integrating survey data with auxiliary information to provide detailed, population-wide insights.

Correcting selection bias in non-probability two-phase payment survey

Chen, H., *Bank of Canada; Tsang, J., University of Ottawa*

We develop statistical inferences for a non-probability two-phase survey sample when relevant auxiliary information is available from a probability survey sample. To reduce selection bias and gain efficiency, both selection probabilities of Phase 1 and Phase 2 are estimated and two-phase calibration is implemented. We discuss both analytical plug-in and pseudo-population bootstrap variance estimation methods that account for the effects of using estimated selection probabilities and calibrated weights. The proposed method is assessed by simulation studies and used to analyze a non-probability two-phase payment survey.

Data-driven sampling: A new framework for correcting political bias in probability-based online panels

Barton, J., Social Research Centre; ***Neiger, D.****, Social Research Centre; Phillips, B., Social Research Centre; Ward, A. C., Social Research Centre*

Survey sampling methodologies must evolve to address the challenges posed by the contemporary survey research environment. The Voting Adjusted Sample Selection (VASS) method is a novel approach designed to mitigate political bias in probability-based online panel surveys. This paper presents an evaluation of VASS in the context of social research demonstrating its efficacy in reducing bias while maintaining weighting efficiency.

Probability-based online panels often exhibit systematic political skew, with participants leaning towards the political left, leading to discrepancies between survey estimates and population benchmarks. VASS corrects for this bias by incorporating a voting-adjusted sampling weight that combines demographic and electoral benchmarks. The method applies a probability-based selection process proportional to these adjusted weights, ensuring that the final survey sample better reflects the electorate's composition without large reductions to the effective sample size that we have observed when applying weighting corrections on their own.

The method was tested on Life in Australia™, the only probability-based online panel of 10,000 randomly recruited members in Australia, by sub-sampling from Life in Australia™ to deliver 2,000 completed surveys. These simulations demonstrate that VASS significantly reduces bias in voting-related estimates without introducing distortions in other outcome variables or major reduction to effective sample size. Compared to traditional stratified random sampling, VASS achieves a bias reduction of up to 9.3 percentage points for key outcomes correlated with political attitudes. Additionally, incorporating voting benchmarks into post-stratification further enhances the accuracy of estimates.

The results indicate that VASS provides a statistically rigorous approach to improving survey accuracy in social research for sub-samples of Life in Australia™. Its application ensures that survey statistics remain relevant in an era where digital data proliferation challenges traditional sampling frameworks. Furthermore, this approach allows for seamless integration with existing weighting methodologies, offering a robust mechanism for addressing political skew without compromising inferential validity. The main limitation of VASS approach is that its impact on political biases reduces as size of the sub-sample approaches the size of the panel and will vary depending on the composition of the specific panel.

Given the increasing reliance on survey data for economic, social, and policy decision-making, methodologies such as VASS will be crucial in shaping the future of survey statistics. By leveraging data-driven techniques, this approach contributes to the broader discourse on adaptive survey methodologies in the digital age. Future research should explore its applicability across diverse domains and other probability-based online panels, to assess its broader impact on data accuracy and representativeness.

The use of meta-information in adjusting non-ignorable selection bias

***Arletti, A.**, University of Padua; Si, Y., University of Michigan; Letizia, M. L., University of Padua; Paccagnella, O., University of Padua*

Non-probability samples such as online panels and social media are goldmines for data, but they suffer from considerable selection bias. We argue that selection bias of this kind is often non-ignorable, that is, it depends on the target variable. Adjusting for nonignorable missingness / selection usually requires some prior knowledge of the selection mechanism. Such knowledge can usually be expressed through a set of assumptions and by setting values of hyper parameters (meta-information). We explore two different approaches in this sense: Proxy-Pattern Mixture Models (PPMM) and Gamma-biased sampling (GbS). PPMM is a relatively simple and effective method which relies on a single meta-information hyperparameter, which indicates how "ignorable" the selection mechanism is believed to be. GbS is a relatively less known approach in survey methodology and has strong links with stochastic optimisation and neural networks. It assumes the sample and population distributions differ up to a known value, and it relies on two meta-information hyperparameters. We employ a set of simulations to compare the two approaches, as well as an empirical example of the Italian National elections of 2022. We examine how these two approaches adjust for different types of selection bias which can be encountered in practice. In addition, we compare the performance of these two approaches against the most common adjustment methods in the literature. Finally, we provide specific recommendations on specific settings in which one approach can be more advantageous than the other.

Sampling strategies for spatial phenomena leveraging auxiliary information

***Bocci, C.**, University of Florence; Rocco, E., University of Florence*

Geographical data often exhibit spatial patterns and an uneven distribution across a study area. Spatial observations are typically not mutually independent; they tend to share similarities with neighboring observations. For this kind of data, selecting the units well spread spatially over the study area allows the collection of more information and consequently provides a more accurate estimate of a mean or total for a target variable. A key sampling problem is thus to spread the sampled units in space as best as possible, a condition described as spatially balanced.

At the same time, the development and diffusion of tools such as satellites, drones, and other systems that enable the collection of wall-to-wall information across a territory provide a wealth of valuable information on statistical units. However, this information often cannot be used directly because it may not be able to answer the specific questions under study and/or could be influenced by measurement errors or self-selection bias. Nevertheless, it can be effectively used as an auxiliary variable for an ad hoc survey to produce estimates of the target variable in a more cost-efficient manner.

Incorporating auxiliary variables in the design and estimation phases relies on assumptions, which are often implicit, regarding the relationship between the response and the auxiliary variables. The effectiveness of enhancing the estimation efficiency through these variables depends on the validity of these assumptions.

