

Data Visualisation with R

Thomas Rahlf

Data Visualisation with R

111 Examples

Second Edition

 Springer

Thomas Rahlf
Rheinische Friedrich-Wilhelms-Universität
Bonn
Bonn, Germany

ISBN 978-3-030-28443-5 ISBN 978-3-030-28444-2 (electronic)
<https://doi.org/10.1007/978-3-030-28444-2>

© Springer Nature Switzerland AG 2017, 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface to the Second English Edition

Due to the continued interest in the book, Springer Verlag has decided to publish an English translation of the second edition. Again, many thanks to Ralf Gerstner from Springer, and also to Tracey Duffy for translating all of the additions.

In 2017, Ulrike Grömping reviewed the English translation of the first edition in the *Journal of Statistical Software*. Inspired by the book, and following the release of the second edition, she has published an R package “preplot” which greatly simplifies the configuration of figure regions for base R graphics. I highly recommend trying this one out.

Meanwhile, the third edition of Paul Murrell’s standard work *R Graphics* has been published. When I read in his foreword that my book—among others—was one inspiration for restructuring part of his book, I felt very honored.

Bonn, Germany

Thomas Rahlf

Preface to the Second German Edition

The feedback on the first edition of this book (published by Open Source Press in 2014 and now out of print), and on the English edition that has been published in the meantime, was so pleasing that I was happy to accept the offer made by Springer Verlag to publish a new, updated edition of the book. The concept of explaining complete examples and the restriction to Base Graphics have been retained. Compared to the first edition of the book, this new, updated edition contains 11 additional, new examples, hence the new subtitle “111 Examples”. There are two main additions: firstly, a section on visualising network relationships has been added to the chapter on categorical data. In addition to examples of classic network diagrams, an adapted heat map, and a multiple bar chart, this section also contains a chord diagram and a riverplot. Although the chord diagram and riverplot may appear unusual at first glance, they can now be found in publications of respected scientific journals such as *Science*, *Nature* or *Cell*. Examples on the use of georeferenced grid formats and on cartograms have been added to the chapter on maps. Three examples for the integration of data created with R in interactive JavaScript illustrations have also been added to the book. R now provides multiple concepts and packages that can be used to create JavaScript visualisations more or less directly. Ultimately, such packages form a type of container in R. In each case, a specific syntax developed by the authors of these packages translates the scripts written in this form into the notation required for the underlying JavaScript library. This means that we have to rely on the scope of the language of the R package and on the quality and flexibility of the translation routines. I am not sure if this is the right path. In this book, I have taken a different path. In three examples in Chap. 12, the data are prepared with R such that they are integrated in existing, only slightly adapted JavaScript code. This is done using *Highcharts* and *Mapael*, two JavaScript libraries that can be used to create very aesthetic illustrations “out of the box” with minimal change effort.

The primary objective of this book is still to explain how to create presentation graphics. For the exploratory visualisation within the scope of data analysis, I refer to the book *Graphical Data Analysis with R* (CRC Press, 2015) by Antony Unwin. The first people I would like to thank are Agnes Herrmann and Iris Ruhmann at

Springer, who have enabled me to create this updated edition of the book. For helpful comments, suggestions, and exchange of ideas for this edition, I would also like to thank Alberto Cairo, Martin S. Fischer, Sebastian Jeworutzki, Nikola Sander, Antony Unwin, January Weiner, and Stefan Fichtel. January Weiner has kindly included a comment in his riverplot package that is helpful for example 6.4.4.

Bonn, Germany

Thomas Rahlf

Preface to the English Edition

This book is a translation of the German book “Datendesign mit R” that was published 2014 by Open Source Press. Due to the encouraging strong interest in the German edition Springer Verlag offered to publish an English translation. First of all I would like to thank Ralf Gerstner from Springer for this and for his helpful suggestions for improvement, as well as Annika Brun for translating most of the text, Colin Marsh for copy editing, and Katja Diederichs for converting all scripts from German to English. Last year I benefited a lot from a communication with Antony Unwin. His book “Graphical Data Analysis with R” can be seen as complementary to my own: while this one focusses on presentation of graphics, you will benefit from his book if you are interested in exploring data graphically.

