METRON (2024) 82:1–3
https://doi.org/10.1007/s40300-024-00270-x

EDITORIAL

Foreword to the special issue on “Survey Methods for Statistical Data Integration and New Data Sources: tools and real data applications for official statistics”

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

Received: 17 January 2024 / Accepted: 3 February 2024 / Published online: 19 March 2024

This special issue contains a second 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 official statistics and applied sciences. In particular, ITACOSM2022 provided a showcase for methods and applications that combine sources of data for better sampling strategies.

Complementarily to the first special issue published last April [6], this selection of papers focuses on the applied challenges of data integration and provides a set of tools and diverse case studies. In particular, three papers provide applications of data integration techniques [1, 2, 7], three provide new tools to face issues when applying data integration to real data [3–5], and all papers face issues and/or apply methods envisioning the production of official statistics. The sampling phase of a survey is considered in [3] and in [1], the analysis of data quality is the core of [2], while the estimation step is treated in [4, 5], and in [7], with a particular emphasis on small area estimation in the former two.

Bernardini et al. [1] provides an overview of the role of data integration in the new Italian Permanent Population and Housing Census carried out by the Italian National Institute of Statistics (Istat). The backbone of the new Census is the statistical Population Register, whose main sources are the local population registers of Italian municipalities, while two sample surveys (the so-called area and list surveys) are conducted annually to evaluate and correct the coverage errors of the Population Register and collect the data needed to produce Census outputs. In 2020 the field surveys were cancelled because of the pandemic, and a process of integration of available data from previous waves with administrative signs-of-life was set up in order to estimate population counts at municipal level for age, sex and citizenship. This

✉ 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

push towards a larger use of administrative data has called for a rethinking of the Census design, where survey data is envisioned to be used for quality assessment of fully register-based population count estimation.

In this context, De Vitiis et al. [3] discuss how Responsive-Adaptive Survey Designs can be suited to the context of large-scale social surveys and, particularly, to the Permanent Population and Housing Census. The latter is the largest and most expensive sample survey carried out yearly by Istat. In order to reduce the overall cost of the survey, data collection is already based on a sequential mixed mode design (CAWI-CAPI). The paper investigates the use of Responsive-Adaptive Survey Designs in order to minimize the expected number of CAPI interviews and allocate them to compensate for under-coverage of some categories of households and individuals. The integration of auxiliary information from administrative data is of greatest importance for the quality of the response models and, therefore, the use of efficient adaptive survey designs.

The paper by Corazziari et al. [2] provides an example of how the integration of the information from two sample surveys can be useful for assessing data quality and adjusting for mode effect. In particular, the focus is on the very sensitive topic of violence against women surveyed by the Italian Women Safety Survey. In dealing with such delicate topics, both the role of the interviewer and the choice of the mode used to collect the information are of paramount importance to guarantee reliable information. In Italy, CATI is the main mode used for this survey. In the last release, the sample was designed to provide also a measure of violence against foreign women and CAPI was used for this target subpopulation. As a consequence, the effect on estimates of the data collection mode is confounded with nationality. To assess whether the chosen mode influences the results in the survey, the Authors use data from another survey (the Citizen's Safety Survey) in which the CATI and CAPI interviews were allocated on respondents regardless of their nationality. In addition, to study and to adjust for the interviewer effect and for the mode effect, multilevel models are applied to the indicators of violence against women in common between the two surveys.

Salvatore et al. [7] propose a general framework for augmenting information by combining traditional and digital trace data sources for the production of smart statistical indicators. In particular, the paper focuses on the case of the construction of a smart composite indicator for measuring Corporate Social Responsibility: it is computed integrating social media textual data (and its metadata) with traditional data sources for business statistics. The paper also outlines and discusses the statistical challenges and errors arising throughout the entire production process, from identification of the units of interest in the digital data source to data collection, pre-processing, analysis, and data augmentation.

Falorsi et al. [4] present the new R Package `mind`, developed to implement small area estimation based on multivariate linear mixed models. Beyond the possibility of considering multivariate dependent variables, the model (and its R implementation) allows marginal random effects to be considered, in addition to the usual area random effects. Adding marginal effects in the classical unit-level mixed model may help reduce the bias of small area estimates when the areas of interest are very small and/or out of sample. The proposed set of tools has been developed to obtain estimates from the Italian Permanent Population and Housing Census, but can be usefully applied every time small area estimation methods need to be applied to estimate multiple contingency tables and estimates need to be computed for more than one type of unplanned domains. The paper discusses also the benefits of the proposed marginal effects through a simulation study on the 2011 Census data.

Marcis et al. [5] face the issue of the effect of the heterogeneity of sampling variances in area-level small area estimation, and in the Fay–Herriot model in particular, on likelihood-based final estimates in terms of bias and variance. The problem outlined by the Authors is more worrisome considering that in practice these variances are unknown and estimated (with considerable error) from survey data. Simulation studies results are presented to evaluate in several empirical scenarios the consequences of the heterogeneity of the sampling variances on the linear predictor, by different shapes of their empirical distribution. A discussion on the pros and cons of using the generalized variance function approach to estimate sampling variances is also proposed.

We would like to thank all the Authors who have contributed to this second selection of papers from ITACOSM2022. Warm thanks go also to the Reviewers of the articles, who provided the authors with valuable comments and suggestions. Special thanks go again to Marco Alfò, Editor-in-Chief of Metron, who has invited us to edit these Special Issues and has supported us in this work.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bernardini, A., Chieppa, A., Zindato, D., Cibella, N.: Improving the design of the italian permanent population and housing census: a transition towards a massive use of administrative data. Metron (2023). https://doi.org/10.1007/s40300-023-00256-1
2. Corazziari, I., Ascari, G., Muratore, M.G.: If the tools to gather information affect data quality: violence against women survey case. Metron (2024). https://doi.org/10.1007/s40300-024-00266-7
3. De Vitiis, C., Falorsi, S., Guandalini, A., Inglese, F., Righi, P., Terribili, M.D.: Adaptive sampling design for the Italian social sample surveys: an application on the population census. Metron (2024). https://doi.org/10.1007/s40300-023-00262-3
4. Falorsi, S., D'Alò, M., Fasulo, A.: Mind, a methodology for multivariate small area estimation with multiple random effects. Metron (2024). https://doi.org/10.1007/s40300-023-00258-z
5. Marcis, L., Pagliarella, M.C., Salvatore, R.: How the sampling variances affect the linear predictor of the Fay–Herriot model. Metron (2024)
6. Ranalli, M.G., Beaumont, J.-F., Bertarelli, G., Shlomo, N.: Foreword to the special issue on "survey methods for statistical data integration and new data sources". Metron **81**, 1–3 (2023)
7. Salvatore, C., Biffignandi, S., Bianchi, A.: Augmenting business statistics information by combining traditional data with textual data: a composite indicator approach. Metron (2024). https://doi.org/10.1007/s40300-023-00261-4

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