METRON (2023) 81:1–3
https://doi.org/10.1007/s40300-023-00248-1

EDITORIAL

Foreword to the special issue on "Survey Methods for Statistical Data Integration and New Data Sources"

M. Giovanna Ranalli[1] · Jean-François Beaumont[2] · Gaia Bertarelli[3] · Nathalie Shlomo[4]

Received: 4 April 2023 / Accepted: 11 April 2023 / Published online: 26 May 2023
© Sapienza Università di Roma 2023

Survey researchers and practitioners have always combined data from various data sources (auxiliary data from Censuses, administrative registers, GIS data) to enhance sampling designs and estimation: from stratification and probability proportional to size sampling to balanced sampling designs (see e.g. [9], for a review), from ratio estimation to model-assisted survey estimation with modern regression techniques [1], from post-stratification to calibration and its extensions [3, 7]. Statistical data integration represents the new frontier of combining information from representative samples with data from multiple sources some of which may have uncontrolled selection mechanisms (non-probability samples, big data, web-scraped data, integrated register data). This special issue contains a first selection of the papers presented at the Seventh Italian Conference on Survey Methodology, ITACOSM2022, held in Perugia in June 2022 with a focus on "Survey Methods for Statistical Data Integration and New Data Sources". ITACOSM is a bi-annual international conference organized by the Survey Sampling Group of the Italian Statistical Society whose aim is promoting the scientific discussion on the theoretical and applied developments of survey sampling methodologies in the fields of economics, social and demographic sciences, official statistics and environmental sciences. In particular, ITACOSM2022 has provided a showcase for methods and applications that combine sources of data for better sampling strategies using classical survey methods or new data science tools such as Machine Learning methods.

The contributions of this Special Issue deal with new challenges in integrating several sources of information in estimation from non-standard sampling designs [2, 4, 5] or for producing official statistics [8, 10], including approaches that use machine learning methods [10], (possibly combined) administrative data [8, 10], or non-probability survey data [2, 4, 6]. All contributions make use of models. The role of probability sample surveys in this new era is discussed in detail in Salvatore [6] and is very diverse among the applications: they are used in a case–control type approach to indirect questioning sensitive items to reduce non-sampling errors [5], integrated with administrative data [10] or with non-probability sample surveys

✉ M. Giovanna Ranalli
giovanna.ranalli@unipg.it

1 Department of Political Science, University of Perugia, Perugia , Italy

2 Statistics Canada, Ottawa, Canada

3 Department of Economics, Università Ca' Foscari, Venice, Italy

4 Department of Social Statistics, University of Manchester, Manchester, UK

[4] to obtain more reliable measurements or estimates, or designed for quality assessment of register based estimates [8].

The Special Issue opens with a study by Salvatore [6] that describes the evolution of the research on integration of probability and non-probability survey data over the years using an original approach based on text mining and bibliometric analysis. A collection of 1,023 documents retrieved from Scopus is analyzed. The literature on this research field is characterized, research trends as well as potential directions for future investigation are discussed together with research gaps which need to be addressed.

The paper by Huang and Breidt [4] fits perfectly in this stream of research. In particular, it proposes to draw inference from a respondent-driven (non-probability) sample using data from a probability sample in a dual-frame estimation setting. Respondent-driven sampling is commonly used for surveying rare, hidden, or otherwise hard-to-reach populations, which is (usually) initiated with a small, non-random sample and relies on respondents themselves to recruit additional participants through their own social networks. Usually inference from respondent-driven samples is based on strong assumptions on the diffusion of sampling through the network. Huang and Breidt [4] consider an alternative setting in which a probability sample is used to initiate the sample and only a few waves of recruitment take place. In this setting, they develop a dual-frame estimator that use both known inclusion probabilities from the initial sampling design and estimated inclusion probabilities from the respondent-driven sample. The proposed dual-frame estimator is then applied to a real respondent-driven sample study of smoking behavior among lesbian, gay, bisexual, and transgender (LGBT) adults.

