# Use R!

*Series Editors:*
Robert Gentleman    Kurt Hornik    Giovanni Parmigiani

# Use R!

*Albert:* Bayesian Computation with R
*Cook/Swayne:* Interactive and Dynamic Graphics for Data Analysis
*Paradis:* Analysis of Phylogenetics and Evolution with R
*Pfaff:* Analysis of Integrated and Cointegrated Time Series with R

Dianne Cook    Deborah F. Swayne

# Interactive and Dynamic Graphics for Data Analysis

## With R and GGobi

With Contributions by Andreas Buja, Duncan Temple Lang,
Heike Hofmann, Hadley Wickham, and Michael Lawrence

Springer

Dianne Cook
Department of Statistics
Iowa State University
325 Snedecor Hall
Ames, IA 50011-1210
dicook@iastate.edu

Deborah F. Swayne
AT & T Labs - Research
Shannon Laboratory
180 Park Avenue
Florham Park, NJ 07932-1049
dfs@research.att.com

*Series Editors:*
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N, M2-B876
Seattle, Washington 981029-1024
USA

Kurt Hornik
Department für Statistik und Mathematik
Wirtschaftsuniversität Wien Augasse 2-6
A-1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

ISBN 978-0-387-71761-6 e-ISBN 978-0-387-71762-3

Library of Congress Control Number: 2007925720

© 2007 Springer Science+Business Media, LLC
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

*Cover illustration*: Brushing a cluster that was found by using the grand tour to explore two-dimensional projections of five-dimensional data on Australian crabs.

HTML HyperText Markup Language; claimed as a trademark or generic term by MIT, ERCIM, and/or Keio on behalf of the W3C. Data Desk is a registered trademark of Data Description, Inc. Spotfire is a registered trademark of Spotfire, Inc. JMP is a registered trademark of SAS Institute Inc. Linux is a registered trademark of Linus Torvalds, the original author of the Linux kernel. Mac OS X - Operating System software - is a registered trademark of Apple Computer, Inc. S-PLUS is a registered trademark of Insightful Corporation. Microsoft® is a registered trademark of Microsoft Corporation. Windows is a registered trademarks of Microsoft Corporation. XML - Extensible Markup Language; Language by W3C - claimed as a trademark or generic term by MIT, ERCIM, and/or Keio on behalf of the W3C. Facebook is a registered trademark of Facebook Inc. MySpace is a Registered Trademark of News Corp. SYSTAT is a registered trademark of Systat Software, Inc. SPSS is a registered trademark of SPSS Inc.

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

# Preface

This book is about using interactive and dynamic plots on a computer screen as part of data exploration and modeling, both alone and as a partner with static graphics and non-graphical computational methods. The area of interactive and dynamic data visualization emerged within statistics as part of research on exploratory data analysis in the late 1960s, and it remains an active subject of research today, as its use in practice continues to grow. It now makes substantial contributions within computer science as well, as part of the growing fields of information visualization and data mining, especially visual data mining.

The material in this book includes:

- An introduction to data visualization, explaining how it differs from other types of visualization.
- A description of our toolbox of interactive and dynamic graphical methods.
- An approach for exploring missing values in data.
- An explanation of the use of these tools in cluster analysis and supervised classification.
- An overview of additional material available on the web.
- A description of the data used in the analyses and exercises.

The book's examples use the software R and GGobi. R (Ihaka & Gentleman 1996, R Development Core Team 2006) is a free software environment for statistical computing and graphics; it is most often used from the command line, provides a wide variety of statistical methods, and includes high–quality static graphics. R arose in the Statistics Department of the University of Auckland and is now developed and maintained by a global collaborative effort. It began as a re-implementation of the S language and statistical computing environment (Becker & Chambers 1984) first developed at Bell Laboratories before the breakup of AT&T.

