Journal of Statistical Software

May 2009, Volume 30, Book Review 7. *http://www.jstatsoft.org/*

Reviewer: Pedro Valero-Mora
University of Valencia

Interactive and Dynamic Graphics for Data Analysis: With R and GGobi

Diane Cook and Deborah F. Swayne
Springer-Verlag, New York, 2007.
ISBN 978-0-387-71762-3. 190 pp. USD 59.95 (P).
http://www.GGobi.org/book/

Interactive and dynamic data visualization is about analysts using the computer to actively explore the data at hand. Indeed, readers interested in this subject should feel very happy lately because three books focusing on such techniques—namely, Cook and Swayne (2007), Unwin, Theus, and Hofmann (2006), and Young, Valero-Mora, and Friendly (2006)—have recently found their way into the press. This must be regarded as a very notable circumstance as there had not been published any book offering a comprehensive view of this topic since Cleveland and McGill (1988) or the manuals of **DataDesk** (Velleman 1995). Besides, even though the three books have many elements in common, there are also sufficient singularities in them to satisfy readers with different interests. Cook and Swayne's book in particular has the special appeal of being linked to **GGobi/rggobi** software that implements very nicely many of the tools discussed in the text.

Writing about interactive and dynamic graphics can be very frustrating because it involves turning into static something whose basic nature is, well, dynamic. Think about reading the manual instead of playing the videogame—not much fun in comparison. In this case, however, Cook and Swayne have managed to produce a book that, for most of the time, can be enjoyed by itself, making it unnecessary to switch on the computer at every moment to understand what is being described in it. This is a notable achievement, attained throughout a combination of good practices: clear prose, compelling figures, compelling figures set in the sequence of steps of an analysis (e.g., the one on page 109), short introductions to theory that do not take you away of the main issues, R code for computing the intermediate steps of some analysis, and did I mention compelling figures?

However, although readers may enjoy the book by itself without being forced to resort constantly to software for understanding, this does not mean that the stuff available electronically is disappointing at all. On the contrary, **GGobi/rggobi** is fine software that provides for an unbeatable price a number of specialized analyses that are very hard to find elsewhere. Additionally, the web site for the software includes complementary material such as movies illustrating the analysis, R code to run examples, and documents. The book has also exercises at the end of each chapter, and solutions to them can be obtained from the publisher and used in a course on statistics.

The book has six chapters. Chapters 1 and 2 are introductory, the first one focusing on the philosophy of data analysis using graphics, and the second one on describing several tools available for this purpose. The book includes a large number of examples and datasets that illustrate the methods. Thus, Chapter 1 uses the Tips dataset to exemplify how using interactive graphics improves the insight that might be obtained using what they call the old-fashioned approach—to fit a regression model. The chapter also provides an overview of the different stages of the data analysis process. They explain that accounts of analysis in textbooks or papers tend to be very straightforward, as if analysts would be able to carry a step after the other because the sequence is logical and necessary, but, on the contrary, the real process of data analysis usually involves searching among several choices, often having to discard many of them before arriving to a satisfactory end. Putting this convoluted process into words is not an easy task but they manage to convey a general idea of it.

Chapter 2 is a listing of the tools usually available in software programs. Many of these tools, but not all, are available in **GGobi**, the freely available software developed by the authors that runs in Linux, Windows and Mac OS X and that is the successor of previous programs that can be traced back to the mid-1980s. The current version can be used with R (via the **rggobi** package) so data manipulations, analysis and visualizations can all be carried out together. Besides, the chapters usually include R code needed for massaging the data before they are actually ready to be visualized. At `http://www.GGobi.org/` there is a short document (4 pages) that introduces **rggobi**. Statistical packages specialized in interactive/dynamic graphics are generally concentrated on taking as much as they can of the most common types of plots and, in this tradition, the chapter only discusses histograms, bar charts, scatter plots, parallel coordinates plots, scatter plot matrices, mosaic plots, and tour plots. Of these, the description of tour plots (Asimov 1985) is exceptionally well done.

Chapter 3 deals with visualization of data with missing values. **GGobi** does not offer specialized plots on visualization of missing values but the authors show how to encode the missing values with convenient values to explore them using scatter plots, tour plots or parallel coordinate plots. However, as this strategy is limited by the appropriateness of the values used for encoding, the authors also use the imputation of the missing values via the R package **norm** (Schafer and Novo 2002). This provides more acceptable values for the visualization.

Supervised classification is the topic of Chapter 4. The emphasis here is in examining in detail the output that the computational algorithms provide. Namely, the idea is to check if the classes are well separated in the data space, if there are clear boundaries between classes, and if there are cases that are suspicious of being misclassified. This is the largest chapter of the book and consequently the topics are described more thoroughly than in other chapters. The main example used is about features of olive oils cultivated in different regions of Italy. Several algorithms for supervised classification are applied to this data, but the most important part is about using interactive graphics to check the assumptions and to explore the results.

Chapter 5 is devoted to cluster analysis, or unsupervised classification. Again, an approach based mainly on graphics but with some support from numerical techniques, is used. Specially interesting is the figure at page 109, which uses a sequence of graphics that resembles McCloud's definition of a comic (McCloud 1994), in order to convey the interactive process of exploring the data and identifying the elements of importance in it. The chapter's main example is PRIM-7, a venerable dataset first used by Friedman and Tukey (1974), but other examples are also used. Graphics are utilized to compare the results of different methods applied, and to characterize the clusters, i.e., to detect the numerical or qualitative differences between the clusters.

Finally, Chapter 6 includes several methods that can be applied to other types of data that could not have been analyzed well using the methods in the previous chapters. The treatment of these topics is rather brief, but the authors promise that additional chapters will be given out for free at the web site of the book. Additionally, the chapter has sections on network data and multidimensional scaling.

References

Asimov D (1985). "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing*, **6**(1), 128–143.

Cleveland WC, McGill ME (1988). *Dynamic Graphics for Statistics*. CRC Press, Inc., Boca Raton, FL, USA.

Cook D, Swayne DF (2007). *Interactive and Dynamic Graphics for Data Analysis*. Springer-Verlag, New York. ISBN 978-0-387-71761-6.

Friedman JH, Tukey JW (1974). "A Projection Pursuit Algorithm for Exploratory Data Analysis." *IEEE Transactions on Computers*, **23**(9), 881–890. ISSN 0018-9340.

McCloud S (1994). *Understanding Comics*. Perennial Currents.

Schafer JL, Novo AA (2002). "**norm**: Analysis of Multivariate Normal Datasets with Missing Values." R package version 1.0-9, URL `http://CRAN.R-project.org/package=norm`.

Unwin A, Theus M, Hofmann H (2006). *Graphics of Large Datasets: Visualizing a Million*. Springer-Verlag, Secaucus, NJ, USA.

Velleman PF (1995). ***DataDesk*** *Version 5*. DataDescription Inc., Ithaca, New York.

Young FW, Valero-Mora PM, Friendly M (2006). *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*. John Wiley & Sons.

Reviewer:

Pedro Valero-Mora
University of Valencia
Department of Methodology of Behavioral Sciences
Valencia, Spain 46010
E-mail: valerop@uv.es
URL: http://www.uv.es/valerop/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 30, Book Review 7 *Published:* 2009-05-11
May 2009