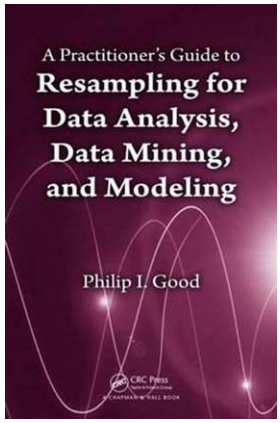
Journal of Ambient Intelligence and Smart Environments 4 (2012) 391–392
DOI 10.3233/AIS-2012-0161
IOS Press

Book review



A practitioner's guide to resampling for data analysis, data mining, and modeling: *A cookbook for starters*

Egon L. van den Broek *
Media and Network Services, TNO Technical Sciences, Delft, The Netherlands
URL: http://www.human-centeredcomputing.com/
E-mail: vandenbroek@acm.org

Abstract. A practitioner's guide to resampling for data analysis, data mining, and modeling provides a gentle and pragmatic introduction in the proposed topics. Its supporting Web site was offline and, hence, its potentially added value could not be verified. The book refrains from using advanced mathematics and as such is useful for undergraduate courses for a broad range of sciences. Moreover, it is a suitable book for some first hands on experience with data mining. However, it is neither suitable as a resource for advanced issues nor has it much added value compared to its predecessors.

Keywords: Practitioner's guide, data analysis, data mining, modeling, course book

A practitioner's guide to resampling for data analysis, data mining, and modeling by P. Good, Boca Raton, FL, USA: CRC Press / Taylor & Francis Group, LCC, 2012, ISBN: 978-1-439855-50-8.

1. Introduction

Ambient Intelligence (AmI) requires abundant storage of various data streams, which vary highly in virtually all possible aspects. However, although advanced processing schemes have been apposed and appealing applications have been posed (e.g., digital preservation of multimedia [5]), "classical" data mining remains to serve as its foundation [4]. So, data mining is and should stay within the focus area of AmI.

Having sufficient books on data analysis, data mining, and modeling and even more than enough books on statistics in general on my shelf, I was intrigued by by the launch of a new book on data mining, which prominently denoted "practitioner's guide" in its title. It suggests to be different from my other books. Moreover, perhaps it could serve as a handbook for undergraduate lectures and as an introductory book. Therefore, I have chosen to assess the value of this book from the perspective of its use in practice.

2. The book's Web site

The preface advertises with a Web site, which should provide "code for most resampling methods" and "code for many of the routines." However, half

*Additional affiliations: Human Media Interaction (HMI), Faculty of EEMCS, University of Twente, Enschede, The Netherlands and Karakter University Center, Radboud University Medical Center (UMC) Nijmegen, Nijmegen, The Netherlands.

way 2012, almost one year after the book was released, this Web site is not online (see URL: http://statcourse.com/PGsoftware.htm). However, the main Web site, http://statcourse.com/, also mentioned in the preface, is online and provides various courses.

Regrettably, for all of the courses provided via the Web site, a considerable fee has to be paid. Not a single line of code is available. Moreover, when a "practitioner's guide" is introduced, sufficient material should be available to "learn by doing" and, as such, counter the deficiencies in the mathematical foundation. Some exercises are available; however, the answers are missing. In addition, the reader is directed to http://www.statcrunch.com/ for some data sets. This Web site is not related to the book; why not provide data sets dedicated to the book?

Perhaps, the book simply aims to attract public to the author's website. He has a consultancy company, which provides commercial statistics courses. This book seems to serve as a very nice advertisement for the author's company and can be used to illustrate the level of his courses, which I will discuss next.

3. Statistics for lay persons

All this being said, I think that Good is a gifted writer. He presents statistics in a way that makes it easy for lay people to grasp the basic ideas behind it. Moreover, he provides sufficient examples and, where possible, presents them in a context that immediately illustrates the relevance of the technique at hand. Good presents the foundation of statistics, simplifies where possible, but does not lose himself in oversimplifying too much. On the one hand, this makes the book in particular suitable for undergraduate courses and for those who want to obtain quick hands on experience with data analysis, data mining, and modeling. On the other hand, for those seeking advanced knowledge on data analysis, data mining, and modeling the book would be disappointing and they are advised to search for another resource; for example, [3,6].

For those interested, Good provides an adequate index and a wealth of references that provide pointers for readers who want to learn more on a specific issue. This is a feature that would also be appreciated by those who are knowledgable in the field. This having said, the book also suffers from minor shortcommings. For example, the typesetting of the formulas is not consistent throughout the book. This will make it unnecessarily difficult for practitioners that are not comfortable with mathematics to grasp the ideas behind it. Moreover, the typesetting and images are of low quality, which one would not expect from a rather expensive hardcover.

4. Conclusion

Overall, Good relies on his vast experience and presents yet another introductory book on statistics, data analysis, data mining, and modeling in particular. It provides the gentle introduction, as is claimed, which can indeed be used in a variety of sciences. Its foundation is good, but in parallel it feels a little outdated; thus, its added value is questionable.

For those who want to save their money, I suggest another book [2], by the same author, as an interesting alternative. That book shows a significant overlap (to say the least) with the book reviewed. There are also other alternatives by the same author [1]. With those books, the publisher provides both an instructor's manual and data sets for the exercises, which would be of value when used as course material.

Taken together, this "practitioner's guide" has given me an ambivalent feeling. On the one hand, it is OK for lay persons in the field and for undergraduate courses. On the other hand, it refrains from going into much details; so, as a source for more advanced issues it is not appropriate. Moreover, older and much cheaper but as valuable books by the same author are available as well. This book feels like a re-release, which not have been necessary.

References

[1] P.I. Good, *Introduction to Statistics Through Sampling Methods and Microsoft Office Excel or R/S-PLUS*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005.

[2] P.I. Good, *Resampling Methods: A Practical Guide to Data Analysis,* 3rd edn., Birkhäuser Boston, New York, NY, USA, 2006.

[3] R. Nisbet, J. Elder, and G. Miner, *Handbook of Statistical Analysis & Data Mining Applications*, Academic Press, 2009.

[4] J. Ponce and A. Karahoca, *Data Mining and Knowledge Discovery in Real Life Applications*, In-Teh/I-Tech Education and Publishing KG, Vienna, Austria, 2009.

[5] E.L. van den Broek, F. van der Sluis, and Th.E. Schouten, User-centered digital preservation of multimedia, *ERCIM (European Research Consortium for Informatics and Mathematics) News* **80** (2010), 45–47.

[6] I.H. Witten, E. Frank, and M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn., Morgan Kaufmann, 2011.