GGobi (Swayne, Temple Lang, Buja & Cook 2003) is free software for interactive and dynamic graphics; it can be operated using a command-line interface or from a graphical user interface (GUI). When GGobi is used as a stand-alone tool, only the GUI is used; when it is used with R, via the `rggobi` (Temple Lang, Swayne, Wickham & Lawrence 2006) package, a command-line interface is used along with the GUI. GGobi is a descendant of two earlier programs: XGobi (Swayne, Cook & Buja 1992, Buja, Cook & Swayne 1996) and, before that, Dataviewer (Buja, Hurley & McDonald 1986, Hurley 1987). Many of the examples that follow might be reproduced with other software such as S-PLUS®, SAS JMP®, DataDesk®, Mondrian, MANET, and Spotfire®. However, GGobi is unique because it offers tours (rotations of data in higher than 3D), complex linking between plots using categorical variables, and the tight connection with R.

*Web resources*

The web site which accompanies the book contains sample datasets and R code, movies demonstrating the interactive and dynamic graphic methods, and additional chapters. It can be reached through the GGobi web site:

`http://www.ggobi.org`

The R software is available from:

`http://www.R-project.org`

Both web sites include source code as well as binaries for various operating systems (Linux®, Windows®, Mac OS X®) and allow users to sign up for mailing lists and browse mailing list archives. The R web site offers a wealth of documentation, including an introduction to R and a partially annotated list of books offering more instruction. [Widely read books include Dalgaard (2002), Venables & Ripley (2002), and Maindonald & Braun (2003).] The GGobi web site includes an introductory tutorial, a list of papers, and several movies.

*How to use this book*

The language in the book is aimed at later year undergraduates, beginning graduate students, and graduate students in any discipline needing to analyze their own multivariate data. It is suitable reading for an industry statistician, engineer, bioinformaticist, or computer scientist with some knowledge of basic data analysis and a need to analyze high-dimensional data. It also may be useful for a mathematician who wants to visualize high-dimensional structures.

The end of each chapter contains exercises to help practice the methods discussed in the chapter. The book may be used as a text in a class on statistical graphics, exploratory data analysis, visual data mining, or information visualization. It might also be used as an adjunct text in a course on multivariate data analysis or data mining.

This book has been closely tied to a particular software implementation so that you can actively use the methods as you read about them, to learn and experiment with interactive and dynamic graphics. The plots and written explanations in the book are no substitute for personal experience. We strongly urge the reader to go through this book sitting near a computer with

GGobi, R, and `rggobi` installed, following along with the examples. If you do not wish to install the software, then the next best choice is to watch the accompanying movies demonstrating the examples in the text.

If you have not used GGobi before, then visit the web site, watch the movies, download the manual, and work through the tutorial; the same advice applies for those unfamiliar with R: Visit the R web site and learn the basics.

As you read the book, try things out for yourself. Take your time, and have fun!

*Acknowledgments*

This work started at Bellcore (now Telcordia), where Jon Kettering, Ram Gnanadesikan, and Diane Duffy carefully nurtured graphics research in the 1990s. It has continued with the encouragement of our respective employers at AT&T (with support from Daryl Pregibon and Chris Volinsky), at Lucent Bell Laboratories (with support from Diane Lambert), and at Iowa State University (with support from chairs Dean Isaacson and Ken Koehler). Werner Stuetzle provided the initial encouragement to write this book. A large part of the book was drafted while Dianne Cook was partially supported for a semester by Macquarie University, Sydney.

The authors would like to thank the peer reviewers recruited by the publisher for their many useful comments.

Proceeds from the sale of each book will be directed to the GGobi Foundation, to nurture graphics research and development.