Based on these considerations, choosing a sampling strategy must take into account two key factors. First, in the context of spatially related data, the relationship between the study and auxiliary variables may be partially or wholly influenced by spatial covariability. Second, the exact relationship between the target and auxiliary variables is never fully known. To address these challenges, Bocci et al. (JABES, 2024) proposed a sampling strategy that involves a two-step spatially balanced sampling process and an estimator based solely on the values of the sampled study variable. In the first step, the sample is used to assess the nature and strength of the relationship between the auxiliary and response variables, evaluating the potential benefits of using auxiliary data in the sampling design. If the data proves useful, it is then incorporated in the second step.

We propose a possible extension of their sampling strategy by using the information gathered during the first step for a broader assessment. Specifically, the data from the first-step sample will be used to determine whether incorporating auxiliary information during the estimation phase would be beneficial, either as a supplement to or as a replacement for its use during the design phase.

Results from an extensive simulation study indicate that when the relationship between the auxiliary and response variables is linear, the optimal approach is to select a well-spread sample and incorporate the auxiliary variable only in the estimation process. However, when the relationship is non-linear, simply including the auxiliary information in the estimation process does not fully compensate for its absence in the sampling design.

An automated approach for tuning precision constraints in an optimal allocation problem

*Barcaroli, G.; **Bombelli, I.**, ISTAT; Fasulo, A., EUROSTAT; Guandalini, A.,ISTAT; Sacco, G., ISTAT; Terribili, M. D., ISTAT*

When performing a stratified sampling design, once strata are defined, the optimal allocation in a multi-domain framework is provided by the methodology proposed by Bethel (1989). The allocation problem is then formulated as a minimization problem, where the overall survey cost—typically measured in terms of sample size—is minimized while satisfying a set of precision constraints. These constraints ensure a specified level of accuracy for the estimates of key variables across different domains, measured in terms of Coefficient of Variation (CV).

In practice, in official statistics, the sample size is often predetermined due to budgetary and logistical constraints. As a result, the optimization problem shifts to identifying the set of precision constraints that leads to the required sample size. This search can be highly time-consuming and depends on the expertise of the survey statistician. In addition, multiple sets of precision constraints may yield the same sample size.

In this work, we propose criteria to guide this selection process, aiming to make it more objective and efficient.

Optimal allocation can deviate significantly from proportional allocation, which may pose challenges for the on-field organization of data collection, as data collection networks are typically structured according to population distribution. Therefore, one valid criterion is to identify the most proportional among the optimal allocations, whether at the stratum level or domain level (i.e., aggregations of strata). Furthermore, when manually searching for the set of precision constraints that yield the required sample size, it is common practice to focus on only one or two key variables among all for simplicity. Therefore, the resulting optimal allocation will be strongly influenced by only these few variables. Instead, all the variables of interest in the allocation problem should be taken into account and should influence the resulting optimal allocation, at least in the absence of further evidence to justify otherwise.

Therefore, another valid criterion could focus on finding an optimal allocation in which all variables have an equal impact on the allocation process. This could be achieved, for example, by examining the sensitivities -that is, the number of additional sampling units needed to reduce the CVs for each variable of interest in each domain by 10%, and aiming to minimize their variability.

To identify the set of precision constraints that, when applied to the allocation problem, leads to an optimal allocation that meets the required sample size, satisfies the precision constraints, and one of the previously described criteria, the genetic algorithm available in the R package QGA (Barcaroli, 2024) has been used. The algorithm utilizes a fitness function that varies based on the chosen criterion.

The genetic algorithm is highly efficient, significantly speeding up the search for the set of precision constraints. It also ensures that the process is automated, totally objective and independent of the expertise of the survey statistician.

The whole procedure is implemented in R and will be embedded in the R package R2BEAT (Barcaroli et al., 2023), an R package for the optimal allocation of a sample already available on the R-CRAN.

Review of procedures for informative sampling

***Eideh, A.A.**, Al-Quds University*

Several approaches to account for an informative design in survey sampling exist. Skinner (1994) extracts the model holding in the population from models identified and fitted to the sample data. Pfeffermann Et. al (1998) identify the sample model from the population model and a model for the conditional expectation of sampling weights given the analysis variable and auxiliary variables. Magee Et. al (1998) consider an estimator that uses more structure on the population density imposed by modelling the process generating the first order inclusion probabilities. Eideh (2002) and Eideh & Nathan (2009) extract the sample models for two-stage cluster sampling from a finite population when the sampling design for both stages is informative, in which the weights at both stages are incorporated via the sample distribution which is determined completely by the population distribution and the conditional distribution of the sampling weights given the response variable. You & Rao (2002) develop a pseudo-empirical best prediction approach that substitutes simple sums with weighted sums when forming predictors and parameter estimators. Pfeffermann & Sverchkov (2007) specify a parametric model for the mean of the survey weight in the context of small area estimation. Verret Et al. (2015) develop an augmented model approach in which the weight is incorporated as a model covariate. Kim and Wang (2023) specify a beta distribution for the first order inclusion probability, they call the model the beta-prime model because the implied distribution for the sampling weight is a beta-prime distribution. Berg and Eideh (2024) develop procedures within a broad class of exponential dispersion families with random small-area effects and consider two models for survey weights. They also construct predictions for means and more general parameters that are nonlinear functions of the model's response variable.

In this paper, we discuss the treatment of nonignorable nonresponse in surveys as informative sampling and we prove the equivalence of some of different procedures. In addition to that, the problem of selection bias in nonprobability samples can be treated in the light of informative sampling.