Bonn, Germany

Thomas Rahlf

Preface to the German Edition

Some 20 years ago, when I reviewed a score of books on statistical graphics and graphic-based data analysis, things were completely different: there were proprietary formats and operating systems, their character sets were incompatible, and graphic and statistical software was expensive. Since the turn of the century, the situation has changed fundamentally: the Internet has come of age, open-source projects have attracted more and more followers, and a handful of enthusiasts provided version 1.0 of the free statistical programming language R. Many developers were inspired to collaborate on this project. R reached version 3 in 2013, and in addition to the basic software, more than 7000 freely available extension packs are currently available. Companies and organisations such as Google, Facebook or the CIA are using R for their data analysis. Its graphic capabilities are again and again emphasised as its strong point. Pretty much all technologies relevant for data visualisation are quickly integrated into R. Through numerous functions, detailed designs of every imaginable figure, creation of maps and much more are made possible. All it takes is to know how—and that is where this book wants to contribute.

What This Book Wants to Be—and What It Doesn't Want to Be

This book is not an introduction that systematically explains all the graphic tools R has to offer. Rather, its aim is to use 100 complete script examples to introduce the reader to the basics of designing presentation graphics, and to show how bar and column charts, population pyramids, Lorenz curves, box plots, scatter plots, time series, radial polygons, Gantt charts, heat maps, bump charts, mosaic and balloon charts, and a series of different thematic map types can be created using

R's Base Graphics System. Every example uses real data and includes step-by-step explanations of the figures and their programming. The selection is based on my personal experiences—it is likely that readers will find one or another illustration lacking, and consider some too detailed. However, a large scope should be covered. This book is aimed at:

- R experts: You can most likely skip Part I. For you, the examples are particularly useful, especially the code.
- Readers that have heard of R and maybe even tried R before and are not daunted by programming; you will profit from both parts.
- Beginners: for you, the finished graphics pictured here will be most helpful. You will see what R can do. Or, in other words: you will realise that there is such a tool as R, and that it can be used to create graphics you have wanted to create for a long time, but merely never knew how. The code will be too complicated for you, but you may be able to commission others to do your graphics programming in R.

Windows, Mac, and Linux

All of the scripts and working steps will yield identical results when executed in Windows, Mac OS X or Linux. All of the examples were created in Mac OS X and then tested in Ubuntu 12.04 and an evaluation copy of Windows 8.1.

Acknowledgements

The following people deserve my thanks for hints, comments, feedback, data, discussions or help: Gregor Aisch, Insa Bechert, Evelyn Brislinger, Giuseppe Casalicchio, Arnulf Christl, Katja Diederichs, Günter Faes, Mira Hassan, Mark Heckmann, Daniel Hienert, Bruno Hopp, Uwe Ligges, Lorenz Matzat, Meinhard Moschner, Stefan Müller, Paul Murrell, David Phillips, Duncan Temple Lang, Martijn Tennekes, Patrick R. Schmid, Thomas Schraitle, Valentin Schröder, Torsten Steiner, Michael Terwey, Katrin Weller, Bernd Weiss, Nils Windisch, Benjamin Zupilko, and Lisa Zhang. This manuscript particularly benefited from discussions with an infographic designer and a data journalist. Stefan Fichtel looked over every figure and provided critical feedback. For selected figures, he designed his own suggestions; this has been an invaluable help. We did not always agree, and I have disregarded his advice here or there. Therefore, any remaining errors and shortcomings are mine. Björn Schwentker went to the trouble of proof reading large parts of the manuscript. I am very grateful for his valuable notes which have surely

made some parts of the text clearer and more readable. Finally, I want to thank Markus Wirtz for tackling the experiment of ultimately printing everything into a book.

On the Internet

The figures are conceived for different final output options. The format of the book implies that some details have become very small, e.g. in maps and radial column charts. Particularly for such cases, please refer to the book's website, on which all figures are available in high resolution or as vector graphics in PDF format:

<http://www.datavisualisation-r.com>

Bonn, Germany

Thomas Rahlf

Contents

1	Data for Everybody	1
1.1	Data Visualisation Between Science and Journalism	1
1.2	Why R?	3
1.3	The Concept of Data Design	3
Part I Basics and Techniques		
2	Structure and Technical Requirements	7
2.1	Terms and Elements	7
2.2	Illustration Grids	7
2.3	Perception	10
2.4	Typefaces	13
2.4.1	Fonts	15
2.4.2	Free Typefaces	16
2.5	Symbols	18
2.5.1	Symbol Fonts	19
2.5.2	Symbols in SVG Format	21
2.6	Colour	21
2.6.1	Colour Models	21
2.6.2	Colour in Statistical Illustrations	23
3	Implementation in R	25
3.1	Installation	25
3.2	Basic Concepts in R	26
3.2.1	Data Structures	27
3.2.2	Import of Data	30
3.3	Graphic Concepts in R	37
3.3.1	The Paper-Pencil-Principle of the Base Graphics System: High-Level and Low-Level Functions	42
3.3.2	Graphic Parameter Settings	44
3.3.3	Margin Settings for Figures and Graphics	50
3.3.4	Multiple Charts: Panels with mfrow and mfcol	51