Dianne Cook
Iowa State University

Deborah F. Swayne
AT&T Labs – Research

July 2007

# Contents

**Preface** .......... V

**Technical Notes** .......... XIII

**List of Figures** .......... XV

**1 Introduction** .......... 1
1.1 Data visualization: beyond the third dimension .......... 1
1.2 Statistical data visualization: goals and history .......... 3
1.3 Getting down to data .......... 4
1.4 Getting real: process and caveats .......... 8
1.5 Interactive investigation .......... 15

**2 The Toolbox** .......... 17
2.1 Introduction .......... 17
2.2 Plot types .......... 19
2.2.1 Univariate plots .......... 19
2.2.2 Bivariate plots .......... 21
2.2.3 Multivariate plots .......... 24
2.2.4 Plot arrangement .......... 34
2.3 Plot manipulation and enhancement .......... 35
2.3.1 Brushing .......... 35
2.3.2 Identification .......... 41
2.3.3 Scaling .......... 41
2.3.4 Subset selection .......... 42
2.3.5 Line segments .......... 43
2.3.6 Interactive drawing .......... 43
2.3.7 Dragging points .......... 43
2.4 Tools available elsewhere .......... 44
2.5 Recap .......... 45
Exercises .......... 45

**3 Missing Values** .......... 47
3.1 Background .......... 48
3.2 Exploring missingness .......... 49
3.2.1 Shadow matrix .......... 49
3.2.2 Getting started: missings in the “margins” .......... 52
3.2.3 A limitation .......... 53
3.2.4 Tracking missings using the shadow matrix .......... 55
3.3 Imputation .......... 55
3.3.1 Mean values .......... 56
3.3.2 Random values .......... 56
3.3.3 Multiple imputation .......... 58
3.4 Recap .......... 61
Exercises .......... 62

**4 Supervised Classification** .......... 63
4.1 Background .......... 64
4.1.1 Classical multivariate statistics .......... 65
4.1.2 Data mining .......... 66
4.1.3 Studying the fit .......... 69
4.2 Purely graphics: getting a picture of the class structure .......... 70
4.2.1 Overview of Italian Olive Oils .......... 70
4.2.2 Building classifiers to predict region .......... 71
4.2.3 Separating the oils by area within each region .......... 73
4.2.4 Taking stock .......... 77
4.3 Numerical methods .......... 77
4.3.1 Linear discriminant analysis .......... 77
4.3.2 Trees .......... 81
4.3.3 Random forests .......... 83
4.3.4 Neural networks .......... 88
4.3.5 Support vector machine .......... 92
4.3.6 Examining boundaries .......... 97
4.4 Recap .......... 99
Exercises .......... 99

**5 Cluster Analysis** .......... 103
5.1 Background .......... 105
5.2 Purely graphics .......... 107
5.3 Numerical methods .......... 111
5.3.1 Hierarchical algorithms .......... 111
5.3.2 Model-based clustering .......... 113
5.3.3 Self-organizing maps .......... 119
5.3.4 Comparing methods .......... 122
5.4 Characterizing clusters .......... 125
5.5 Recap .......... 126
Exercises .......... 127

**6 Miscellaneous Topics** . . . . . 129
6.1 Inference . . . . . 129
6.2 Longitudinal data . . . . . 134
6.3 Network data . . . . . 139
6.4 Multidimensional scaling . . . . . 145
Exercises . . . . . 151

**7 Datasets** . . . . . 153
7.1 Tips . . . . . 153
7.2 Australian Crabs . . . . . 154
7.3 Italian Olive Oils . . . . . 155
7.4 Flea Beetles . . . . . 157
7.5 PRIM7 . . . . . 157
7.6 Tropical Atmosphere-Ocean Array (TAO) . . . . . 159
7.7 Primary Biliary Cirrhosis (PBC) . . . . . 161
7.8 Spam . . . . . 162
7.9 Wages . . . . . 164
7.10 Rat Gene Expression . . . . . 166
7.11 Arabidopsis Gene Expression . . . . . 168
7.12 Music . . . . . 171
7.13 Cluster Challenge . . . . . 172
7.14 Adjacent Transposition Graph . . . . . 172
7.15 Florentine Families . . . . . 173
7.16 Morse Code Confusion Rates . . . . . 174
7.17 Personal Social Network . . . . . 175

**References** . . . . . 177

**Index** . . . . . 185

# Technical Notes

*R code*

The R code in this book, denoted by `typewriter font`, and the more extensive code on the web site, has been tested on version 2.4.0 of R, version 2.1.5 of GGobi, and version 2.1.5 of `rggobi`. Updates will be available on the web site as they are needed.

