Journal of Statistical Software

September 2004, Volume 11, Book Review 9. http://www.jstatsoft.org/

Reviewer: Nicholas J. Cox
University of Durham

Exploratory Data Mining and Data Cleaning

Tamrapani Dasu and Theodore Johnson
John Wiley, Hoboken, NJ, 2003.
ISBN 0-471-26851-8. xii + 203 pp. $69.95.

Data mining has been an area looming just beyond statistical science for several years, and even an area that some statisticians evidently regard as overlapping with their territory. Yet many people may be unclear on how it differs from statistics applied to large or very large datasets, together to be sure with a lot of data management. Data cleaning is certainly something people in statistical science should care about, and perhaps care more about than we do, even if we mostly make a living out of analysing other people's datasets. Sometimes it seems that statisticians have largely ignored it. J.I. Naus' book *Data quality control and editing* (Marcel Dekker, New York, 1975) is one counter-example, which appears to have had very little impact. On the other hand, everybody I know has war stories about cleaning up lousy datasets, which usually boil down to the need for very careful scrutiny while applying common sense or subject-matter knowledge (women 90 years old cannot have babies, or whatever). Such methods do not lend themselves to systematic exposition. It is much easier to write another text on linear models.

A book that introduces these two areas is thus an attractive proposition, particularly when there are hints of an exploratory approach, reminiscent of John W. Tukey's exploratory data analysis, and the whole comes in at moderate length. In this case the standard issue of who the book is for and what prior knowledge is assumed is particularly crucial. A book with such a title in the Wiley Series in Probability and Statistics might be intended for at least three groups of readers. First come statistically-minded people who may know less about computing with large datasets than they would like, especially if that involves databases with many string or text fields, rather than datasets with mostly or entirely numeric variables. Perhaps many of the readers of this review, and certainly its author, fall into this group. Second come the opposite group, computing people who know less statistics than they would like. Third might come managers who need to know more, possibly because they have groups of technical people working for them.

Dasu and Johnson's book is structured to proceed from exploratory data mining through to data quality. Statistical material is densest in the early part of the book and database concepts densest in the later part. The authors appear to have experience in interacting with all these

groups of readers. Nevertheless, they seem to aim at varying targets throughout the book, and all too commonly their exposition is an uneven mishmash. Some very elementary statistical concepts are introduced at length, while several more advanced or more esoteric concepts are covered briefly and cryptically. Moreover, the explanations and the recommendations for further reading often seem pitched at the wrong level. If you didn't know what a mean is, you get an explanation on p.27, first with an integral and an (unexplained) expectation operator, and then with the usual formula for a sample mean. Who wants this kind of explanation? Theoretical physicists migrating into statistics? Similarly, medians, modes, histograms and chi-square tests are introduced as if new to the reader (while boxplots go unexplained on p.153). The last chapter takes it that you might appreciate a definition of a relational database, but do not need explanations of regular expressions, hash functions or the $O()$ notation applied to algorithms.

At another extreme, there are several discursive sections with flow charts, bulleted lists and lots of keywords in bold face which seem derived from presentations to management. These sound very sensible but also, to be blunt, are often elaborations of the obvious. Take the reminder that problems ensue if missing values are encoded by a number that is a possible data value. Everybody has to learn this for the first time, but I guess that most readers of this review learned it indirectly when they came across people to whom it was not self-evident.

Some other problems arise from a lack of care in presentation. Examples of unnecessary repetition and minor disorganisation indicate that the book has been pasted together from a variety of documents in the authors' files (witness a tell-tale 'In this paper' on p.95). The section 'Why are data dirty?' (pp.165–166) revisits questions covered earlier. Copy editing and proof reading have also been skimped. The graph on p.10 is repeated on p.38; the definitions of Poisson distributions and R^2 are incorrect on pp.53 and 94; the histogram on p.60 is in fact a bar chart. Spelling errors like Hausdorf and Serpinski seem expectable compared with a mass of punctuation errors and other small typos. Careless statements vary from the surreal ('We will not discuss regression methods in this book', p.91, in the middle of a section on regression), through the merely awkward (unreported or dropped data, unintended duplicate records and switched fields are described as 'mega phenomena' on p.104), to the backward ('Clear and accurate information about the data is the biggest DQ [data quality] problem in practice', p.118).

In addition, the examples are weak. Problems of confidentiality presumably explain why good examples cannot be reproduced directly from the authors' experiences with real large datasets, but instead we often get toy examples based on fabricated data. One concerns snarks, gryphons, and unicorns: even devotees of Lewis Carroll may find this wearing thin after several encounters. The authors use themselves too as examples, with various little jokes about New York and New Jersey, which may not translate far from their origin. A more realistic case study is based on the insensitively-named Holy Cow Corp. (pp.180–187).

On the whole, software solutions are referred to only briefly and obliquely. SAS seems to get most mentions among proprietary programs. I sense an associated presumption that most readers will be in business or government rather than academic or research institutions. Perl and CGI scripts are also mentioned. I think it is fair to say that readers will learn little in any detail about appropriate software solutions or computing styles from this book, even though many might expect this to be largely what the book is all about.

There are, to be sure, several sections on bigger issues: data depth (how deeply embedded a data point is in the data cloud); the importance of one-pass algorithms with very large

datasets; the role of approximate matching. Anyone wanting to proceed further has a Bibliography of 130 references. A large fraction of the data base literature recommendations appear to be in fugitive conference proceedings.

I think the authors are right: there is a major need for books in this intersection of statistics and computing. Unfortunately, their own contribution tries too hard to appeal to all potential groups of readers and is too carelessly presented to receive my recommendation.

Reviewer:

Nicholas J. Cox
Department of Geography
University of Durham, UK
E-mail: n.j.cox@durham.ac.uk

Journal of Statistical Software *Published:* 2004-09-27
September 2004, Volume 11, Book Review 9.
http://www.jstatsoft.org/