3.3.5	More Complex Assembly and Layout	53
3.3.6	Font Embedding	55
3.3.7	Output with <code>cairo_pdf</code>	56
3.3.8	Unicode in Figures	57
3.3.9	Colour Settings	60
3.4	R Packages and Functions Used in This Book	61
3.4.1	Packages	61
3.4.2	Functions	66
3.4.3	Schematic Approach	75
4	Beyond R	77
4.1	Additions with LaTeX	77
4.2	Manual Post-processing and Creation of Icon Fonts Using Inkscape	82
4.2.1	Post-processing	82
4.2.2	Creation of Icon Fonts	84
5	Regarding the Examples	89
5.1	An Attempt at a Systematics	89
5.2	Getting the Scripts Running	91

Part II Examples

6	Categorical Data	95
6.1	Bar and Column Charts	95
6.1.1	Bar Chart Simple	96
6.1.2	Bar Chart for Multiple Response Questions: First Two Response Categories	101
6.1.3	Bar Chart for Multiple Response Questions: All Response Categories	105
6.1.4	Bar Chart for Multiple Response Questions: All Response Categories, Variant	108
6.1.5	Bar Chart for Multiple Response Questions: All Response Categories (Panel)	110
6.1.6	Bar Chart for Multiple Response Questions: Symbols for Individuals	113
6.1.7	Bar Chart for Multiple Response Questions: All Response Categories, Grouped	116
6.1.8	Column Chart with Two-Line Labelling	121
6.1.9	Column Chart with 45° Labelling	123
6.1.10	Profile Plot for Multiple Response Questions: Mean Values of the Responses	124

6.1.11	Dot Chart for Three Variables	127
6.1.12	Column Chart with Shares	130
6.2	Pie Charts and Radial Diagrams	132
6.2.1	Simple Pie Chart	133
6.2.2	Pie Charts, Labels Inside (Panel)	135
6.2.3	Seat Distribution (Panel)	136
6.2.4	Spie Chart	139
6.2.5	Radial Polygons (Panel)	142
6.2.6	Radial Polygons (Panel): Different Column Arrangement	144
6.2.7	Radial Polygons Overlay	145
6.3	Chart Tables	146
6.3.1	Simplified Gantt Chart	147
6.3.2	Simplified Gantt Chart: Colours by People	151
6.3.3	Bump Chart	152
6.3.4	Heat Map	155
6.3.5	Mosaic Plot (Panel)	158
6.3.6	Balloon Plot	160
6.3.7	Tree Map	162
6.3.8	Tree Maps for Two Levels (Panel)	164
6.4	Network Relationships	168
6.4.1	Undirected Network	169
6.4.2	Chord Diagram	174
6.4.3	Directed Network	178
6.4.4	Riverplot	182
6.4.5	Heat Map for Relationships	185
6.4.6	Multiple Bar Chart	188
7	Distributions	191
7.1	Histograms and Box Plots	191
7.1.1	Histograms Overlay	191
7.1.2	Column Charts Coloured with ColorBrewer (Panel)	193
7.1.3	Histograms (Panel)	196
7.1.4	Box Plots for Groups: Sorted in Descending Order	199
7.1.5	Box Plots for Groups: Sorted in Descending Order, Comparison of Two Polls	203
7.2	(Population) Pyramids	207
7.2.1	Pyramid with Multiple Colours	208
7.2.2	Pyramids: Emphasis on the Outer Areas (Panel)	210
7.2.3	Pyramids: Emphasis on the Inner Areas (Panel)	213
7.2.4	Pyramids with Added Line (Panel)	216
7.2.5	Aggregated Pyramids	217
7.2.6	Bar Charts as Pyramids (Panel)	220
7.3	Inequality	223
7.3.1	Simple Lorenz Curve	224