*Figures*

The figures in this book were produced in a variety of ways, and the files and code to reproduce them are all available on the book's web site. Some were produced directly in R. Some were produced using both GGobi and R, and the process of converting GGobi views into publication graphics deserves an explanation.

When we arrive at a GGobi view we want to include in a paper or book, we use the `Save Display Description` item on GGobi's `Tools` menu to generate a file containing an S language description of the display. We read the file into R using the R package `DescribeDisplay` (Wickham 2006*b*), like this:

```
> library(DescribeDisplay)
> d <- dd_load("fig.R")
```

We create the publication-quality graphic using either that package's plot method or another R package, `ggplot` (Wickham 2006*c*), like this:

```
> plot(d)
```

or

```
> p <- ggplot(d)
> print(p)
```

Figure 0.1 illustrates the differences with a trio of representations of the same bivariate scatterplot. The picture at left is a screen dump of a GGobi display. Such images are not usually satisfactory for publication for several

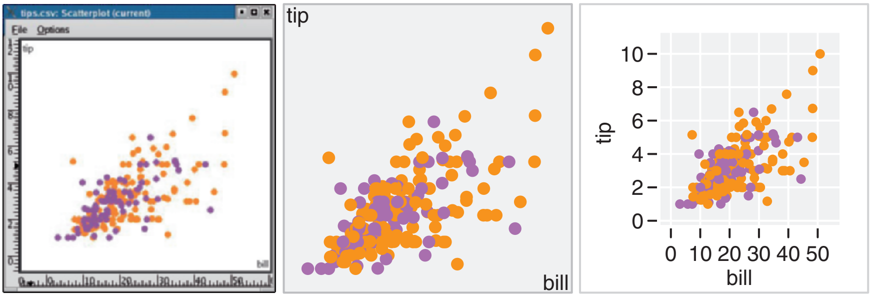

**Fig. 0.1.** Sample plots produced from GGobi in different ways: **(left)** a simple screen dump; **(middle)** a plot produced using the plot method of the R package `DescribeDisplay`; **(right)** a plot made using the R package `ggplot`.

reasons, the most obvious of which is the lack of resolution. The second picture was produced using `DescribeDisplay`'s plot method, which reproduces the plotting region of the view with pretty good fidelity. We used this method to produce most of the one–dimensional and two-dimensional tour pictures in this book. The third picture was produced using `ggplot`, which adds axis ticks, labels and grid lines. We used it to produce nearly all the bivariate scatterplots of GGobi views in this book.

# List of Figures

0.1 Sample plots produced from GGobi . . . . . . . . . . . . . . . . . . . . . . . . XIV

1.1 Histograms of tip with differing bin width . . . . . . . . . . . . . . . . . . . 6
1.2 Scatterplot of tip vs. bill . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
1.3 Scatterplot of tip vs. bill conditioned by sex and smoker . . . . . . . 9
1.4 Factors affecting tipping behavior . . . . . . . . . . . . . . . . . . . . . . . . . . 12
1.5 Histogram of tip linked to 2D mosaic plots of sex and smoker . . 16

2.1 Univariate displays . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.2 Barchart and spine plot of the Tips data . . . . . . . . . . . . . . . . . . . . 22
2.3 Bivariate scatterplots for the Blue female Australian Crabs . . . . . 22
2.4 A mosaic plot of day and sex for the Tips data . . . . . . . . . . . . . . . 23
2.5 Parallel coordinate plot of the Australian Crabs data . . . . . . . . . . . 24
2.6 Scatterplot matrix of the real-valued variables in Australian Crabs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
2.7 One-dimensional projections of the Crabs data . . . . . . . . . . . . . . . 27
2.8 Two-dimensional projections of the Crabs data . . . . . . . . . . . . . . . 28
2.9 Two-dimensional projections found by projection pursuit . . . . . . 32
2.10 One-dimensional projections found by projection pursuit . . . . . . 33
2.11 The relationship between a 2D tour and the biplot . . . . . . . . . . . . 35
2.12 A dynamic query posed using linked brushing . . . . . . . . . . . . . . . . 36
2.13 Categorical brushing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
2.14 Linking by subject identifier . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
2.15 Linking between a point in one plot and a line in another, using Arabidopsis Gene Expression . . . . . . . . . . . . . . . . . . . . . . . . . . 40
2.16 Transient brushing contrasted with persistent painting . . . . . . . . 40
2.17 Identifying points in a plot using three different labeling styles . 41
2.18 Scaling a plot reveals different aspects of the River Flow data . . . 42

