Chapman & Hall/CRC

Computer Science and Data Analysis Series

The interface between the computer and statistical sciences is increasing, as each discipline seeks to harness the power and resources of the other. This series aims to foster the integration between the computer sciences and statistical, numerical, and probabilistic methods by publishing a broad range of reference works, textbooks, and handbooks.

Proposals for the series should be sent directly to one of the series editors above, or submitted to:

Chapman & Hall/CRC
23-25 Blades Court
London SW15 2NU
UK

Published Titles

Bayesian Artificial Intelligence
Kevin B. Korb and Ann E. Nicholson

Pattern Recognition Algorithms for Data Mining
Sankar K. Pal and Pabitra Mitra

Exploratory Data Analysis with MATLAB®
Wendy L. Martinez and Angel R. Martinez

Clustering for Data Mining: A Data Recovery Approach
Boris Mirkin

Correspondence Analysis and Data Coding with Java and R
Fionn Murtagh

R Graphics
Paul Murrell

Computer Science and Data Analysis Series

R Graphics

Paul Murrell

The University of Auckland
New Zealand

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York Singapore

Published in 2006 by
Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-486-X (Hardcover)
International Standard Book Number-13: 978-1-58488-486-6 (Hardcover)
Library of Congress Card Number 2005046278

Library of Congress Cataloging-in-Publication Data

Murrell, Paul.
R graphics / Paul Murrell.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-486-X
1. Computer graphics. 2. R (Computer program language) I. Title.

T385.M9 2005
006.6—dc22 2005046278

T&F informa
Taylor & Francis Group
is the Academic Division of T&F Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Preface

R is a popular open source software tool for statistical analysis and graphics. This book focuses on the very powerful graphics facilities that R provides for the production of publication-quality diagrams and plots.

What this book is about

This book describes the graphics system in R. The first chapter provides an overview of the R graphics facilities. There are many pictures that demonstrate the variety and complexity of plots and diagrams that can be produced using R. There is a description of the different output formats that R graphics can produce and there is a description of the overall organization of the R graphics facilities, so that the user has some idea of where to find a function for a particular purpose.

The most important feature of the R graphics setup is the existence of two distinct graphics systems within R: the traditional graphics system and the grid graphics system. Section 1.2.2 offers some advice on which system to use. Part I of this book is concerned with the traditional graphics system, which implements many of the "traditional" graphics facilities of the S language[11][5] (originally developed at Bell Laboratories and available in a commercial implementation as S-PLUS). The majority of R graphics functions available at the time of writing are based upon this system. The chapters in this part of the book describe how to work with the traditional graphics functions, with a particular emphasis on how to modify or add output to a plot to produce exactly the right final output. Chapter 2 describes the functions that are available to produce complete plots and Chapter 3 focuses on how to customize the details of plots, combine multiple plots, and add further output to plots.

Part II describes the grid graphics system, which is unique to R and is much more powerful than the traditional system. At the time of writing, there are fewer functions based on grid for producing complete plots, but there is more power to produce a wider range of final results. Most of the functions that produce complete plots using grid come from Deepayan Sarkar's lattice package, which implements Bill Cleveland's Trellis graphics system. This is described in Chapter 4. The remaining chapters describe how the grid system can be used to produce graphical scenes starting from a blank page. In particular, there is a discussion of how to develop new graphical functions

that are easy for other people to use and build on.

Appendix A provides a very brief introduction to the R system in general and Appendix B discusses ways in which the traditional and grid graphics systems can be combined.

The main part of the book assumes a basic familiarity with the R language and environment. For more detailed information, the reader is directed to the home page of the R Project (the URL is given below), which has links to on-line documents and references to printed material.

There are a number of projects working on graphical user interfaces to R, but the common underlying method of interaction is via a command line. This book focuses on the production of graphical output by entering R code interactively at the command-line interface to R and writing code in scripts to load into R or to run as a batch job.

What this book is not about

This book does *not* contain discussions about which sort of plot is most appropriate for a particular sort of data, nor does it contain guidelines for correct graphical presentation. In fact, instructions are provided for producing some types of plots and graphical elements that are generally disapproved of, such as pie charts and cross-hatched fill patterns.

