Antony Unwin, Chun-houh Chen, Wolfgang Härdle

Introduction to the Handbook of Computational Statistics and Data Visualization

Preprint Nr. 028/2007 — 11. Juli 2007

Institut für Mathematik, Universitätsstraße, D-86 135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:

Institut für Mathematik
Universität Augsburg
86135 Augsburg
http://www.math.uni-augsburg.de/forschung/preprint/

ViSdP:

Antony Unwin
Institut für Mathematik
Universität Augsburg
86135 Augsburg

Introduction

Antony Unwin[1], Chun-houh Chen[2] and Wolfgang Härdle[3]

[1] Augsburg University, Germany unwin@math.uni-augsburg.de
[2] Institute of Statistical Science, Academia Sinica, Taiwan
cchen@stat.sinica.edu.tw
[3] Humboldt-Universität zu Berlin, Germany haerdle@wiwi.hu-berlin.de

1 Computational Statistics and Data Visualization

This book is the third volume of the Handbook of Computational Statistics and covers the field of Data Visualization. In line with the companion volumes, it contains a collection of chapters by experts in the field to present readers with an up-to-date and comprehensive overview of the state of the art. Data Visualization is an active area of application and research and this is a good time to gather together a summary of current knowledge.

Graphic displays are often very effective at communicating information. They are also very often not effective at communicating information. Two important reasons for this state of affairs are that graphics can be produced with a few clicks of the mouse without any thought, and that the design of graphics is not taken seriously in many scientific textbooks. Some people seem to think that preparing good graphics is just a matter of common sense (in which case their common sense cannot be in good shape) and others believe that preparing graphics is a low-level task, not appropriate for scientific attention. This volume of the Handbook of Computational Statistics takes graphics for Data Visualization seriously.

1.1 Data Visualization and Theory

Graphics provide an excellent approach for exploring data and are essential for presenting results. Although graphics have been used extensively in statistics for a long time, there is not a substantive body of theory about the topic. Quite a lot of attention has been paid to graphics for presentation, particularly since the superb books of Edward Tufte. However, this knowledge is expressed in principles to be followed and not in formal theories. Bertin's work from the 1960s is often cited, but has not been developed further. This is a curious state of affairs. Graphics are used a great deal in many different fields and one might expect more progress to have been made along theoretical lines.

Sometimes in science the theoretical literature for a subject is considerable while there is little applied literature to be found. The literature on Data Visualization is very much the opposite. Examples abound in almost every issue of every scientific journal concerned with quantitative analysis. There are occasionally articles published in a more theoretical vein about specific graphical forms, but little else. Although there is a respected statistics journal called the Journal of Computational and Graphical Statistics, most of the papers submitted there are in computational statistics. Perhaps this is because it is easier to publish a study of a technical computational problem than it is to publish work on improving a graphic display.

1.2 Presentation and Exploratory Graphics

The differences between graphics for presentation and graphics for exploration lie in both form and practice. Presentation graphics are generally static and a single graphic is drawn to summarise the information to be presented. These displays should be of high quality and include complete definitions and explanations of the variables shown and of the form of the graphic. Presentation graphics are like proofs of mathematical theorems, they may give no hint as to how a result was reached, but they should offer convincing support for its conclusion. Exploratory graphics, on the other hand, are used for looking for results. Very many of them may be used and they should be fast and informative rather than slow and precise. They are not intended for presentation, so that detailed legends and captions are unnecessary. One presentation graphic will be drawn for viewing by potentially thousands of readers while thousands of exploratory graphics may be drawn to support the data investigations of one analyst.

Books on visualization should make use of graphics. Figure 1 shows some simple summaries of data about the chapters in this volume, revealing that over half the chapters had more than one author and that more authors does not always mean longer papers.

1.3 Graphics and Computing

Developments in computing power have been of great benefit to graphics in recent years. It has become possible to draw precise, complex displays with great ease and to print them with impressive quality at high resolution. That was not always the case and initially computers were more a disadvantage for graphics. Computing screens and printers could at best produce clumsy line-driven displays of low resolution without colour. These offered no competition to careful, hand-drawn displays. Furthermore, even early computers made many calculations much easier than before and allowed fitting of more complicated models. This directed attention away from graphics and it is only in the last twenty years that graphics have come into their own again.