3.1 Assigning constants to missing values . . . . . . . . . . . . . . . . . . . . . . . 52
3.2 Tour views with missings set to 10% below minimum . . . . . . . . . 54

3.3 Parallel coordinates plot with missings set to 10% below minimum . . . 54
3.4 Exploring missingness in the TAO data . . . 56
3.5 Comparison of simple, widely used imputation schemes . . . 57
3.6 Conditional random imputation . . . 58
3.7 Two different imputations using simulation from a multivariate normal distribution of all missing values . . . 60
3.8 Tour projection of the data after multiple imputation . . . 61

4.1 Evaluating model assumptions for 2D data . . . 67
4.2 Classification of the data space for the Flea Beetles . . . 68
4.3 Differentiating the oils from the three regions in the Olive Oils data in univariate plots . . . 71
4.4 Separation between the Northern and Sardinian oils . . . 72
4.5 Finding variables for classification using a parallel coordinate plot of Olive Oils . . . 73
4.6 Separating the oils of the Northern region by area . . . 75
4.7 Separating the oils of the Southern region by area . . . 76
4.8 Evaluating model assumptions for the 6D Flea Beetles data . . . 78
4.9 Misclassifications from an LDA classifier of the Olive oils. . . . 79
4.10 Improving on the results of the tree classifier using the manual tour . . . 82
4.11 Examining the results of a forest classifier of Olive Oils . . . 84
4.12 Examining the results of a random forest after classifying the oils of the South . . . 87
4.13 Misclassifications of a feed-forward neural network classifying the oils from the South . . . 91
4.14 Misclassifications of a support vector machine classifying the oils of the South . . . 96
4.15 Using the tour to examine the choice of support vectors . . . 96
4.16 Classification boundaries for different models . . . 98

5.1 Structures in data and their impact on cluster analysis . . . 104
5.2 Clustering the example data . . . 106
5.3 Stages of “spin and brush” on PRIM7 . . . 109
5.4 The PRIM7 model summarized . . . 110
5.5 Hierarchical clustering of the particle physics data . . . 112
5.6 Model-based clustering of a reduced set of Australian Crabs . . . 115
5.7 Comparing the Australian Crabs data with results of model-based clustering . . . 118
5.8 Comparison of clustering music tracks using a self-organizing map versus principal components . . . 120
5.9 Unsuccessful SOM fit . . . 122
5.10 Successful SOM fit . . . 123

5.11 Comparing two five-cluster models of the Music data using confusion tables linked to tour plots . . . 125
5.12 Characterizing clusters in a parallel coordinate plot . . . 126

6.1 Comparison of various pairwise relationships with the same correlation . . . 130
6.2 Testing the assumption that species labels correspond to groups in the data . . . 132
6.3 Testing the assumption that species labels correspond to groups in simulated data . . . 133
6.4 Exploring the longitudinal Wages data . . . 135
6.5 Using linked brushing to explore subject profiles . . . 138
6.6 Business and marital relationships among Renaissance Florentine families . . . 141
6.7 Highlighting the links involving the greatest number of ties . . . 142
6.8 The $n = 3$ adjacency transposition graph . . . 143
6.9 The $n = 4$ adjacency transposition graph in various layouts . . . 144
6.10 A grid and its distance matrix . . . 146
6.11 Snapshots from MDS animation . . . 148
6.12 Applying MDS to the Morse Code confusion data . . . 149
6.13 Studying MDS results . . . 150
6.14 3D MDS layouts of the Morse Code . . . 150