Fellows and Handcock [2] address networked populations, as well, and focus on the relationship between the individual attributes of interest and the the social structure of the ties in the network. In particular, they model the joint distribution of social ties and individual attributes when the population is only partially observed, as when (i) a network sampling design is used or (ii) (possibly non-ignorable) missing data arise for the attributes and/or the ties. To this end, the paper develops a theory of inference for exponential-family random network models when only part of the network is observed. In addition, the proposed inferential framework is applied to data collected via contact tracing, that is of considerable importance to infectious disease epidemiology and public health.

The paper by Quatember [5] addresses the issue of surveying sensitive variables using indirect questioning designs, which protect the respondents' privacy by masking the sensitive information. In particular, the paper looks at the *item count technique* that integrates information from two independent samples from the population: respondents in the control sample are asked to report the number of a set of non-key items that apply to them, while for those in the treatment sample, the list additionally includes the key item that surveys the sensitive information. The paper proposes to further integrate prior auxiliary information in the estimation process, to improve the estimation accuracy and, at the same time, reduce the respondents' task.

The last two papers of this special issue look at the production of multi-source official statistics. In particular, Varriale and Alfò [10] investigate methods for producing estimates on employment integrating survey data and administrative sources by means of Machine learning methods. The former are drawn from the Labour Force survey conducted by the Italian National Statistical Office, Istat, while the latter include several administrative sources that Istat usually acquires, such as those related to social security and to fiscal data. Machine learning methods, such as decision trees and random forests, are used to predict the individual employment status. The proposed approach proves to be particularly useful to detect instances in which the different sources of data do not agree. Solari et al. [8] focus, on the other hand, on

the extremely relevant issue of developing a statistical framework to evaluate the accuracy of combined and/or fully register based censuses. To this end, the Authors argue that, in this setting, probability sample surveys should be designed for quality assessment and for improving the quality of the register based estimation process, rather than for estimation purposes only. Drawing on similar experiences, they provide a formalization of the population size estimation process fully based on administrative data and an application to the estimation process implemented at Istat.

We would like to thank all the Authors who have contributed to this first selection of papers from ITACOSM2022. Another issue of Metron in the coming months will include a second selection of papers. Warm thanks go also to the Reviewers of the articles, who provided the authors with valuable comments and suggestions. Special thanks go to Marco Alfò, Editor-in-Chief of Metron, who has invited us to edit this Special Issue and has supported us in this work.

References

1. Breidt, F.J., Opsomer, J.D.: Model-assisted survey estimation with modern prediction techniques. Stat. Sci. **32**, 190–205 (2017)
2. Fellows, I.E., Handcock, M.S.: Modeling of networked populations when data is sampled or missing. Metron. **81**(1) (2023)
3. Haziza, D., Beaumont, J.-F.: Construction of weights in surveys: a review. Stat. Sci **32**, 220–226 (2017)
4. Huang, C.-M., Breidt, F.J.: A dual-frame approach for estimation with respondent-driven samples. Metron. **81**(1) (2023)
5. Quatember, A.: Efficient item count techniques with one or two lists. Metron. **81**(1) (2023)
6. Salvatore, C.: Inference for non-probability samples and survey data integration: mapping the literature through text mining approaches. Metron. **81**(1) (2023)
7. Särndal, C.-E.: The calibration approach in survey theory and practice. Surv. Methodol. **33**(2), 99–119 (2007)
8. Solari, F., Bernardini, A., Cibella, N.: Statistical framework for fully register based population counts. Metron. **81**(1) (2023)
9. Tillé, Y., Wilhelm, M.: Probability sampling designs: principles for choice of design and balancing. Stat. Sci. **32**, 176–189 (2017)
10. Varriale, R., Alfò, M.: Multi-source statistics on employment status in Italy, a machine learning approach. Metron. **81**(1) (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.