Calibration with bagging of the principal components on a large number of auxiliary variables

***Hasler, C.**, University of Neuchâtel & University of Zürich; Tripet, A., University of Neuchâtel; Tillé, Y., University of Neuchâtel*

Calibration, introduced by Deville and Särndal (1992), is a widely used method in survey sampling to adjust weights so that estimated totals of some chosen calibration variables match known population totals or totals obtained from other sources. When a large number of auxiliary variables are included as calibration variables, the variance of the total estimator can increase, and the calibration weights can become highly dispersed. To address these issues, we propose two approaches inspired by bagging and principal component decomposition.

In both of our approaches, samples of calibration variables are selected without replacement from among the candidate calibration variables. For each sample, a system of weights is obtained. The final weights are the average weights of these different weighting systems. Two alternatives are proposed for the selection of the samples of calibration variables: a selection with equal probability among all the auxiliary variables, and a selection with unequal probabilities among the principal components of the auxiliary variables. For both alternatives, it is possible to calibrate exactly for some of the main auxiliary variables. For the other auxiliary variables, the weights cannot be calibrated exactly. While the first proposed alternative is not conclusive, the second proposed alternative allows us to obtain a total estimator whose variance does not explode when new auxiliary variables are added and to obtain very low scatter weights. Finally, this alternative allows us to obtain a single weighting system that can be applied to several variables of interest of a survey. An estimator of the variance of the total estimator is also proposed.

We evaluate the proposed total and variance estimators using a simulation study on real survey data from the Swiss Survey on Income and Living Conditions (SILC). The results show that the second alternative significantly reduces the weight variability and the variance of the total estimator compared to competing total estimators for some variables of interest. In particular, this alternative reduces the variance of the total estimator by up to about 25% compared to the Horvitz-Thompson estimator in the case of the SILC data. In addition, our variance estimator accurately estimates the variance of the proposed total estimator.

Two-step calibration of design weights under dual frame surveys

***Kumar, S.**, Banaras Hindu University; Rai, P. K., Banaras Hindu University*

The calibration approach in survey sampling, introduced by Deville and Särndal (1992), adjusts the design weights used by the Horvitz-Thompson estimator for the population total and has been widely adopted over the last three decades. Building on this, Singh and Sedory (2016) and Alka et al. (2019) have developed a two-step calibration approach under a single-frame survey.

Multiple-frame surveys help reduce sampling costs and address under-coverage issues in single-frame designs. Elkasabi et al. (2015) introduced the Joint Calibration Estimator (JCE) under dual-frame setup, and Rai et al. (2020) have extended this work to find a more efficient estimator of the population total.

This paper introduces a novel two-step calibration estimator for dual-frame surveys to estimate the population total. In the first step, calibration weights are set proportional to the design weights; in the second step, the proportionality constants are calculated using bias reduction and mean squared error (MSE) minimization techniques. The proposed approach provides a flexible and efficient way to integrate auxiliary information across both frames while improving estimation accuracy. The performance of the proposed estimator is also evaluated through simulation studies and a real-world application. Results demonstrate that the two-step method yields lower bias and MSE than existing estimators, making it a valuable tool for practitioners working with complex survey data.

References

Deville, J. C., & Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. Journal of The American Statistical Association, 87(418), 376-382.

Singh, S., & Sedory, S. A. (2016). Two-step calibration of design weights in survey sampling. Communications in Statistics-Theory and Methods, 45(12), 3510-3523.

Alka, Rai, P. K., & Qasim, M. (2019). Two-step calibration of design weights under two auxiliary variables in sample survey. Journal of Statistical Computation and Simulation, 89(12), 2316-2327.

Elkasabi, M. A., Heeringa, S. G., & Lepkowski, J. M. (2015). Joint Calibration Estimator for Dual Frame Surveys. Statistics in Transition New Series, 16(1).

Rai, P.K., Tikkiwal, G. C., & Alka, (2020). A Joint Calibration Estimator of Population Total Under Minimum Entropy Distance Function Based On Dual Frame Surveys. In Statistical Methods and Applications in Forestry and Environmental Sciences, 125-150.

A neutrosophic general class of calibration estimators for population median using stratified sampling in presence of non-sampling errors

Pandey, M. K., *Indian Institute of Technology - Indian School of Mines; Singh, G. N., Indian Institute of Technology - Indian School of Mines*

Neutrosophic estimation presents a pioneering methodology for addressing uncertainty and imprecision in statistical data analysis. Traditional estimation techniques often struggle with data affected by non-response, measurement errors, or inherent vagueness in observations. By leveraging the neutrosophic framework, which extends classical probability by incorporating degrees of truth, indeterminacy, and falsity, we develop a robust approach to estimation.

This paper introduces a neutrosophic general class of estimators designed to estimate the population median while explicitly accounting for the presence of random non-response and measurement errors. The neutrosophic observation is mathematically defined as: Z_N = Z_L + Z_U · I_N, where I_N ∈ [I_L, I_U] and Z_N ∈ [Z_L, Z_U]. Here, Z_N represents the neutrosophic observation, bounded by the lower and upper limits Z_L and Z_U, respectively. The term I_N is a neutrosophic component, reflecting the indeterminate nature of the observation, and it takes values within the interval [I_L, I_U]. This formulation enables a flexible representation of uncertainty, distinguishing between known and indeterminate parts of the data.

To enhance the efficiency of median estimation under a stratified two-phase sampling framework, we employ auxiliary variables and implement calibration techniques. These methods adjust the strata weights dynamically, optimizing the estimation process by incorporating available auxiliary information. The calibration approach ensures that the estimator remains robust even in the presence of substantial uncertainty, random non-response, and measurement errors.

Furthermore, we validate the proposed methodology through extensive simulation studies and a real-world application concerning the impact of climate change on rice production, utilizing the dataset from Aslam (2022). The results indicate that the neutrosophic estimator consistently reduces the mean squared error (MSE) compared to conventional estimation techniques, demonstrating its practical effectiveness.

