Foreword to the special issue on “Advances in Survey Statistics”

Jean D. Opsomer¹ · M. Giovanna Ranalli² ·
Maria Michela Dickson³

Received: 2 October 2017 / Accepted: 7 October 2017
© Sapienza Università di Roma 2017

This Special Issue was inspired by the desire to bring together valuable and modern contributions in the field of Survey Statistics, which shed light on some of the challenging tasks of designing and analyzing surveys. Traditionally, survey statistics has been seen as a separate branch to the rest of statistics, because of its reliance on the design-based approach to inference, as opposed to the model-based approach used in other applications of statistics. The former approach is certainly sensible for survey data, because it is common for design features such as strata, clusters and unequal probabilities to induce strong biasing effects in the data. While these effects can be incorporated in models, an often simpler and more robust approach is to rely on randomization theory and inverse-probability weighting to produce estimates and perform inference.

However, there is also a need to recognize the many interactions between randomization-based inference and statistical modeling. The role of modeling in regression estimation is well understood and has been incorporated into the design-based framework for several decades. At the other extreme, small area estimation methods are fully model-based, with the design effects incorporated through the model or, when appropriate, ignored. But there are many statistical research topics that occur at the intersection of randomization-based and model-based inference, as illustrated by many of the contributions to this Special Issue. These include non-response and measurement errors in survey data, estimation under non-measurable designs, and fitting of statistical models under design informativeness. The contributions of this Special Issue deal with topics in either design ([1, 7]) or estimation ([2–6]), also including approaches

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

Jean D. Opsomer
jopsomer@stat.colostate.edu

Maria Michela Dickson
mariamichela.dickson@unitn.it

¹ Department of Statistics, Colorado State University, Fort Collins, US

² Department of Political Science, University of Perugia, Perugia, Italy

³ Department of Economics and Management, University of Trento, Trento, Italy

that target informativeness ([2, 6, 7]) and nonsampling errors such as nonresponse ([4, 6]). All contributions account for the complexity of the sampling design and make use of models. By presenting articles on these disparate survey-related topics in one Special Issue, it is our hope to spur further interest in this exciting research area within statistics.

When a population has a structure that can be described through a model, that information can be used to improve the effectiveness of sampling designs. The most common applications of this idea are in spatial sampling, but other structures such as networks also fall under this category. In the spatial setting, Benedetti et al. [1] review, compare and combine several methods to select spatially balanced samples. These methods use the geographical coordinates of the units in the frame as auxiliary information and produce samples that are well spread over the population. The comparison is based on an extensive simulation study: the performance of several spatially balanced designs is extremely good under a variety of population characteristics and the gain in efficiency depends on the strength of the dependence among closer units. In addition, significant differences are found in the performance of the different methods evaluated and guidelines are provided to choose one, according to the features of the population.

Espa et al. [3] deal with spatial sampling as well, but address an estimation topic. In particular, they look at variance estimation for the estimator of the mean from a systematic sample. This design is non-measurable, so that no unbiased design-based variance estimator is available. However, a model-based approach can be used to explicitly take the spatial structure of the population into account, and obtain a valid variance estimator, jointly with respect to the design and the spatial model. Here, a flexible class of spatial auto-regressive models is considered for the population units, and their applicability in spatial surveys is illustrated in simulations.

Zhang and Patone [7] look at sampling when units exhibit a network structure. In particular, they provide a systematic review of the theory for graph sampling and propose a definition of sample graph together with the relevant observation procedures that enable sampling in a graph. In addition, a general approach to Horvitz-Thompson estimation under arbitrary T -stage snowball sampling is developed. Finally, they provide a graph sampling reformulation of several unconventional sampling methods that are widely used in practice when classical list sampling tools like stratification and clustering are not effective, like multiplicity sampling, indirect sampling, and adaptive cluster sampling.

A number of the contributions in this Special Issue deal with approaches to account for design informativeness as part of a statistical model. The goal of the analysis is to fit a model and perform inference for it, but the data are obtained through a survey. Bonnry et al. [2] deal with estimation of distribution functions for continuous random variables. In particular, they consider the problem of kernel density estimation of the probability density function (pdf) of a response variable under informative selection from a finite population. Properties of the standard kernel density estimator are described under an asymptotic framework that covers a wide range of informative selection mechanisms, and accounts for both the target (superpopulation) model and the model induced by the sampling design. A variety of parametric and non-parametric adjustments to account for the informative selection are proposed, and investigated via simulation.