7.3.2	Lorenz Curves Overlay	226
7.3.3	Lorenz Curves (Panel)	229
7.3.4	Comparison of Income Proportions with Bar Chart (Quintile)	231
7.3.5	Comparison of Income Proportions with Bar Chart (Decile)	234
7.3.6	Comparison of Income Proportion with Panel-Bar Chart (Quintile)	236
8	Time Series	239
8.1	Short Time Series	239
8.1.1	Column Chart for Developments	239
8.1.2	Column Chart with Percentages for Growth Developments	242
8.1.3	Quarterly Values as Columns	245
8.1.4	Quarterly Values as Lines with Value Labels	247
8.1.5	Short Time Series Overlayed	249
8.2	Areas Underneath and Between Time Series	252
8.2.1	Areas Between Two Time Series	252
8.2.2	Areas as Corridor with Time Series (Panel)	254
8.2.3	Forecast Intervals (Panel)	256
8.2.4	Forecast Intervals Index (Panel)	259
8.2.5	Time Series with Stacked Areas	262
8.2.6	Areas Under a Time Series	264
8.2.7	Time Series with Trend (Panel)	267
8.3	Presentation of Daily, Weekly and Monthly Values	270
8.3.1	Daily Values with Labels	270
8.3.2	Daily Values with Labels and Week Symbols (Panel)	272
8.3.3	Daily Values with Monthly Labels	276
8.3.4	Time Series from Weekly Values (Panel)	277
8.3.5	Monthly Values (Panel)	280
8.3.6	Monthly Values with Monthly Labels	282
8.3.7	Monthly Values with Monthly Labels (Layout)	285
8.4	Exceptions and Special Cases	288
8.4.1	Time Series as Scatter Plot (Panel)	288
8.4.2	Time Series with Missing Values	290
8.4.3	Seasonal Ranges (Panel)	293
8.4.4	Seasonal Ranges Stacked	295
8.4.5	Season Figure (Seasonal Subseries Plot) with Data Table	297
8.4.6	Temporal Ranges	300
9	Scatter Plots	303
9.1	Variants	305
9.1.1	Scatter Plot Variant 1: Four Quadrants Differentiated by Colour	305
9.1.2	Scatter Plot Variant 2: Outliers Highlighted	308

9.1.3	Scatter Plot Variant 3: Areas Highlighted	311
9.1.4	Scatter Plot Variant 4: Superimposed Ellipse	313
9.1.5	Scatter Plot Variant 5: Connected Points	316
9.2	Exceptions and Special Cases	319
9.2.1	Scatter Plot with Few Points	319
9.2.2	Scatter Plot with User-Defined Symbols	321
9.2.3	Map of Germany as Scatter Plot	324
10	Maps	327
10.1	Introductory Examples	327
10.1.1	Maps of Germany: Local Telephone Areas and Postcode Districts.....	327
10.1.2	Filtered Postcode Map.....	329
10.1.3	Map of Europe NUTS 2006 (Cut-out).....	331
10.2	Points, Diagrams, and Symbols in Maps.....	333
10.2.1	Map of Germany with Selected Locations and Outline (Panel).....	333
10.2.2	Map of Germany with Selected Locations (Pie Charts) and Outline	336
10.2.3	Map of Germany with Selected Locations (Columns) and Outline	338
10.2.4	Map of Germany as Three-Dimensional Scatter Plot	341
10.2.5	Map of North Rhine-Westphalia with Selected Locations (Symbols) and Outline	345
10.2.6	Map of Tunisia with Self-Defined Symbols	347
10.3	Choropleth Maps	350
10.3.1	Choropleth Map of Germany at District-Level	350
10.3.2	Choropleth Map of Germany at District-Level (Panel) ...	353
10.3.3	Choropleth Map of Europe at Country-Level	358
10.3.4	Choropleth Map of Europe at Country-Level (Panel).....	360
10.3.5	World Choropleth Map: Regions	364
10.4	Exceptions and Special Cases	366
10.4.1	World Map with Orthodromes	366
10.4.2	City Maps with OpenStreetMap Data (Panel).....	369
10.4.3	Georeferenced Map in Grid Format	373
10.4.4	Cartogram (Panel)	380
11	Illustrative Examples	385
11.1	Table with Symbols of the “Symbol Signs” Type Face.....	385
11.2	Polar Area Charts with Labels (Panel)	387
11.3	Polar Area Charts Without Labels (Panel)	394
11.4	Polar Area Chart (Poster)	397
11.5	Nighttime Map of Germany as Scatter Plot	401
11.6	Scatter Plot Gapminder	403
11.7	Map of Napoleon’s Russian Campaign in 1812/13, by Charles Joseph Minard, 1869	408

12 Interactive Visualisation with JavaScript: Highcharts and Mapael ..	413
12.1 Scatter Plot in Highcharts	414
12.2 Time Series in Highcharts	423
12.3 Choropleth Maps with Mapael	429
Appendix	447
A Data	447
B Bibliography	449