The information in this book is meant to be used to produce a plot once the format of the plot has been decided upon and to experiment with different ways of presenting a set of data. No plot types are deliberately excluded, partly because no plot type is all bad (e.g., a pie chart can be a very effective way to present a simple proportion) and partly because some graphical elements, such as cross-hatching, are sometimes required by a particular publisher.

The flexibility of R graphics encourages the user *not* to be constrained to thinking in terms of just the traditional types of plots. The aim of this book is to provide lots of useful tools and to describe how to use them. There are many other sources of information on graphical guidelines and recommended plot types, some of which are mentioned below.

Most introductory statistics text books will contain basic guidelines for selecting an appropriate type of plot. Examples of books that deal specifically with the construction of effective plots and are aimed at a general audience are "Creating More Effective Graphs" by Naomi Robbins[51] and Edward Tufte's "Visual Display of Quantitative Information"[60] and "Envisioning Information"[61]. For more technical discussions of these issues, see "Visualizing Data" and "Elements of Graphing Data" by Bill Cleveland[12][13], and "The Grammar of Graphics" by Leland Wilkinson[67].

For ideas on appropriate graphical displays for particular types of analysis or particular types of data, some starting points are "Data analysis and graphics using R" by John Maindonald and John Braun[37], "An R and S-Plus Companion to Applied Regression" by John Fox[20], "Statistical Analysis and Data Display" by Richard Heiberger and Burt Holland[29], and "Visualizing Categorical Data" by Michael Friendly[25].

This book is also *not* a complete reference to the R system. Appendix A provides a very brief introduction to R, but there are many freely-available documents that provide both introductory and in-depth explanations of the R system. The best place to start is the "Documentation" section on the home page of the R project web site (see "On the web" on page ix). Two examples of introductory texts are "Introductory Statistics with R" by Peter Dalgaard[18] and "Using R for Introductory Statistics" by John Verzani[65]; the standard advanced text is "Modern Applied Statistics with S" by Bill Venables and Brian Ripley[64].

Finally, this book does *not* describe in any detail the many graphics functions that are available in add-on packages for R that are *not* part of the standard R installation. This book only focuses on the graphics facilities that are distributed with R by default — in particular, functions in the `grDevices`, `graphics`, `grid`, and `lattice` packages. No attempt is made to enumerate all existing graphics functions for R or even to list all add-on packages that contain graphics functions; the list is very long and growing all the time. Except where specified, all add-on packages mentioned in this book are available from CRAN*, the main download site for R.

Differences with S-PLUS

The traditional graphics system in R is a reimplementation of the traditional graphics system in the original S language. This means that much of what is said about the traditional system in Part I of this book is also true for the traditional graphics in the commercial distribution of S, S-PLUS. However, there are some important differences between traditional R graphics and traditional S graphics, such as the specification of colors and line types by character strings, the concept of layouts for arranging plots, and the availability of mathematical annotation in text. These differences mean that graphics code written for R is not guaranteed to produce the same result (or even run) in S-PLUS. Furthermore, the grid graphics system described in Part II is not available in S-PLUS (just as the S-PLUS editable graphics are not available in R).

This book focuses on the graphics systems available in R so specific differences

*The Comprehensive R Archive Network; `http://cran.r-project.org`

with S-PLUS are not highlighted in the main text. However, much of what is said in Part I will also apply to traditional graphics in S-PLUS.

Who should read this book

This book should be of interest to a variety of R users. For people who are new to R, this book provides an overview of the graphics system, which is useful for understanding what to expect from R's graphics functions and how to modify or add to the output they produce. For this purpose, Chapter 1 and Chapter 2 are a good starting point from which to begin producing standard plots, but you will soon need to start dipping into Chapter 3 in order to fine tune your plots. It would also be worthwhile to take a look at Chapter 4 to see what Trellis plots can do.

