The Survey Statistician, 2022, Vol. 85, 30-32.

Book and Software Review

Big Data Meets Survey Science. A Collection of Innovative Methods

Alina Matei

Institute of Statistics, University of Neuchâtel, Switzerland. E-mail : alina.matei@unine.ch

Wiley published in 2021 the book entitled Big Data Meets Survey Science. A Collection of Innovative Methods, edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg. The book includes selected papers presented at the first conference on Big Data Meets Survey Science (BigSurv18), hosted by the University Pompeu Fabra's Research and Expertise Centre for Survey Methodology in Barcelona, Spain in 2018, and conducted under the auspices of the European Survey Research Association.

Featuring a broad range of topics, the book includes 24 chapters organized in six sections, and offers a wide range of tools, methods, and approaches that illustrate how Big Data sources and methods are used in survey and social sciences to improve official statistics and estimates. Finding answers to the two following questions is essential for survey statisticians/official statisticians:

1) How are Big Data defined?
2) What does '*survey science'* mean?

First, the editors provide in the book introduction the following definition: 'In contrast to censuses or surveys that generate designed and sample data, we view Big Data as nonsampled data that are organic or found in sources for which the primary purpose is not statistical inference *per se*. In particular, in this book we use the term of Big Data to refer to a collection of datasets so large, complex, and rapidly changing that they become difficult to process using extant database management tools or traditional data processing applications.' Second, it seems that a formal definition of 'survey science' is not given in the book. One understands that it represents a mixture between survey methodology and data science.

The information provided in the book is huge, reason for which I focus my discussion on some parts related to official statistics and survey estimation, mostly presented in Section 3 ('Big Data in Official Statistics'). Compared to the definition advocated above, census and administrative data are also seen as Big Data in Section 3, when referring to large populations. It is difficult to define what is 'large'. Chapter 11 (Tam et al., 2021), for instance, provides an application in Subsection 11.6, where the population size is 1,000,000. Does this represent Big Data? Numerous existing surveys around have been dealing with such or even larger amount of data.

Following Holt (2007), Chapter 9 (Japec and Lyberg, 2021) advocates and illustrates several possible issues for official statistics using Big Data: 'wider, deeper, quicker, better, and cheaper'. The aspect 'cheaper' is illustrated, for example, by the estimation of the Consumer Price Index using scanner data from retail stores, first used by Statistics Sweden, and employed now by several countries. Chapter 9 also discusses the very important aspect of the Big Data quality. 'Selectivity, in that Big Data subpopulations often do not coincide with target populations studied in official statistics' represents one of the most important challenges for National Statistical Offices.

The future will tell us if 'wider, deeper, quicker, better, and cheaper' will represent characteristics of Big Data statistics. Definitely, Big Data provide nowadays a 'cheaper' solution in some cases. Nevertheless, this may not always be the case in the future. Big Data statistics are quicker to obtain, yet they may not be wider or deeper if they do not represent, for instance, the whole population under study, or if a carefully designed study is not in place. Last, but not least, the aspect 'better' will depend upon all the other mentioned characteristics.

Chapter 10 (Braaksma et al., 2021) offers a very good description on how Big Data are used at Statistics Netherlands, and how they may be used in general in official statistics. It furthermore describes innovative experiments useful to develop other methods to deal with Big Data. This chapter also presents two strategies to handle Big Data in official statistics: 1) use Big Data as an indicator of the society, assuming however some imperfections due to the way these data are collected, and 2) use Big Data into a model or model-assisted approach and conduct similar analyses as for classical survey data. The authors underline the high degree of changes in Big Data, and a lack of information about the data-generating process. They also advocate the link between multiple sources, and provide the following example among others: individual social media information may be combined to some information known from registers for instance. Nevertheless, many users keep personal information private, to protect their privacy on social media. In this context, it is quite difficult to successfully use such a data combination.

Chapter 11 (Tam et al., 2021) offers a more technical content. It is dedicated to data combination between a nonprobability sample or Big Data set and a probability sample. The method is based on calibration, assuming that different totals are known. Several limitations of the method are underlined by the authors. The method is, however, promising, and opens the way for new methodological developments.

It is important to mention the use of machine learning in clustering and prediction models. Chapter 1 (Buskirk and Kirchner, 2021) provides an informative review on the use of machine learning methods (MLMs) in surveys. The use of algorithms is not new in survey statistics. Methods based on clustering algorithms are used for a while, for example, to create imputation classes; see for instance Haziza and Beaumont (2007). 'Compared to traditional statistical methods, MLMs are more prone to overfitting the data, that is, to detecting patterns that might not generalize to other data', underline the authors of Chapter 1. Overfitting is not, however, an MLM issue, and traditional statistical criteria and methods may lead to overfitted models (especially when cross-validation techniques are not used during the model selection process). Recently, several papers in survey estimation have used model-assisted estimators, that attenuate the potential impact of overfitting produced by MLMs (McConville et al., 2017; Mehdi Dagdoug et al., 2022).

I welcome the attempt of the cited chapters' authors to also underline the drawbacks of Big Data sources and methods. However, my impression in reading the book is that a gap between survey statisticians and Big Data defenders is still present. More research must be done from the methodological point of view to accommodate Big Data in our 'routine' as survey statisticians. In any case, Big Data should be used together with conventional statistical sources and methods whenever they can bring new insights.

In conclusion, the book includes a large diversity of topics, making it informative for a broad audience including survey and social science researchers, and survey statisticians/official statisticians. Without any doubt, the book represents an important contribution to survey science. The Big Data debate still continues, but I hope that the book will help to diminish the mentioned gap between survey statisticians and Big Data defenders.

References

Braaksma, B., Zeelenberg, K., and de Broe, S. (2021). Big Data in Official Statistics: A Perspective from Statistics Netherlands. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 303-338. Wiley.

Buskirk, T. D. and Kirchner, A. (2021). Why Machines Matter for Survey and Social Science Researchers: Exploring Applications of Machine Learning Methods for Design, Data Collection,

and Analysis. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 9-62. Wiley.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. International Statistical Review, 75(1):25-43.

Holt, D. T. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. *The American Statistician*, 61(1):1-8.

Japec, L. and Lyberg, L. (2021). Big Data Initiatives in Official Statistics. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 275-302. Wiley.

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-Assisted Survey Regression Estimation with the Lasso. *Journal of Survey Statistics and Methodology*, 5(2):131-158.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, to appear.

Tam, S. M., Kim, J.-K., Ang, L., and Pham, H. (2021). Mining the New Oil for Official Statistics. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 339-357. Wiley.