The Survey Statistician, 2025, Vol. 91, 44–46.

Book and Software Review

Advances in Business Statistics, Methods and Data Collection Section 7 "Data Integration, Linking and Matching"

María Bugallo[1]

[1]Center of Operations Research, Miguel Hernández University of Elche, Spain.
mbugallo@umh.es

Abstract

The book "*Advances in Business Statistics, Methods and Data Collection*" represents a comprehensive and up-to-date contribution to the field. It covers a wide range of currently available methods, recent developments, and challenges in the production of modern, high-quality business statistics. The book is divided into seven self-contained sections, each clearly stating and pursuing its objectives. All the topics addressed are essential for enhancing decision-making and efficiency within national statistical offices and private institutions. By improving data collection processes and employing advanced statistical techniques, statisticians can gain deeper insights and achieve greater accuracy of results. A good example of the applicability of these techniques is found in Section 7, entitled "*Data Integration, Linking and Matching*", which is the focus of this review. It is divided into four chapters, the first two being more conceptual, discussing linked data and methods for estimating the quality of multisource statistics, and the last two dealing with case studies.

Keywords: Record linkage, matching data, multisource statistics, data integration, data quality.

856 pages of the book "*Advances in Business Statistics, Methods and Data Collection*" contain comprehensive review papers and relevant case studies, with an eye to the future, for increasingly global economies and larger volumes of data. Modern societies are changing faster and faster and the scientific community must respond appropriately to these new dynamics. The authors expertly guide the reader through a variety of real survey data examples, complemented by diagrams that help visualise the database structure, an aspect that can often pose significant challenges in practice. In this report we look at Section 7, which focuses on "*Data Integration, Linking and Matching*" and is organized into four self-contained chapters, from Chapter 33 to Chapter 36.

Chapter 33 (Larsen & Herning, 2023) deals with record linkage so we must first define what we are talking about. In this context, the purpose of record linkage is to identify distinct entities across two or more databases by establishing associations based on the analysis of variables from the different files. Even in cases where unique and accurate identification numbers are lacking, strong evidence could be provided to support the conclusion that the representations refer to the same entity. There

is no denying that gathering information from different surveys, censuses or external files potentially enriches statistical analyses. It is unique identification numbers that are the most effective variables for linking individuals and organizations across different databases. However, they are often not available and more sophisticated methods must be provided. In this chapter, exact, deterministic and probabilistic matching methods are covered and illustrated with an application to business data linking using Australian data sources. One-to-one matching and the computational part of these algorithms are also discussed. Both the subject matter results as well as the variety of examples are of high interest to the public.

As previously mentioned, Chapter 33 includes a practical example applied to the Business Longitudinal Analysis Data Environment. The methods are widely used in different fields of statistics. Here, I would like to emphasize the importance of record linkage for Small Area Estimation (SAE). In an SAE context, survey data is collected at a predetermined level of aggregation; however, the accuracy of the estimates for specific domains cannot be guaranteed. Consequently, statistical models are developed to incorporate auxiliary information (covariates) along with spatial and temporal dependency structures (Morales et al., 2021). Predictors derived from unit-level models, which use individual survey responses as the target variables, require supported census files to be calculated. It is for this reason that data linking can be highly beneficial.

While linked data offers numerous advantages, the potential impact of errors arising from record linkage should not be underestimated. This leads us to Chapter 34 (van Delden et al., 2023), which focuses on methods for estimating the quality of multisource statistics. Clearly, this chapter is about an important challenge. Multisource statistics refers to the collection and analysis of data from different sources or databases. One form of multisource statistics is to produce integrated microdata sets. There are several reasons why it is important to quantify the output accuracy of multisource statistics: first, to inform users about the quality of the results, and second, to enhance the accuracy of regularly published statistics. This chapter focuses on methods for calculating the quality statistics of the results.