Finally, the paper provides actionable recommendations for survey practitioners and statisticians on how to integrate neutrosophic estimation into real-world survey designs. By adopting this approach, researchers can improve the reliability of statistical inferences drawn from incomplete or imprecise data.

References

Aslam, M. (2022). Aggregative effect on rice production due to climate change using index number under indeterminate environment: a case study from Punjab, Pakistan. Theoretical and Applied Climatology, 147(1), 283–290.

Accurate Inference for SAE models

***Victoria Feser, M.-P.**, University of Bologna; Orso, S., University of Geneva; Guerrier, S., University of Geneva; Karemera, M., University of Geneva*

Firstly defined through the nested error linear regression model, unit-level SAE models have evolved into more sophisticated and flexible forms such as Generalized Linear Mixed Models (GLMM). While model-based estimators have smaller variances (compared to direct survey estimators), they tend to be biased, mainly affecting inferential procedures based on them, such as bootstrap methods or the ones based on asymptotic Mean Squared Prediction Error (MSPE). The latter, which needs to be estimated, can provide quite liberal inference, especially when the number of areas is relatively small. We propose instead an alternative inferential approach by adapting recent results in [1]\cite{IB-inference:24} who derive a simulation-based framework for inference based on a method called Implicit Bootstrap (IB). Building upon this approach, we propose a general method that (1) facilitates the construction of accurate CIs and (2) maintains flexibility to apply across different models without requiring model-specific derivations of asymptotic MSPE. Basically, the IB produces an empirical distribution from which, for example, Confidence Intervals (CI) can be constructed using the percentile method which is invariant to parameter transformations and is shown to be second order accurate. It is based on estimators for the chosen model (e.g. the GLMM), with importantly reduced finite sample bias, which are used to compute the predicted responses. Since the latter are functions of these estimators, the IB is particularly suited to SAE models.

References

[1] Orso, S., M. Karemera, M.-P. Victoria-Feser, and S. Guerrier (2024). An accurate percentile method for parametric inference based on asymptotically biased estimators. arXiv:2405.05403v2.

A dynamic adaptive sampling approach for spatial data

***Altieri, L.**, University of Bologna; Cocchi, D., University of Bologna*

Adaptive sampling methodologies represent a promising avenue for improving the efficiency and precision of spatial surveys, particularly when traditional spatial balanced methods falter in the presence of non-compact or repulsive spatial patterns. This study addresses the substantive challenge of designing adaptive spatial sampling strategies that dynamically account for spatial association, enhancing estimates of spatial metrics and population totals.

We propose and evaluate the Dynamic Adaptive Sampling Approach (DASA), a novel methodology, starting from Spatial Correlated Poisson Sampling, that iteratively refines sampling probabilities based on observed spatial correlation structures. DASA incorporates measures such as Moran's Index and Spatial Information to adaptively guide the selection of sampling units, prioritizing areas of high heterogeneity while reducing redundancy in more homogeneous regions. By comparison, we also analyze a number of spatially balanced sampling methods, Simple Random Sampling and the two-phase Preliminary Sampling Approach (PSA), which pre-determines spatial correlation metrics in an initial sampling phase before conducting sequential sampling.

The DASA methodology begins with an initial random selection phase to establish baseline estimates of spatial association metrics. It then sequentially updates the inclusion probabilities for unsampled units, leveraging measures of spatial correlation observed during the survey. This iterative process adapts sampling efforts to the spatial heterogeneity of the study area, concentrating resources in regions where spatial association measures indicate greater variability or significance. By dynamically refining inclusion probabilities, DASA ensures a more precise alignment of sampling efforts with the underlying spatial structure of the data.

Our application focuses on point pattern data from Norwegian spruce trees, characterized by a weakly repulsive spatial structure. Through a large number of data samples on a discretized grid, we compare DASA and PSA against conventional spatially balanced sampling methods. Key performance metrics include the precision of spatial association estimates, bias and variability in population total estimates, and computational efficiency.

The two-phase PSA is substantially more efficient than spatially balanced methods and SRS in providing good estimates of the total or mean of the study variable over a variety of scenarios, including the specific case study considered in this work. The results demonstrate that DASA consistently outperforms PSA, offering more precise and less biased estimates for both Moran's Index and Spatial Information across varying distance classes. DASA leverages all sampled data for sequential refinement, eliminating the need for a preliminary phase and thereby reducing field survey costs and time. Although computationally intensive, DASA's iterative updates produce superior precision in estimating spatial metrics and population totals, even under complex spatial dependencies.

In conclusion, DASA emerges as a robust and resource-efficient methodology for adaptive spatial sampling. Its flexibility and precision make it suitable for diverse applications, including environmental monitoring, urban planning, epidemiology, and socio-economic studies. Future research could further optimize computational efficiency and explore extensions to clustered or heterogeneous spatial patterns, broadening the applicability of adaptive sampling strategies.

