Journal of Statistical Software

June 2017, Volume 78, Book Review 2. *doi: 10.18637/jss.v078.b02*

Reviewer: Stefano M. Iacus
University of Milan

Big Data and Social Science – A Practical Guide to Methods and Tools

Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, Julia Lane
Chapman & Hall/CRC, Boca Raton, 2016.
ISBN 978-1-498-75140-7. 356 pp. GBP 25.95 (H).
http://www.bigdatasocialscience.com/

Social science practitioners and researchers face more and more large and complex datasets. There is no way to turn around the "big data" issue, big data are now a commodity in the social sciences. Still, the social data scientist is not yet a common profile in academia and research institutions. The reasons for this uncompleted transition, from social scientist to social data scientist, is the lack of access to proper statistical methods and computer science tools. This book tries to fill the gap between the willingness to perform a big data analysis in the social sciences and the actual competence of doing it. The authors did a great effort in this direction and they succeed to some extent. As data science is a mix of skills and background knowledge from different fields, it is clearly impossible to fill all the gaps. Therefore, this book, at times, must remain on the surface.

The book is divided into three parts. The first one, "Model and Curation", explains how to deal with simple web scraping and then describes the more complex use of API, although it doesn't really explain the details of some basic, yet necessary, steps and problems faced by practitioners, like authentication, tokens, etc. Apart from that, the Python code presented is sufficient to understand the API concept. Some practical examples of interactions with ORCID and Twitter API's are also explained. The ORCID to Twitter case study also introduces the other problem of big data which is the record linkage issue, i.e., how to put together data from different sources? This chapter describes applications of survey methodology to the big data context, like probabilistic record linkage and other useful techniques. Then, a standard chapter follows about data base systems (where do I store the data once I got them?). Finally, the authors present the distributed computing paradigm to solve elementary but scalable tasks on huge amounts of data. To present, this is one of the non-specialist books which treats the topic with sufficient detail. I think this is a plus.

The second part of the book, called "Modelling and Analysis", goes through the standard machine learning topics but avoids to talk about deep learning, which is quite trendy these days and available through several open source frameworks. The next two chapters contain a non exhaustive review of some text analysis techniques. This part is probably too elementary, and recent approaches like Word2Vec or aggregated sentiment analysis approaches are completely missing, although quite popular in the social sciences. This second part of the book ends with the basic ideas of social network analysis.

The third part of the book is dedicated to "Inference and Ethics" but actually starts with effective data visualization, another fundamental topic. Inference is considered in terms of errors more than in terms of statistical models (although some are presented). This chapter clearly addresses the problem of how and where errors arise in big data analysis. This is an often underestimated problem in social data science. Some solutions to these problems are also, very briefly, presented. The last chapter discusses Privacy and Confidentiality. This is a topic usually neglected in technical books but a real concern for the social data scientist and especially if he or she works in an governmental authority or public institution. This problem is two-fold: on the one hand, it is pretty legal (how to keep the privacy of the subject under investigation provided that we can mix many sources of data); on the other hand, this issue is about the distribution of data for replicability which is more and more common in the social sciences.

The book also contains a "Workbook" chapter, which is a collection of **Jupyter** notebooks that explain how to replicate most of the examples presented in the book by skipping the burden of learning everything from scratch but allowing the practitioner to work through these examples at first and only later to dig deeply in the code.

In summary, although there is a growing number of books related to social science and big data, this volume contains several non-trivial aspects which make it worth to have in the library, possibly along with other similar textbooks as a good complement to them. Not all subjects are treated in full detail, but when this is the case, most of the time, the overview offered is valuable and the reader can always examine other specialized texts later on.

Reviewer:

Stefano M. Iacus
Department of Economics, Management and Quantitative Methods
University of Milan
Via Conservatorio 7, I-20123 Milan, Italy
E-mail: stefano.iacus@unimi.it

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
June 2017, Volume 78, Book Review 2 *Published:* 2017-06-01
doi:10.18637/jss.v078.b02