Journal of Statistical Software

December 2006, Volume 17, Book Review 1. http://www.jstatsoft.org/

Reviewer: Paul Murrell
The University of Auckland

Graphics of Large Datasets: Visualizing a Million

Antony Unwin, Martin Theus, and Heike Hoffman
Springer-Verlag, New York, 2006.
ISBN 0-387-32906-4. xiv + 271 pp. USD 84.95 (H).
http://stats.math.uni-augsburg.de/GOLD/

In the preface to "Graphics of Large Datasets" the authors set out two goals for the book:

> We hope that our book will help others interested in visualizing large datasets to find out more easily what has been done and to contribute themselves. More especially, we hope it will help data analysts in analysing their data.

The book definitely succeeds on the first point, providing a report on the state-of-the art in advanced statistical graphics. However, it has some weaknesses in making that information available to practitioners.

The overall organisation of the book is in increasing order of specificity and technical detail. Chapter 1 provides a general introduction to the ideas of visualizing data and what constitutes a large data set. This chapter includes some interesting historical perspectives, makes some useful general points like the fact that "large" can mean many variables as well as many cases, and discusses the ways that data size can impact on the process of data analysis. There is also an acknowedgement that, despite rapid advances in hardware and software, data sets are growing larger at a greater rate than our capability to deal with them. As the authors put it (page 18):

> Large will usually mean large rather than LARGE in this book.

Chapter 2 provides a taxonomy of basic types of plots. This includes a description of very standard plot types such as barplots, histograms, and scatterplots, but also includes a number of plots that are not as widely known, such as mosaic plots, spline plots, fluctuation diagrams, and structure plots. There are also specific sections on maps, time series plots, and representations of missing data. The authors' background and expertise in statistical graphics, particularly interactive graphics, is reflected in the variety of plots and in the examples used. However, for some of the more modern plots there is perhaps too little detail given for a non-expert reader to understand how the plots work.

There is one painful error (for this reader) in the caption of Figure 2.22; Maunga Whau is in Auckland, New Zealand, not Hawaii.

Chapter 3 looks at how the plot types from the previous chapter can be applied to visualizing large data sets. Where problems arise, such as excessive overplotting of symbols in scatterplots, various solutions are discussed, such as using alpha-blending (semitransparent plotting symbols). This is a very interesting chapter, with useful discussions of ideas such as "red marking", (logical) zooming, density estimation and binning, and ordering or sorting elements of a plot.

Chapter 4 looks specifically at the importance of interactive graphical techniques in plots of large data. The initial sections describe the basic interactive techniques of querying, selecting and linking between plots and the later sections look at how the techniques may be used in a large data setting. Again, the chapter contains many interesting and useful ideas, most of which are not as widely used in practice as they deserve to be. Unfortunately, this is where interested readers may become slightly frustrated with a lack of information about how to try it for themselves.

This concludes the first part of the book. In Part II, various guest authors have a chapter each to discuss issues of visualizing large data for a specific type of data, or type of plot: Heike Hoffman discusses categorical data and several variations on the barplot and mosaic plot; Dianne Cook and Leslie Miller look at scaling up the **GGobi** software system for massive data sets; Rida Moustafa and Ed Wegman describe several variations on the parallel coordinate plot; Graham Wills looks at the problems of visualizing very large (node-and-edge) graphs; Simon Urbanek describes a number of novel plots for visualizing large tree models; and Bárbara González-Arévalo, Félix Hernández-Campas, Steve Marron, and Cheolwoo Park discuss plots of internet traffic data (mice and elephant plots).

Each of these chapters is quite specific and may be of interest to a smaller audience. Some provide theoretical background material that will only really be of interest to other researchers in the field.

Chapter 11 stands alone as an extended case study of a particular data set. This chapter looks at a data set of High-Tech company information, which was used for the InfoVis 2005 conference contest, and demonstrates some of the techniques described earlier in the book being used in anger. This chapter is a good read and helps to make a number of ideas more obvious and concrete.

Unsurprisingly, for a book on graphics, there are many figures and they are generally of a very high quality. This implies that considerable time and care was spent on preparing the figures, given that the software used to produce most of the plots is designed for exploratory and interactive graphics, rather than presentation graphics. Colour is used extensively, often simply to highlight a selected group with the colour red, but again generally to good effect.

Overall, the book is more research report than how-to guide. Even the early, general chapters contain some information that is more speculation than concrete advice (a comment on one proposed solution reads: "now all that is needed is to design and implement the method", page 95). Some of the later chapters are very focused on specific applications and represent a report on a research project rather than a description of techniques or tools that have been developed for a general audience.

Anyone interested in modern techniques for visualizing data will be well rewarded by reading this book. There is a wealth of important plotting types and techniques. Data analysts who

read this book will be exposed to many new and valuable techniques, but will have to look elsewhere for information on how perform these techniques themselves.

Reviewer:

Paul Murrell
Department of Statistics
The University of Auckland
Auckland, New Zealand
E-mail: paul@stat.auckland.ac.nz
URL: http://www.stat.auckland.ac.nz/~paul/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 17, Book Review 1 *Published:* 2006-12-05
December 2006