ISTAT study on new spatial sampling methods to improve survey efficiency

*Benedetti, R., University of Chieti-Pescara "G. D'Annunzio"; **Falorsi, S.**, ISTAT; Loriga, S., ISTAT; Piersimoni, F., ISTAT; Russo, M., ISTAT*

The set of ISTAT's Household Social Surveys, HSS, is currently conducted using a standard sample design, which is common to most of the large-scale surveys on the population carried out by the most important National Statistical Offices (NSOs). Its success lies on the simplicity of implementation and on its ability to guarantee efficient estimates for multi-purpose and multi-domain surveys such as HSS. However, in the wake of the now ten-year decline in response rates that have made the conduct of sample surveys increasingly more expensive, the various NSOs are conducting studies aimed at improving the overall efficiency of this sampling design. One of these ISTAT's studies concerns the possibility of adopting spatially balanced sampling designs (SBSD) for the selection of the first-stage units of the HSS. The underlying idea is that variables that are characterized by positive spatial autocorrelation should benefit, in terms of increased sampling efficiency, from the adoption of a sampling design in which municipalities are selected through a SBSD. This advantage should be greater the higher the spatial autocorrelation coefficient for a specific target variable. In this regard, it is considered important to underline that the Coefficients of Variation (CV) of some of the target estimates are bound by the EU regulations and improving their CVs can be strategic for a NSO. SBSD have traditionally been studied and applied to improve the efficiency of surveys in the agricultural field. However, this type of design had never been studied for surveys in the social field. This aspect therefore represents an important value added of this work. Another important aspect of SBSD is that of ensuring better sample coverage of unplanned domains in the design phase can be very important in order to be able to produce good quality direct estimates for this type of domains. This work, therefore, represents an important innovation and advancement in applied research on this topic. The work uses an aggregate database of known municipal totals from administrative sources (containing a large spectrum of socio-economic variables) and on this it carries out a simulation study, for the sample design of the LFS survey (representative example of the entire set of ISTAT's HSS), aimed at comparing the empirical properties, in terms of bias and variability, of the standard design and of different SBSD proposed in the literature. In particular, the SBSD methods compared are the Local Pivotal Method (Grafström et al. 2012) and the Proportional to the Within Sample Distance (See Benedetti et al. 2017). As expected, SBSD are effective in reducing the CVs of the estimates. These are reduced proportionally to the increase in the value of the Moran index of the target variables considered. These positive results open further research questions e.g. related to defining consistent estimators of the sample variance.

Spatial gradient-oriented transects for sampling rare populations

Carrer, F., *Swedish University of Agricultural Sciences; Prentius, W., Swedish University of Agricultural Sciences; Grafström, A., Swedish University of Agricultural Sciences*

In survey statistics, a large sample is not always required to obtain reliable inferences. However, in the case of a rare population, a larger sample is often required to guarantee a minimum number of observations. The integration of auxiliary variables into the design can help improve the efficiency of the estimators.

We consider the problem of a hard-to-detect population within an area frame. A sufficiently smooth auxiliary variable defines a gradient field on the frame. If the sample collection can be performed as a sequence of units forming a path, a natural choice for trajectories is the field lines tangent to the vector field at any point. We define spatial Gradient-Oriented Transects (GOT) to sample rare populations on the area frame. GOT consists of sequences of units derived by following the gradient field line from a unit initially selected. The design takes the form of an asymmetric link-tracing design, where initially selecting a population unit leads to the inclusion of all the units along the gradient field line. GOT wind around the area frame following the direction of the steepest ascent of the auxiliary variable, resulting in an unequal probability sampling design that assigns higher inclusion probabilities to units with a locally higher value of the auxiliary variable. Since gradient field lines converge to the local maxima of the auxiliary variable, GOT induce a natural stratification of the population, where each stratum consists of the units converging to the same local maximum. By moving along the gradient lines, GOT produces a sample that spans a range of values of the auxiliary variable, capturing variability across different levels.

We derive unbiased Horvitz-Thompson estimators for the population total and the variance. Adaptive Cluster Sampling (ACS) is extended to the case of primary units chosen according to the GOT. The estimators are revised to account for the resulting multiplicity of the population units.

We conduct numerical simulations in the context of forest damage. Forest damage often occurs as rare and clustered populations in the landscape, and a quick response is required to quantify the disturbance for ecological and financial purposes. Since GOT takes the form of trajectories on the area frame, the design is compatible with data collection with the use of drones for data collection over forest stands. Risk maps for forest damage, such as spruce bark beetle infestations or fallen trees due to wind and storms, are used as auxiliary variable. Numerical results show that GOT can provide a lower Mean Squared Error (MSE) than traditional transects with a fixed shape selected with probability proportional to size.

A non-probabilistic sample survey to study the regional variability in attitudes towards homosexuality: Survey design and weighting aspects

***Moretti, A.**, Utrecht University; Salvatore, C., Utrecht University; Meitinger, K., Utrecht University*

Previous research revealed a puzzling finding in the measurement of attitudes towards homosexuality: In countries with extensive discrimination towards homosexuals, respondents were not aware of discrimination. In contrast, respondents in more tolerant countries report discrimination more frequently. At the same time, the countries also showed regional variability in the perceived discrimination of homosexuals. Respondents across and within countries might interpret the questions differently and survey measures are potentially misleading. Web probing is a crucial tool to reveal variations in respondents' associations and/or silent misinterpretations when answering survey items. So far, web probing studies compared respondents' associations across countries but disregarded potential regional variabilities of results. The combination of web probing with regional analyses provides a novel approach to study within and across country variability of respondents' associations.

In our project, we considered the following countries: Germany, France, Italy, Poland, The Netherlands, and the United States. The data collection was based on a quota (non-probabilistic) sample; therefore, important points need to be addressed here. In this presentation we will discuss the study from a total survey error framework perspective. In particular, we will pay particular attention to the data editing and weighting steps for cross-country comparisons. We will compare different weighting approaches we adopted in a cross-country context. These are based on European probabilistic sample surveys and/or known population benchmarks available via Official Statistics archives. For the United States we consider the General Social Survey and population tables available generated from the Census. This research is part of the Dutch Research Council (NWO) funded project "Regional validity of survey questions on attitudes towards LGBT: combining web probing with small area estimation" (Grant number: 406.XS.01.081).