Modeling under design informativeness is also considered in Lawson and Skinner [6], in the linear regression context. In the setting considered here, the nonresponse occurs at the level of the individual respondents, but the regression model is specified at the cluster level. This allows for specification of response mechanism models that can be incorporated as part of the overall regression model. Several response models and estimators are proposed, and

their properties are studied in two simulation studies. The methods are illustrated in real data from a survey of employees, where the clusters consist of workplaces.

Nonresponse is addressed by Haziza and Chen [4] as well. Differently from [6], item non-response is considered and treated by means of multiple imputation. The focus is on a survey variable that exhibits a large number of zero-valued observations. A main concern is robustness to model failure: this is achieved by using a procedure that is based on multiple imputation models and/or multiple nonresponse models. In this way, the proposed estimator is consistent when at least one model is not misspecified.

Kim et al. [5] deal with the very relevant issue of combining information from several surveys from the same target population. In particular, they propose a measurement error model approach to integrate estimates of the mean obtained from two independent surveys. The situation where two surveys have a common measurement for auxiliary variables, but different measurements for the variable(s) of interest is considered. The survey-specific difference can occur due to differences in survey questions or survey modes. It is assumed that measurement error occurs in one survey, and the other survey is considered as the gold standard. The imputed values for the variable(s) of interest obtained from the measurement error model are combined with direct observations from the gold standard sample to compute a composite estimator. The paper is motivated by data from the Food and Nutrition Technical Assistance Project of the U.S. Agency for International Development and the proposed procedure is applied to produce integrated estimates from two independent surveys conducted in Guatemala, where the geographic areas covered by the two surveys have substantial overlap.

We would like to thank all the Authors who have contributed to this Special Issue: their availability, commitment, and kindness made it possible to realize this great selection of pieces of research. Note that the papers by [1, 3], and [7] were presented at the Workshop on “Advances in Sampling Methods” that was held at the University of Trento on February, 15th, 2017. Warm thanks go also to the reviewers of the articles, who provided the authors with valuable comments and suggestions. Special thanks go to Giovanni Maria Giorgi, Editor-in-Chief of *Metron*, who has invited us to edit this Special Issue and to Marco Alfò, Executive Editor of *Metron*, who has supported us in this work. Last, we thank the whole Editorial Board and the Staff of the journal for giving us the chance of putting together this Special Issue.

References

1. Benedetti, R., Piersimoni, F., Postiglione, P.: Alternative and complementary approaches to spatially balanced samples. *Metron* (2017). doi:[10.1007/s40300-017-0123-1](https://doi.org/10.1007/s40300-017-0123-1)
2. Bonnéry, D., Breidt, F.J., Coquet, F.: Kernel estimation for a superpopulation probability density function under informative selection. *Metron* (2017). doi:[10.1007/s40300-017-0127-x](https://doi.org/10.1007/s40300-017-0127-x)
3. Espa, G., Giuliani, D., Santi, F., Taufer, E.: Model-based variance estimation in two-dimensional systematic sampling. *Metron* (2017). doi:[10.1007/s40300-017-0125-z](https://doi.org/10.1007/s40300-017-0125-z)
4. Haziza, D., Chen, S.: Multiply robust imputation procedures for zero-inflated distributions in surveys. *Metron* (2017). doi:[10.1007/s40300-017-0128-9](https://doi.org/10.1007/s40300-017-0128-9)
5. Kim, J.-K., Park, S., Stukel, D.: A measurement error model approach to survey data integration: combining information from two surveys. *Metron* (2017). doi:[10.1007/s40300-017-0124-0](https://doi.org/10.1007/s40300-017-0124-0)
6. Lawson, N., Skinner, C.: Estimation of a cluster-level regression model under nonresponse within clusters. *Metron* (2017). doi:[10.1007/s40300-017-0120-4](https://doi.org/10.1007/s40300-017-0120-4)
7. Zhang, L.-C., Patone, M.: Graph sampling. *Metron* (2017). doi:[10.1007/s40300-017-0126-y](https://doi.org/10.1007/s40300-017-0126-y)