STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. III–VI

Preface

The demand for statistics on a range of socio-economic, agricultural, health, transportation, and other topics is steadily increasing at a time when government agencies are desperately looking for ways to reduce costs to meet fixed budgetary requirements. A single data source may not be able to provide all the data required for estimating the statistics needed for many applications in survey and official statistics. However, information complied through different data linkage or integration techniques may be a good option for addressing a specific research question or for multi-purpose uses. For example, information from multiple data sources can be extracted for producing statistics of desired precision at a granular level, for a multivariate analysis when a single data source does not contain all variables of interest, for reducing different kinds of nonsampling errors in probability samples or self-selection biases in nonprobability samples, and other emerging problems.

The greater accessibility of administrative and Big Data and advances in technology are now providing new opportunities for researchers to solve a wide range of problems that would not be possible using a single data source. However, these databases are often unstructured and are available in disparate forms, making data linkages quite challenging. Moreover, new issues of statistical disclosure avoidance arise naturally when combining data from various sources. There is, therefore, a growing need to develop innovative statistical data integration tools to link such complex multiple data sets. In the US federal statistical system, the need to innovate has been emphasized in the following report: National Academies of Sciences, Engineering, and Medicine. (2017), *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24652.

The idea of organizing an international week-long workshop on statistical data integration arose in 2017. I joined Dr. Sanjay Chaudhuri, a faculty member at the National University of Singapore (NUS), Dr. Danny Pfeffermann, National Statistician of Israel, and Dr. Pedro Silva of the Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil, and former President of the International Statistical Institute, to organize this international workshop. Eventually, with generous funding from the Institute for Mathematical Sciences at the National University of Singapore, the workshop was held on the NUS campus during August 5–8, 2019. The World Statistics Congress Satellite meeting on Current Trends in Survey Statistics took place at the same venue in the following week, August 13–16, 2019. We had great success with

participants and speakers from more than 18 countries in these two meetings, at which a number of papers on statistical data integration were presented.

A few months before the two Singapore events, in February of 2019, I had a fruitful lunch meeting in the Washington DC area with Professor Wlodzimierz Okrasa, Editor-in-Chief, and Dr. Graham Kalton, a member of the Editorial Board, of the Statistics in Transition (SiT) New Series. During that meeting they invited me to edit a special issue for the journal. We discussed a few options for the focus of the special issue. Our discussions led to the idea of focusing on statistical data integration, in view of the current importance of the topic, and the value of disseminating the findings from current research. We felt the issue would be timely, given the emphasize on this topic in the two Singapore workshops that were to be held later that year. We agreed that anyone, including the participants of the two Singapore meetings, could submit papers for possible publication in the special issue, and all papers would go through a thorough review process.

Out of the nineteen papers submitted for possible publication in this special issue, we finally accepted ten papers, after they went through a referring and revision process. In addition, this special issue features an invited discussion paper on a selective review of small area estimation by Professor Malay Ghosh, which is based on his 2019 Morris Hansen lecture delivered in Washington DC on October 30, 2019. We are pleased to have seven experts, including Professor J. N. K. Rao and Dr. Julie Gershunskaya – the two invited discussants of Professor Ghosh's Morris Hansen lecture – as discussants of Professor Ghosh's paper.

For over 75 years, survey statisticians have been using information from multiple data sources in solving a wide range of problems. One early example of combining surveys can be traced back to a 1943 Sankhya paper (www.jstor.org/stable/25047787) by Mrs. Chameli Bose. Mrs Bose developed the regression estimation for double sampling used by Professor P.C. Mahalanobis in 1940–41 to estimate the yield of cinchona bark in the Government Cinchona Plantation at Mungpoo, Bengal, India. Over the years, we have witnessed tremendous progress in such research topics as small area estimation, probabilistic record linkage, combining multiple surveys, multiple frame estimation, microsimulation, poststratification, all of which incorporate multiple data sources and can be brought under the broader umbrella of statistical data integration or data linkages. In a 2020 Sankhya B paper (doi 10.1007/s13571-020-00227-w), Professor J. N. K. Rao provides an excellent review of a selected subtopics of statistical data integration.

