DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 141–142, http://dx.doi.org/10.1515/JOS-2015-0008

Book Review

Dean M. Resnick[1]

Anders Wallgren and Britt Wallgren. (Eds.) *Register-based Statistics*. 2014 New York: Wiley, ISBN 978-1-119-94213-9, 320 pp, $120.

"Register-based Statistics" by Anders and Britt Wallgren is a how-to cookbook for creating a national statistical register from scratch. The type of register envisioned is one along a Nordic model that continuously tracks a set of entities such as persons, households, or businesses by the compilation and updating of existing data from administrative sources. Created in this manner, this kind of register would allow the development of consistent, policy-relevant statistics on an ongoing basis or as new research questions arise without having to field a new survey, add new questions to an existing survey, or requiring the recompilation and reintegration of administrative record data from multiple sources. Based on the authors' experience of developing registers like this for Sweden, the authors provide tools, recommendations, caveats, and the rudiments of an administrative data system theory (which they correctly suggest is presently much less developed than sampling or survey theory).

For an American reviewer, this book presents something of a conundrum. This is because, at least in terms of person, family, or household-specific data, the development of a statistical register is not countenanced legally or socially, particularly under the coordination of a government entity. To some degree, this concern is obviated by the book's coverage of non-person-based registers (as of businesses), but more generally, the book takes on more relevance (for an American reader) if considered more as a guide to the use of administrative record data as combined from multiple sources, including survey data.

Here, this book provides a useful overview of the technical issues encountered in this type of processing (i.e., combining data from multiple sources). However, in this regard, this book should be considered more as an introductory presentation rather than a thorough explication of the more advanced data-management and statistical techniques needed for this. For example, the book discusses issues related to record linkage, imputation, entity duplication, and undercoverage, but in regard to these topics, a reasonably experienced data analyst or statistician would probably be seeking a much fuller treatment. Thus, it seems, this book is best suited for someone fairly new to the field, such as a manager or a policymaker. Here, the book lays out some very useful guiding principles, such as the need for subject-matter expertise, comprehensive metadata, and carefully thought-out data integration approaches.

[1] Health Policy Center, The Urban Institute. 2100 M Street NW Washington, DC 20037, U.S.A.
Email: DResnick@urban.org

Certainly, there are some areas that would be more valuable to a more experienced analyst. Particularly appreciated is the extensive treatment of multilevel variables. By this, the book means a categorical data item for which a given entity (e.g., a business, person or household) can be fairly considered as having more than one value, at least over the course of time. Here the authors rightly indicate the dangers associated with the selection and representation of only one of these values, such as the biasing of derived estimates, and provide thorough guidance on how multiple-level data can be retained and used for estimation. The recommended treatment of these data seems quite extendable to imputation results (although it is not clear this is intended by the authors).

In addition, this book provides a nice treatment of the integration of administrative and survey data, suggesting that some entities (i.e., businesses or households) may be represented on one of these sources and not another and therefore their concatenation allows a fuller picture of the represented situation than either alone. This would be advice well heeded for someone working to develop comprehensive statistical estimates from available data sources.

In terms of the treatment of error within administrative data, this book certainly provides good guidance on how to minimize these, but it is rather rudimentary in presenting a theoretical framework for quantifying them – suggesting the appropriate measurement techniques are not well developed. Here, the statistical comparison of data elements from different sources (i.e., comparing administrative data to survey data) seems a useful area for exploration and a natural extension of the treatment of data integration. Still, it is greatly appreciated that the authors stress that sampling error is only a small part of estimation errors (albeit readily treatable by known statistical techniques). If quality comparisons are made between survey and administrative data, the existence of nonsampling error in survey data should be recognized.

In terms of readers for whom this book would be most helpful, obviously, someone newly assigned to the task of creating a statistical register would be the greatest beneficiary. To some degree, persons with experience in this area would also benefit from the identification and systematization of methods relevant for this type of work. For those not involved in register-development *per se*, but seeking to develop competence in the integration and use of administrative record data, this book could be a useful introduction to and reference for applicable methods and their systematization and a source of best-practice principles.