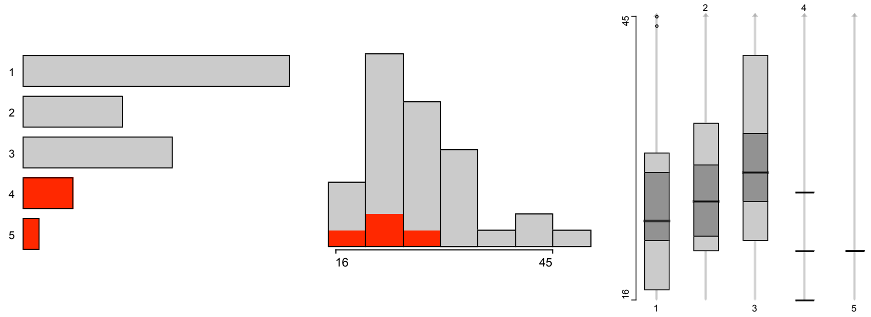

Fig. 1. A barchart of the number of authors per paper, a histogram of the number of pages per paper, and parallel boxplots of length by number of authors. Papers with more than three authors have been selected.

These comments relate to presentation graphics, that is graphics drawn for the purpose of illustrating and explaining results. Computing advances have benefitted exploratory graphics, that is graphics drawn to support exploring data, far more. Not just the quality of graphic representation has improved but also the quantity. It is now trivial to draw many different displays of the same data or to riffle through many different versions interactively to look for information in data. These capabilities are only gradually becoming appreciated and capitalized on.

The importance of software availability and popularity in determining what analyses are carried out and how they are presented will be an interesting research topic for future historians of science. In the business world no one seems to be able to do without the spreadsheet Excel. If Excel does not offer a particular graphic form, then this will not be used. (In fact Excel offers many graphic forms, though not all that a statistician would want.) Many scientists, who only rarely need access to computational power, also rely on Excel and its options. In the world of statistics itself, the packages SAS and SPSS were long dominant. In the last fifteen years first S and S-plus and now R have emerged as important competitors. None of these packages currently provide effective interactive tools for exploratory graphics, though they are all moving slowly in that direction as well as extending the range and flexibility of the presentation graphics they offer.

Data Visualization is a new term. This expresses the idea that it involves more than just representing data in a graphical form (instead of using a table). The information behind the data should also be revealed in a good display, the graphic should aid readers or viewers in seeing the structure in the data. The term Data Visualization is related to the new field of Information Visualization. This includes visualization of all kinds of information, not just of data, and is closely associated with research by computer scientists. Up till now the work in this area has tended to concentrate just on presenting information, rather than on what may be deduced from it. Statisticians tend to be concerned more with variability and to emphasise the statistical properties of results. The closer linking of graphics with statistical modelling can make this more explicit and is a promising research direction that is facilitated by the flexible nature of current computing software. Statisticians have an important role to play here.

2 The Chapters

Needless to say, each Handbook chapter uses a lot of graphic displays. Figure 2 is a scatterplot of the number of figures against the number of pages. There is an approximate linear relationship with a couple of papers having rather more figures per page and one rather less. The scales have been chosen to maximise the data-ink ratio. An alternative version with equal scales makes clearer that the number of figures per page is almost always less than one.

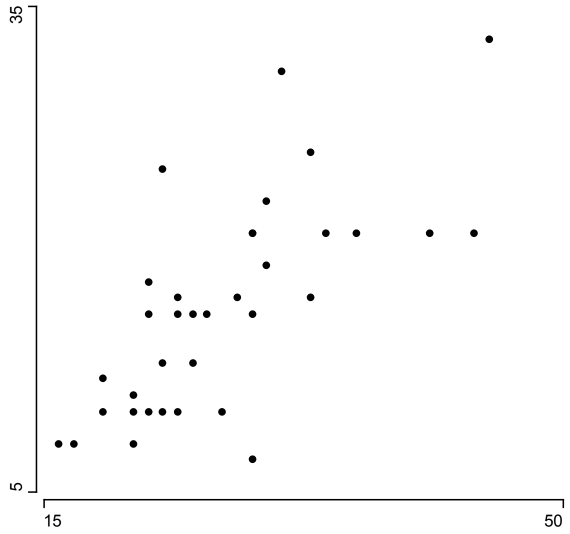

Fig. 2. A scatterplot of the number of figures against the number of pages for the Handbook's chapters.

The Handbook has been divided into three sections: Principles, Methodology, and Applications. Needless to say, the sections overlap. Figure 3 is a binary matrix visualization using Jaccard coefficients for both chapters (rows) and index entries (columns) to explore links between chapters. In the raw data map (lower-left portion of Figure 3) there is a banding of black dots from the lower-left to upper-right corners indicating a possible transition of chapter/index combinations. In the proximity map of indices (upper portion of Figure 3) group B indices are dominated by the History chapter and overlap with group C indices dominated by the Good Graphics chapter. Both chapters deal with general concepts of graphics.