Alternative selection mechanisms in online surveys

Prücklmair, F., Universität Bamberg; ***Rendtel, U.****, Freie Universität Berlin*

The rapid changes in internet accessibility and the evolving ways it is used could raise doubts about whether earlier findings on selection effects in internet surveys still hold true. We address the following questions: 1) Is internet access alone still a reasonable self-selection criterion? 2) If this is no longer true, how should alternative self-selection processes be modeled? 3) How can these selection processes be controlled? Are demographic control variables sufficient to establish the Missing at Random (MAR) condition, or is it possible to establish the MAR condition with more powerful control variables? 4) To what extent do weighting procedures correct self-selection bias? We investigate these questions in the setting of a simulation study where we assume four different selection models. These involve the length of internet use, posting behavior, and interest in politics and are based on theoretical considerations. We use the European Social Survey (ESS) as a simulation environment, which contains these variables and demographic background variables. It also includes our outcome variable, the vote in the 2017 Bundestag election in Germany. In order to judge the differences between the non-probability results and the simulated universe, we compare the ESS estimates of the 2017 Bundestag election with the real election results.

Harnessing social media for innovative data collection: A Reddit scraping and data-cleaning pipeline

Stracqualursi, L., *University of Bologna*

This paper introduces a novel approach to data collection from social media by presenting a fully automated pipeline that scrapes and cleans data from Reddit. Designed to complement traditional survey methods and administrative data, the workflow efficiently extracts large volumes of user-generated content, identifies relevant submissions via targeted keywords, and systematically cleans the resulting text for subsequent analysis. Through customized modules that remove duplicates, standardize linguistic features, and handle domain-specific stopwords, the pipeline produces a high-quality dataset readily applicable to sentiment analysis, topic modeling, or other statistical techniques. By capturing timely, organic discussions on rapidly evolving topics, this method offers a valuable supplement to conventional data-collection strategies for researchers, policymakers, and statistical agencies seeking richer, more immediate insights from online communities.

Assigning unit weights for linked data

***Liu, A.-C.**, Utrecht University; Lugtig, P., Utrecht University; Scholtus, S., Statistics Netherlands; De Waal, T., Statistics Netherlands, Tilburg University*

Linked data from different data sources have been widely used in many research areas. Often in data infrastructure, the data linkers are separated from the secondary analysis researchers for disclosure reasons. The researchers may only see the final linked data set but have little or no idea of the original data sources. However, the quality of the linked data is dependent on the quality of the original data sources and the quality of the linking process. If, for example, the probability for a unit being included in the linked data set is related to the study variables, treating the linked data as a simple random sample may lead to biased inference.

To avoid biases, one choice may be producing some quality indicators, such as (estimated) success linked rate, of the linked data set. However, the relation between the quality indicators and the final inference is not straightforward and may be affected by the research questions in mind. For example, if the goal is to detect possible fraud in the linked data set, having a set of correct links with certainty is crucial. On the other hand, if the goal is to observe the relationship between variables, having a complete range of the variables may be more of interest.

Therefore, to offer flexibility for secondary analysis, we propose producing a set of unit weights alongside the linked data set. The weights are constructed based on the representativeness and the linkage quality of a unit. The researchers may then incorporate the weights in the secondary analysis by, for example, having a sub-selection of units given the quality or performing a weighted estimator.

Machine learning approaches for the estimation of response probabilities in the weighting procedure of EU-SILC in Austria

***Glaser, T.**, Statistics Austria*

The EU-SILC (EU Statistics on Income and Living Conditions) survey is the main source of data on poverty, social inclusion and income of private households in Austria. From 2004 onwards, a rotational sampling design with four sub-samples has been carried out on a yearly basis as a survey with voluntary participation. For the first waves of each rotation, a probability sample is drawn from a sampling frame based on the central residence register. The weighting procedure of EU-SILC accounts for sampling (design weights), participation (unit non-response weights) as well as coherence with external data sources (calibration). Estimating the probability of response is a central part of the weighting procedure since the inverse of these estimates is used as factors to adjust so-called “base weights” from the previous year to unit non-response. This weighting step is intended to counteract selective unit non-response due to non-participation of certain socio-economic groups in the survey. A broad variety of characteristics from the sampling frame or the results of the previous years are available to estimate the probability of response using logistic regression models. The selection of characteristics to estimate response is carried out automatically. The presentation compares different machine learning procedures (such as stepwise selection or lasso) that have been applied to estimating the probability of response in different waves of EU-SILC in Austria. An important challenge hereby is posed by selecting models with good predictive performance that also do not lead to overly dispersed base weights. Especially very small estimated probabilities signify large adjustment weights that may also lead to extreme base weights. Although some extreme values may be truncated, a feasible approach is to build models that explain and estimate response probabilities well and at the same time lead to a comparatively low coefficient of variation of the resulting weights. The main motivation for aiming at a lower dispersion of base weights is that the precision of the EU-SILC survey is measured by the standard error of the main indicator: the rate of people at risk of poverty or social exclusion (AROPE). Highly dispersed base weights usually also lead to larger standard errors of the weighted results of the main indicators of EU-SILC. In addition to taking the inverses of the estimated response probabilities as weight adjustment factors, the score method and response propensity class adjustment are presented as options. These methods are combined with grouping estimated response probabilities into homogeneous classes by k-means clustering before applying them to unit non-response adjustment weights. Results speak in favor of more sparse models for estimating response probabilities based on variable selection by lasso that still perform well in terms of model evaluation metrics. Regarding the adjustment of base weights to unit non-response, a combination of the lasso method for logistic regression and the score method is preferred.