It is difficult to cover all interesting statistical data integration topics in a single issue of SiT. But we are happy that the invited discussion review paper plus the ten contributed papers published in this special issue collectively cover a broad spectrum of topics in statistical data integration. The papers can be broadly classified into the following subtopics: 1) small area estimation, 2) advances in probabilistic record

linkage and analysis of linked data, 3) statistical methods for longitudinal data, multiple-frame, and data fusion, and 4) synthetic data for microsimulations, disclosure avoidance and multi-purpose inferences.

Professor Ghosh's paper, along with the discussions, provide an excellent review of some topics in small area estimation and they should prove to be a valuable reference for those working on small area estimation. In addition, this issue features two more papers on small area estimation by (i) Cai, Rao, Dumitrescu, and Chatrchi, and (ii) Neves, Silva, and Moura that address variable selection and modeling to capture uncertainties of sampling errors of survey estimates, respectively. These are indeed important and yet understudied problems in small area estimation.

This special issue includes two papers that advance knowledge on probabilistic record linkage. Consiglio and Tuoto investigate potential advantages of using probabilistic record linkage in small area estimation. Bera and Chatterjee discuss a problem of probabilistic record linkage on high-dimensional data. This is a novel approach to the probabilistic record linkage methodology that can be applied in absence of any common matching field among the data sets.

The three papers by (i) Saegusa, (ii) Zhang, Pyne, and Kedem, and (iii) Bonnery, Cheng, and Lahiri investigate potential benefits of using nonparametric and semi-parametric methods to combine information from multiple data sources. The nature of the available multiple data sources differs between the three papers. Saegusa develops a nonparametric method to construct confidence bands for a distribution function using multiple overlapping data sources – this is an advancement in the multiple-frame theory. To overcome a relatively small sample of interest, Zhang et al. propose a semi-parametric data fusion technique for combining multiple spatial data sources using variable tilts functions obtained by model selection. Bonnery et al. carefully devise a complex simulation study, using the U.S. Current Population Survey (CPS) rotating panel survey data, to evaluate different possible estimators of levels and changes in the context of labor force estimation.

The three papers by (i) Bugard, Dieckmann, Krause, Münnich, Neufang, and Schmaus, (ii) Alam, Dostie, Drechsler, and Vilhuber, and (iii) Lahiri, and Suntornchost demonstrate how the synthetic data approach can be useful for solving seemingly unrelated problems. Bugard et al. discuss microsimulations that are used for evidence-based policy. Using a general framework for official statistics, they use synthetic data created from multiple data sets to approximate a realistic universe. The synthetic data discussed in the Alam et al. paper relates to statistical data disclosure. The authors consider a feasibility study to understand if the synthesis method for longitudinal business data used in a US project can be effectively applied to two other longitudinal business projects, in Canada and Germany. In the context of poverty estimation for small geographic areas, Lahiri and Suntornchost point out the inappropriateness of using point estimates for all inferential purposes. Using a Bayesian approach,

they demonstrate how synthetic data can be created for multipurpose inferences in small area estimation problems.

I would like to thank Professor Wlodzimierz Okrasa and Dr. Graham Kalton for encouraging me to take a lead on this project. I appreciate all the help I received from Professor Okrasa and his editorial staff. Thanks are also due to the anonymous referees who offered many constructive suggestions to improve the quality of the original submissions. Last but not the least, I would like to thank my distinguished guest co-editors Drs. Jean-Francois Beaumont, Sanjay Chaudhuri, Jörg Drechsler, Michael Larsen, and Marcin Szymkowiak for their diligent editorial work. Without their enormous help, we would not have this high quality special issue.

Partha Lahiri,
Guest Editor-in-Chief