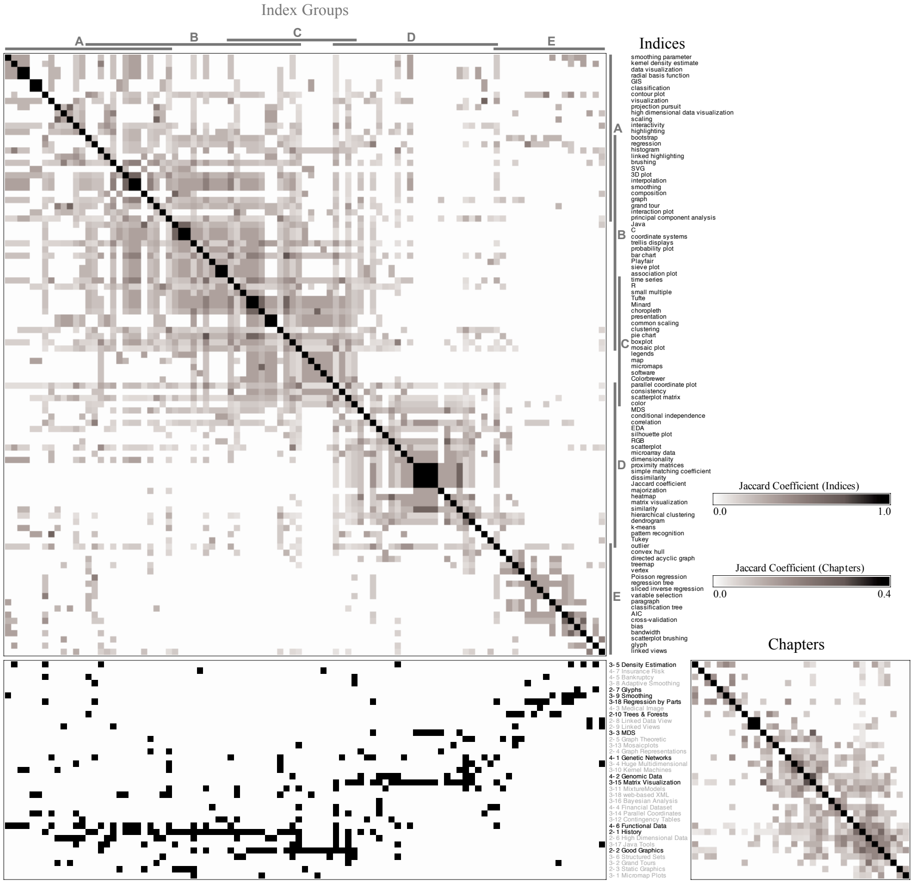

Fig. 3. Matrix visualizations of the Handbook with chapters in the rows and index entries in the columns.

2.1 Summary and Overview; Part II

The ten chapters in Part II are concerned with principles of Data Visualization. First there is an historical overview by Michael Friendly, the custodian of the internet Gallery of Data Visualization, outlining the developments in graphical displays over the last few hundred years and including many fine examples.

In the next chapter Antony Unwin discusses some of the guidelines for the preparation of sound and attractive data graphics. The question mark in the chapter title sums it up well: whatever principles or recommendations are followed, the success of a graphic is a matter of taste, there are no fixed rules.

The importance of software for producing graphics is incontrovertible. In Paul Murrell's chapter he summarises the requirements for producing accurate and exact static graphics. He emphasises both the need for flexibility in customising standard plots and the need for tools which permit the drawing of new plot types.

Structure in data may be represented by mathematical graphs. George Michailidis pursues this idea in his chapter and shows how this leads to another class of graphic displays associated with multivariate analysis methods.

Lee Wilkinson approaches graph-theoretic visualizations from another point of view and his displays are concerned predominantly, though by no means exclusively, with trees, directed graphs and geometric graphs. He also covers the layout of graphs, a tricky problem for large numbers of vertices, and raises the intriguing issue of graph matching.

Most data displays concentrate on one or two dimensions. This is frequently sufficient to reveal striking information about a dataset. To gain insight into multivariate structure, higher dimensional representations are required. Martin Theus discusses the main statistical graphics of this kind that do not involve dimension reduction and compares their possible range of application.

Everyone knows about Chernoff faces, though not many ever use them. The potential of data glyphs for representing cases in informative and productive ways has not been fully realised. Matt Ward gives an overview of the wide variety of possible forms and of the different ways they can be utilized.