For intermediate-level R users, this book provides all of the information necessary to perform sophisticated customizations of plots produced in R. As with many software applications, it is possible to work with R for years and remain unaware of important and useful features. This book will be useful in making users aware of the full scope of R graphics, and in providing a description of the correct model for working with R graphics. Sections 1.2, 1.3, and Chapters 3 and 4 should be read first. Chapters 5, 6, and 7 should be read by users interested in experimenting with novel graphical displays.

For advanced R users, this book contains vital information for producing coherent, reusable, and extensible graphics functions. Advanced users should pay particular attention to Part II.

Conventions used in this book

This book describes a large number of R functions and there are many code examples. Samples of code that could be entered interactively at the R command line are formatted as follows:

```
> 1:10
```

where the > denotes the R command-line prompt and everything else is what the user should enter. When an expression is longer than a single line it will look like this:

```
> plot(1:10, 1:10, col="blue", lty="dashed",
       axes=FALSE, type="l")
```

with the additional lines indented appropriately.

Often, the functions described in this book are used for the side-effect of producing graphical output, so the result of running a function is represented

by a figure. In cases where the result of a function is a value that we might be interested in, the result will be shown below the code that produced it and will be formatted as follows:

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

In some places, an entirely new R function is defined. Such code would normally be entered into a script file and loaded into R in one step (rather than being entered at the command line), so the code for new R functions will be presented in a figure and formatted as follows:

```
1  myfun <- function(x, y) {
2    plot(x ,y)
3  }
```

with line numbers provided for easy reference to particular parts of the code from the main text.

When referring to a function within the main text, it will be formatted in a `typewriter` font and will have parentheses after the function name, e.g., `plot()`.

When referring to the arguments to a function or the values specified for the arguments, they will also be formatted in a `typewriter` font, but they will not have any parentheses at the end, e.g., `x`, `y`, or `col="red"`.

When referring to an S3 class, statements will be of the form: "the `"classname"` class," using a typewriter font with the class name in double-quotes. However, when referring to an object that is an instance of a class, statements will be of the form: "the `classname` object," using a typewriter font, but without the double-quotes around the class name.

On the web

There is a web site (URL below) with errata and links to pages of PNG versions of all figures from the book and the R code used to produce them.

http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html

There is also an `RGraphics` package containing functions to produce the figures in this book and all functions, classes, and methods defined in the book (see especially Chapter 7). This package is available from CRAN (see the footnote on page vii).

Version information

Software development is an ongoing process and this book can only provide a snapshot of R's graphics facilities. The descriptions and code samples in this book are accurate for R version 2.1.0 and above. Apart from a couple of places, mostly in Chapter 7, code examples are also accurate for R version 2.0.1. In each of these cases, there is a footnote to highlight the difference and, if possible, to provide information about how to modify the code so that it will work in R version 2.0.1. Much of the content of Part I is also accurate for earlier versions of R, but specific areas of incompatibility are not indicated in the text.

A new "minor" version of R is released approximately every six months. The most up-to-date information on the most recent versions of R and grid are available in the on-line help pages and at the home pages for the R Project and the grid package:

http://www.R-project.org/
http://www.stat.auckland.ac.nz/~paul/grid/grid.html

Acknowledgements

R graphics could not exist without R itself, so the first thanks go to Ross Ihaka and Robert Gentleman for starting the whole thing. Thanks to the R Core Team in general for making R such a reliable, high-quality piece of software, and to the wider R community for making working with R so rewarding and enjoyable.

The traditional R graphics system owes most of its success and popularity to the excellence of the design of the original S graphics system. Most credit for the R-specific extensions to the traditional system is due to Ross Ihaka. For the grid system, I am almost entirely to blame.

With regard to this book in particular, I would like to thank John Chambers, Ross Ihaka, Duncan Murdoch, Stefano Iacus, Deepayan Sarkar, and the anonymous reviewers for valuable feedback on the manuscript.*

Last, and most, thank you Ju.

Auckland, *Paul Murrell*
New Zealand,

*This manuscript was generated on a Fedora Core 1 Linux system using the LaTeX document preparation system, Friedrich Leisch's `Sweave` package, several of the GNU software tools, and of course R.

Contents

List of Figures

List of Tables