Using administrative data and machine learning for nonresponse weighting in official statistics: Evidence from the icelandic labour force survey

Einarsson, H., University of Iceland; ***Sakshaug, J.W.****, IAB & LMU-Munich*

Declining response rates in official surveys, such as labor force surveys, increase the risk of nonresponse bias in the estimation of key target variables. To address this problem, survey practitioners often rely on high-quality administrative records, such as data drawn from administrative registers, in survey design and estimation. In this paper, we analyze data from the Icelandic Labour Force Survey (IS-LFS), a register-based telephone sample survey of named individuals conducted over a period of 21 years (2003-2023), to evaluate the effectiveness of several machine learning algorithms for nonresponse bias adjustment relative to standard practice that relies on parametric logistic regression. Furthermore, given that the effectiveness of weighting adjustments hinges on the availability of auxiliary information that correlates with both the propensity to respond to surveys and the target variables of interest, we examine whether nonresponse bias is reduced by the inclusion (or omission) of register-based covariates. As response rates in the IS-LFS have declined during the study period, we also investigate whether the effectiveness of survey weights has changed over time. Our findings suggest while the use of machine learning yields only modest gains over logistic regression methods, identifying and selecting predictors that correlate with both the propensity to respond, and the target variable is more important when deriving weights for official surveys.

Consistent estimation of linear regression models from different data sources with many variables in common

***Hirukawa, M.**, Ryukoku University*

When conducting regression analysis using a dataset of anonymous surveys ("primary dataset"), researchers often face the situations in which some regressors are unavailable. Examples of missing regressors in economics include parental income in the study of intergenerational income mobility, ability measure in estimating returns to schooling, housing wealth in the determination of consumer expenditures, and media exposure when explaining consumers' purchase behaviors, to name a few. Suppose that there is yet another dataset of anonymous surveys ("auxiliary dataset") that contains 'missing' regressors as well as other variables common across two datasets ("overlapping variables"). Under this environment, it is possible to estimate regression coefficients consistently by combining primary and auxiliary datasets. Examples of such estimation procedures are the matched-sample indirect inference ("MSII"; Hirukawa and Prokhorov, 2018: J. Econom., 203, 344-358) and the plug-in least squares ("PILS"; Hirukawa et al., 2023: Econom. Rev., 42, 1-27). Although both MSII and PILS are built on ordinary least-squares ("OLS") estimation, there is a difference in the way of constructing proxies for missing regressors nonparametrically. MSII directly imputes values of each missing regressor from the auxiliary dataset that are chosen via the nearest-neighbor matching ("NNM"), whereas PILS replaces each missing regressor with a kernel-smoothed estimate of its conditional expectation given overlapping variables.

The curse of dimensionality is a common concern across MSII and PILS. If the primary dataset with the sample size n were complete, we could run OLS and attain root-n consistency in regression coefficient estimators. If proxies for missing regressors are generated nonparametrically using original overlapping variables, then we must confine the number of the latter to three or less to make MSII and PILS estimators root-n consistent. Then, we extend the scope of MSII and PILS so that both can restore the parametric convergence rate when two datasets have many overlapping variables. The extension takes three steps. The first step is dimension reduction. There are two possible approaches, depending on the structural assumption on the conditional expectation of each missing regressor given overlapping variables that we impose to mitigate the curse of dimensionality. If a semiparametric single-index model with an unknown link function is presumed, then index coefficients can be estimated with the assistance of a few algorithms of sufficient dimension reduction ("SDR"). Alternatively, predicted values of each missing regressor may be generated under the assumption of additive nonparametric regression ("ANPR"). The second step is imputation. Estimated indices by SDR can be used either as matching variables for NNM in MSII or as regressors for kernel regression estimates in PILS. Predicted values by ANPR become proxies of missing regressors in PILS. The third and final step is estimation of the regression model of interest. This can be done by MSII or PILS, depending on the imputation method in the previous step. Convergence properties of extended MSII and PILS are explored in conjunction with covariance estimation. Monte Carlo simulations confirm their nice finite-sample properties, and a real data example of intergenerational income mobility is also presented.

Including the spatial balance criterium into the estimator

*Benedetti, R., University of Chieti-Pescara "G. D'Annunzio"; **Pantalone, F.**, University of Southampton; Piersimoni, F., ISTAT*

Spatially balanced sampling designs have been introduced to capture the spatial heterogeneity of the target population. These designs aim to randomly select a sample well spread over the region of interest, and simulations showed that the Horvitz-Thompson estimator performs considerably better in terms of variance when these designs are employed (compared to simple random sampling which does not take into account any spatial structure) and when the population has a spatial structure. Additionally, investigations showed that better performances could be achieved when the spread increases, which can be measured by the Spatial Balance Index, which is an index based on Voronoi polygons. Therefore, spatially balanced designs provide samples with a good level of spatial balance.

In this paper, we include the spatial balance criterium into the estimator. To this end, once a sample is selected, a Voronoi polygon is constructed for each sample unit, and a new weight is then computed for each sample unit based on the corresponding Voronoi polygon. This can be seen as an endogenous post-stratification, where the strata are random as they depend on the selected sample.

Monte Carlo simulations were carried out to investigate the performance of the proposed estimator in different settings, where different spatial trends, different levels of spatial dependence and different spatial distributions of the population units were considered. Results suggest that the proposed estimator performs well when population units are spatially dependent, with performance increasing when the spatial dependence increases (and the trend is stronger), while efficiency is lost if there is no spatial dependence. Therefore, in case of spatially dependent population units, the proposed estimator provides an improvement upon the Horvitz-Thompson estimator coupled with a spatially balanced sample, and it can also provide a way to increase efficiency when the sampling design employed is not spatially balanced.