Recently developed methods to estimate bias and variance of outputs affected by representation error, linkage error and measurement error are introduced in Chapter 34. Of particular interest is Section 34.4.2, which deals with estimating the effect of measurement error on the results. However, this presents a dual challenge, as its mathematical rigor may overwhelm less experienced readers. In addition, all methods are illustrated with applications to real data and valuable explanations. Simulation studies are also included. The techniques are of advanced complexity and require theoretical knowledge and computational skills to develop and implement. The authors refer to de Waal et al. (2021) for a comprehensive overview of the types of methods currently in use.

Although all of the procedures in the previous two chapters are accompanied by real-world examples, Chapters 35 and 36 are particularly guided toward the application to real data.

Chapter 35 (Young et al., 2023) addresses the integration of previously reported data into the 2022 Census of Agriculture. It highlights the challenges faced by the US Department of Agriculture's National Agricultural Statistics Service (NASS), which conducts over 100 surveys each year and is experiencing increasing nonresponse rates due to the burden placed on large producers by frequent survey requests. The Census of Agriculture has also seen declining response rates. This chapter outlines NASS's plan to incorporate previously reported data in the 2022 Census of Agriculture and describes the tests being conducted to assess its impact on data quality and the response rates. It begins with a description of the Agricultural Production Survey and its sources of previously reported data, followed by an overview of the study's design and results.

Chapter 36 (Duval et al., 2023) explores the integration of alternative and administrative data into

monthly business statistics, highlighting Statistics Canada's modernization efforts to address a complex economy and growing user demands for timely information. This strategy has effectively enhanced data quality, relevance, and cost efficiency while reducing the burden on respondents across various monthly business programs. The chapter explores three applications of alternative and administrative data in three monthly economic programs. The first replaces survey data with Goods and Services Tax data in the Monthly Survey of Food Services and Drinking Places. The second uses scanner data in the Retail Commodity Survey, while the third application involves using SAE techniques that combine survey data with administrative data in the Monthly Survey of Manufacturing. Each project includes context, methodology, assessment, and results, along with a brief discussion on the pandemic's impact and future initiatives. Together with Chapter 35, this chapter offers clear examples of how to tackle practical case studies with rigor and expertise.

As a final remark, the book as a whole, and Section 7 as an example, begins with detailed introductory descriptions of the topics to be covered, ensuring that it can be understood by a wide range of readers, from statisticians to practitioners and researchers. With its rigorous approach, it aims to equip them with the tools necessary for deriving actionable insights from complex datasets. My small reservation is the lack of mathematical developments in some sections, as it seems that the book mainly targets and emphasizes the intuition behind the application of the various statistical techniques rather than the mathematical aspects.

References

Advances in Business Statistics, Methods and Data Collection (2023). Snijkers, G., Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J. W., Thompson, K. J. & van Delden A. (eds.), Wiley. https://doi.org/10.1002/9781119672333.

Duval, M. C., Laroche, R. & Landry, S. (2023). Integrating Alternative and Administrative Data into the Monthly Business Statistics: Some Applications from Statistics Canada. In *Advances in Business Statistics, Methods and Data Collection* (pp. 821–838). Wiley.

Larsen, M. D. & Herning, A. (2023). Record Linkage for Establishments: Background, Challenges, and an Example. In *Advances in Business Statistics, Methods and Data Collection* (pp. 757–780). Wiley.

Morales, D., Esteban, M. D., Pérez, A. & Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models Methods: Theory and Applications in R*. Springer.

van Delden, A., Scholtus, S., de Waal, T. & Csorba, I. (2023). Methods for Estimating the Quality of Multisource Statistics. In *Advances in Business Statistics, Methods and Data Collection* (pp. 781–804). Wiley.

de Waal, T., van Delden, A. & Scholtus, S. (2021). Commonly used methods for measuring outputquality of multisource statistics. *Spanish Journal of Statistics*, **2**, 79–107. https://doi.org/10.37830/SJS.2020.1.05.

Young, L. J., Rodhouse, J. B., Terner, Z., & Corral, G. (2023). Adopting Previously Reported Data into the 2022 Census of Agriculture: Lessons Learned from the 2020 September Agricultural Survey. In *Advances in Business Statistics, Methods and Data Collection* (pp. 805–820). Wiley.