There are two chapters on linking. Adalbert Wilhelm describes a formal model for linked graphics and the conceptual structure underlying it. He is able to encompass different types of linking and different representations. Graham Wills looks at linking in a more applied context and stresses the importance of distinguishing between views of individual cases and aggregated views. He also highlights the variety of selection possibilities there are in interactive graphics. Both chapters point out the value of linking simple data views over linking complicated ones.

The final chapter in this section is by Simon Urbanek. He describes the graphics that have been introduced to support tree models in statistics. The close association between graphics and the models (and collections of models in forests) is particularly interesting and has relevance for building closer links between graphics and models in other fields.

2.2 Summary and Overview; Part III

The middle and largest section of the Handbook concentrates on individual area of graphics research.

Geographical data can obviously benefit from visualization. Much of Bertin's work was directed at this kind of data. Juergen Symanzik and Daniel Carr write about micromaps (multiple small images of the same area displaying different parts of the data) and their interactive extension.

Projection pursuit and the Grand Tour are well-known, but not easy to use. Despite the availability of attractive free software, it is still a difficult task to analyse datasets in depth with this approach. Dianne Cook, Andreas Buja, Eun-Kyung Lee and Hadley Wickham describe the issues involved and outline some of the progress that has been made.

Multidimensional scaling has been around for a long time. Michael Cox and Trevor Cox (no relations, but an MDS would doubtless place them close together) review the current state of research.

Advances in high-throughput techniques in industrial projects, academic studies and biomedical experiments and the increasing power of computers for data collection have inevitably changed the practice of modern data analysis. Real life data sets become larger and larger in both sample size and numbers of variables. Francesco Palumbo, Alain Morineau, and Domenico Vistocco illustrate principles of visualization for such situations.

Some areas of statistics benefit more directly from visualization than others. Density estimation is hard to imagine without visualization. Michael Minnotte, Steve Sain and David Scott examine estimation methods in up to three dimensions. Interestingly there has not been much progress with density estimation in even three dimensions.

Sets of graphs can be particularly useful for revealing the structure in datasets and complement modelling efforts. Richard Heiberger and Burt Holland describe an approach primarily making use of Cartesian products and the Trellis paradigm. Wei-Yin Loh describes the use of visualization to support the use of regression models, in particular with the use of regression trees.

Instead of visualizing the structure of samples or variables in a given dataset, researchers may be interested in visualizing images collected with certain formats. Usually the target images are collected with various types of noise pattern and it is necessary to apply statistical or mathematical modeling to remove or diminish the noise structure before the possible genuine images can be visualized. Jörg Polzehl and Vladimir Spokoiny present one such novel adaptive smoothing procedure in reconstructing noisy images for better visualization.

The continuing increase in computer power has had many different impacts on statistics. Computationally intensive smoothing methods are now commonplace, although they were impossible only a few years ago. Adrian Bowman gives an overview of the relations between smoothing and visualization. Yuan-chin Chang, Yuh-Jye Lee, Hsing-Kuo Pao, Mei-Hsien Lee and

Su-Yun Huang investigate the impact of kernel machine methods on a number of classical techniques: principal components, canonical correlation and cluster analysis. They use visualizations to compare their results with those from the original methods.

Cluster analyses have often been a bit suspect to statisticians. The lack of formal models in the past and the difficulty of judging the succes of the clusterings were both negative factors. Fritz Leisch considers the graphical evaluation of clusterings and some of the possibilities for a sounder methodological approach.

Multivariate categorical data were difficult to visualize in the past. The chapter by David Meyer, Achim Zeileis and Kurt Hornik describes fairly classical approaches for low dimensions and emphasises the link to model-building. Heike Hofmann describes the powerful tools of interactive mosaicplots that have become available in recent years, not least through her own efforts, and discusses how different variations of the plot form can be used for gaining insight into multivariate data features.

Alfred Inselberg, the original proposer of parallel coordinate plots, offers an overview of this approach to multivariate data in his usual distinctive style. Here he considers in particular classification problems and how parallel coordinate views can be adapted and amended to support this kind of analysis.

Most analyses using graphics make use of a standard set of graphical tools, scatterplots, barcharts, histograms, for example. Han-Ming Wu, ShengLi Tzeng and Chun-houh Chen describe a different approach, built around using colour approximations for individual values in a data matrix and applying cluster analyses to order the matrix rows and columns in informative ways.

For many years Bayesians were primarily theoreticians. Thanks to MCMC methods they are now able to also apply their ideas to great effect. This has led to new demands in assessing model fit and the quality of the results. Jouni Kerman, Andrew Gelman, Tian Zheng and Yuejing Ding discuss graphical approaches for tackling these issues in a Bayesian framework.