Producing representative data in a monthly employment panel in Germany: Challenges and approaches in a longitudinal context

***Seesing-Coelho, C.**, Federal Statistical Office of Germany & University of Hamburg; Jaeger, C., Federal Statistical Office of Germany; Preising, M., Federal Statistical Office of Germany*

Germany's monthly earnings survey (Verdientserhebung) is a key resource for labor market research and will soon be made available to the scientific community also as a longitudinal dataset. This panel, based on reports from approximately 55 000 firms covering all their employees, provides monthly information on about 9 million workers and offers a highly reliable source of income data, as it is mandatory and derived directly from pay slip records. The availability of such high-frequency, high-quality wage data opens new possibilities for analyzing labor market dynamics and conducting causal inference studies. However, transforming this survey into a panel suitable for economic analysis presents several methodological challenges.

One of the primary challenges arises from the sample design. The Verdiensterhebung follows a disproportionate sampling strategy combined with a rotational panel structure, making appropriate weighting indispensable for producing unbiased estimates. Another crucial challenge is the definition of a longitudinal population that allows for meaningful calibration to external data sources while maintaining representativeness over time.

To achieve representativeness, the panel undergoes a generalized calibration procedure that adjusts firm weights so that key totals match external benchmarks from the German Federal Employment Agency (Bundesagentur für Arbeit). This calibration is designed to ensure alignment in the number of firms, of workers subject to social security contribution, and of those in marginal employment (currently earning less than €556 per month), across federal states, firm size classes, and economic sectors. A key feature of this approach is the possibility of introducing a small tolerance threshold in matching totals, which in this case is used only for the totals of marginally employed. This flexibility is necessary due to slight inconsistencies between the definitions of employment used in the panel and in the registry data. Allowing for minor deviations in these totals enables the algorithm to converge efficiently without requiring further manual alterations, optimizing the data production process.

A further adjustment is necessary to correct for non-response bias. Since some active firms fail to report consistently, logistic models are applied monthly to adjust weights. This approach ensures proper weighting for businesses that stay in the sample and those that return during the panel, compensating for systematic non-response patterns.

Beyond these weighting challenges, the normal dynamics of firm evolution introduce additional complexities. Businesses experience structural changes such as openings, closures, mergers, splits, and fluctuations in workforce size. When firms transition between strata, for example going from a smaller size class to a larger one, their original sample weights become invalid, potentially leading to misrepresentation in employment and wage statistics. Addressing these issues requires meticulous case-by-case analysis and continuous refinement of the weighting strategy to prevent over- or under-representation of specific firm types and their employees.

By systematically tackling these methodological challenges, the panel will offer researchers a robust tool for studying labor market dynamics and wage developments in Germany. Ensuring its representativeness through careful weighting and calibration is key to unlocking its full potential for high-quality empirical research in the data-driven era.

STATENT Flash: Rapid estimates for the Swiss structural business statistics

Chalimourda, A., *Swiss Federal Statistical Office; Vallon, N., Swiss Federal Statistical Office; Nedyalkova, D.; Swiss Federal Statistical Office; Assoulin, D., Swiss Federal Statistical Office*

The structural business statistics STATENT (Statistique structurelle des entreprises, in French) provide key information on the structure of the Swiss economy. STATENT covers all enterprises that are obliged to pay contributions to the Old-Age and Survivors Insurance (OASI) either for employees or self-employed persons, with an annual income of at least CHF 2300. It is based mainly on data from the OASI register and information from the business register of the Swiss Federal Statistical Office (FSO), supplemented by ongoing surveys on enterprises. The STATENT which refers to employment in December of a specific year, is currently published 20 months later, partly due to the gradual, quarterly delivery of the OASI data to the FSO during the following year. The aim of the FSO's STATENT Flash innovation project is to provide rapid estimates on enterprises and employment one year before the official STATENT publication, i.e. shortly after the second quarterly delivery of the OASI data.

In the present work, we describe two approaches to provide rapid estimates for the STATENT. The first approach estimates the probability that a unit of the Swiss Business Register at the end of a reference year belongs to the STATENT published 20 months later. No restrictions on the annual income or other conditions are imposed on the initial population of the business register, making this a data-driven approach. Important auxiliary information includes indicator variables such as whether a unit belongs to the second OASI data delivery and to the STATENT of the previous reference year. Assuming that this statistical classification model is also valid for the following year, we use it to predict which units of the considered year's business register will be part of the future unknown STATENT, for which we want to produce the rapid estimates. In an alternative approach, we estimate the probability that a STATENT unit has already been delivered to the FSO one year earlier with the second OASI delivery, i.e. a type of delivery probability. Assuming that the model is also valid for the following year, we use it to assign a probability to each unit present in the second OASI delivery. The inverse of the estimated delivery probabilities is then used to weight these units to account for units not yet delivered. This weighting approach could be improved by adopting practices used in business surveys such as identification and specific treatment of large units, variance estimation and calibration. Finally, we compare the two approaches and discuss their advantages and disadvantages.

Keywords: swiss structural business statistics, rapid estimates, different data sources, prediction, weighting approach.

On some composite estimators for domain estimation under systematic sampling design

***Yadav, G.**, Banaras Hindu University*

Domain Estimation in Survey Sampling refers to the process of estimating population parameters (such as totals, means, or proportions) for specific subgroups, or "domains," within the larger population using data collected from a sample. In many instances, the focus is not on the entire population but on specific subgroups that are of particular relevance. This study presents a family of composite estimators for domain estimation using auxiliary information under a systematic sampling scheme. The proposed estimators are based on modifications of the Bahl and Tuteja (1991) estimator in survey sampling, leading to a generalized class of composite estimators. These estimators integrate exponential-type direct and synthetic ratio estimators, along with different forms of the proposed domain estimators. Theoretical properties of the proposed estimator are also discussed, including expressions for the weights used. Furthermore, the study demonstrates the application of these composite estimators for domain estimation and compares their relative performance to the individual constituent estimators through a simulation study.