Without software to draw the displays, graphic analyis is almost impossible nowadays. Junji Nakano, Yamamoto Yoshikazu and Keisuke Honda are working on Java-based software to provide support for new developments and they outline their approach here. Many researchers are interested in providing tools via the web. Yoshiro Yamamoto, Masaya Iizuka and Tomokazu Fujino discuss using XML for interactive statistical graphics and explain the issues involved.

2.3 Summary and Overview; Part IV

In the final section there are seven chapters on specific applications of data visualization. There are, of course, individual applications discussed in earlier chapters, but here the emphasis is on the application rather than principles or methodology.

Genetic networks are obivously a promising area for informative graphic displays. Grace Shieh and Chin-Yuan Guo describe some of the progress made so far and make clear the potential for further research.

Microarray data analysis has from the first relied on graphical displays. Florian Hahne, Wolfgang Huber and Robert Gentleman illustrate a variety of different tools and convey both the importance of visualization in this field and the range of interesting problems to be tackled.

Modern medical imaging systems have made significant contributions to diagnoses and treatments. Henry Lu discusses the visualization of data from positron emission tomography, ultrasound, and magnetic resonance.

Two chapters examine company bankruptcy datasets. In the first one, Antony Unwin, Martin Theus and Wolfgang Härdle use a broad range of visualization tools to carry out an extensive exploratory data analysis. No large dataset can be analysed cold and this chapter shows how effective data visualization can be in assessing data quality and revealing features of the dataset. The other bankruptcy chapter employs graphics to visualize SVM modelling. Wolfgang Härdle, Rouslan Moro and Dorothea Schäfer use graphics to display results that cannot be presented in a closed analytic form.

The astonishing growth of eBay has been one of the big success stories of recent years. Wolfgang Jank, Galit Shmueli, Catherine Plaisant and Ben Shneiderman have studied data from eBay auctions and describe the role graphics played in their analyses.

Krzysztof Burnecki and Rafal Weron consider the application of visualization in insurance. This is another example, where the value of graphics lies in providing insight into the output of complex models.

2.4 The authors

The editors would like to thank the authors of the chapters for their contributions. It is important for a collected work of this kind to cover a broad range and to gather many experts with different interests together. We have been fortunate in receiving the assistance of so many excellent contributors.

The mixture at the end remains, of course, a mixture. Different authors take different approaches and have different styles. It early became apparent that even the term Data Visualization means different things to different people! We hope that the Handbook gains rather loses by this eclecticism.

Figure 1 and Figure 2 earlier in the chapter showed that the chapter form varied between authors in various ways. Figure 4 reveals another aspect. The scatterplot shows an outlier with a very large number of references (the historical survey of Michael Friendly) and that some papers referenced the work of their own authors more than others. The histogram is for the rate of self-referencing.

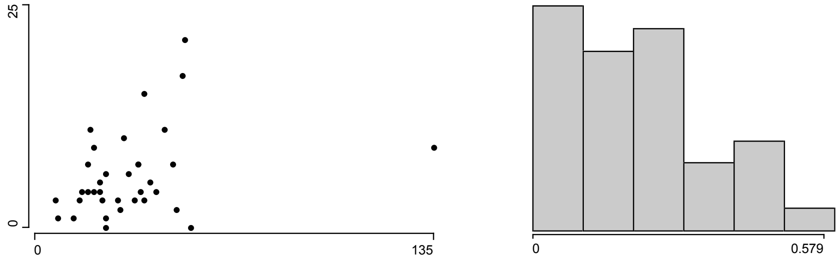

Fig. 4. A scatterplot of the number of references to papers by a chapter's authors against the total number of references and a histogram of the rate of self-referencing.

3 Outlook

There are many open issues in Data Visualization and many challenging research problems. The datasets to be analysed tend to be more complex, and are certainly becoming larger all the time. The potential of graphical tools for exploratory data analysis has not been fully realised and the complementary interplay between statistical modelling and graphics has not yet been fully exploited. Advances in computer software and hardware have made producing graphics easier, but also contributed to raising the standards expected.

Future developments will undoubtedly include more flexible and powerful software and better integration of modelling and graphics. There will probably be individual new and innovative graphics and some improvements in the general design of displays. Gradual gains in knowledge about the perception of graphics and the psychological aspects of visualization will lead to improved effectiveness of graphic displays. Ideally there should be progress in the formal theory of Data Visualization, but that is perhaps the biggest challenge of all.