Journal of Statistical Software

November 2015, Volume 68, Book Review 4. *doi: 10.18637/jss.v068.b04*

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

Using R and RStudio for Data Management, Statistical Analysis and Graphics (2nd Edition)

Nicholas J. Horton and Ken Kleinman
CRC Press, Boca Raton, 2015.
ISBN 978-1-4822-3736-8. 313 pp. USD 59.46 (H).
http://www.amherst.edu/~nhorton/r2/

This book is the second edition of "R for Data Management, Statistical Analysis and Graphics", which was quite well-received as a reference book for working with R at the time (see, e.g., Geissler 2011). The second edition has been substantially restructured and extended, which is already reflected in the addition of **RStudio** to the title. The twelve chapters cover the three topics mentioned in the book title: Chapter 1 discusses "Data input and output" (9 pages) and Chapter 2 "Data management" (21 pages) in a more narrow sense. Chapter 3 ("Statistical and mathematical functions", 11 pages) and Chapter 4 ("Programming and operating system interface", 6 pages) provide some technical prerequisites in preparation for the statistical analysis chapters. The "Statistical analysis" part of the book consists of Chapter 5 ("Common statistical procedures", 16 pages), Chapter 6 ("Linear regression and ANOVA", 23 pages) and Chapter 7 ("Regression generalizations and modeling", 32 pages). The graphical content is delivered in Chapter 8 ("A graphical compendium", 21 pages) and Chapter 9 ("Graphical options and configuration", 21 pages). Three further chapters, Chapter 10 ("Simulation", 11 pages), Chapter 11 ("Special topics", 20 pages) and Chapter 12 ("Case studies", 24 pages), discuss diverse topics. Four appendices give an introduction to R and **RStudio** (Appendix A, an extension of the former Chapter 1) and to the most important example data set used throughout the book (Appendix B), provide references (Appendix C) and two indexes (Appendix D). The book is accompanied by a website with example code, example data and a convenient command for installing the packages needed for running the entire example code; an errata list is also provided.

In comparison to the first edition, the one wide-ranging chapter on data management has been split into Chapters 1 to 4, only two of which I would consider data management, as listed above. The chapters on statistical analysis are quite similar to their counterparts in the first edition, the methods described range from simple descriptive methods to advanced regression modeling techniques; only the advanced chapter (Chapter 7) has been restructured and slightly extended. The previous "Graphics" chapter has been split into Chapters 8 and 9, and the graphical compendium given in Chapter 8 has been extensively restructured, much to its benefit. The previous chapter “Advanced applications” has been split into three parts (Chapters 10 to 12), restructured and somewhat extended. While modernization was likely the motive for creating a new edition, the authors have also used the opportunity for restructuring the book in other respects, almost always for the better. Thus, the very positive overall evaluations of the first edition remain valid for the second edition, in spite of a few less positive aspects I detail below.

The preface clearly states that the book is not intended as an introductory text to be read from beginning to end, but should be used as a reference text. The authors hope to “bolster the analytic abilities of a relatively new user” with this reference approach, and they support easy access to the content with the two indexes (topical and R commands), like in the first edition. The primary target readers are new R users, who have “an understanding of statistics at the level of multiple-regression analysis”. According to the book authors, sophisticated users of other statistical software might specifically benefit from the book; this ties in with the fact that Horton and Kleinman also wrote “SAS and R” (Kleinman and Horton 2014). Experienced R users are referred to the case studies; these are most likely of benefit to almost all readers. With large systems like R, the experience of an experienced user is of course also limited to some parts of the language, so that most experienced users will also benefit from other parts of the book. I found the introductory Appendix A offers an interesting addition: the swirl system (`swirlstats.com`), which can be used for teaching R programming and data science within the R console. The authors confess to have made no attempt at completeness or elegance, in favor of keeping things simple for new users.

It is a bit strange that the preface of the first edition was adapted in some places to changes in numbering (e.g. Appendix B instead of A for the example data) but not in one very important place: the instructions on how to use the book (which are given only in that preface) still recommend that new users should start with the first chapter, even though the content of the previously first chapter was moved to Appendix A. I suppose that it might have been more transparent for readers to keep the preface to the first edition unchanged except for visibly new insertions on numbering changes.

I started inspection of the book by looking at its use of R packages, which can be easily done via the R command index that references packages as `library(pkgname)`: the book uses a few R packages a lot more than others; in particular **dplyr** (“next iteration” of **plyr** according to its author), **lattice**, **MASS**, **mosaic** and **survival** are used quite frequently. Package **mosaic** depends on **car** and **ggplot2**, which are also frequently used. Package use was adapted vs. the first edition, for which **mosaic** and **dplyr** were not yet available. Packages **dplyr** and **ggplot2** are instances of the “grammar of” approach, which is not to my liking. Starting out reading the book in detail after this preliminary package survey, I was curious whether I would come to appreciate that “grammar of” approach for some purposes. Actually, I remain a fan of doing things with base R commands, wherever easily possible; this is of course a matter of taste. I cannot resist providing simplified base R code for the calculation of group wise regression parameters (`params` in Chapter 11.1.2), which the book does with 14 lines of code (pp. 168/169): The code from lines 3 to 12 can be replaced by the following much shorter solution, which is very readable, once the concept of the `sapply` function is understood (`uniquevals` contains the unique levels of the grouping factor):

```
R> params <- sapply(uniquevals, function(obj)
+    coef(lm(i1 ~ age, subset = (female == obj), data = ds)))
```

In a similar spirit, even the subsequent code on p. 169 that uses package **plyr** for group wise processing can be replaced by only slightly longer and no more complicated base R code (not shown). This is an example of the benefits obtainable by efficient use of base R, which the book does not emphasize as much as I would like.

Based on the book's title, I had hoped to learn something substantial about **RStudio**. When starting to use **RStudio** in 2013, I often ended up with a vast number of open windows and loss of control over many parallel work streams, because I did not go about its usage in a structured way. Therefore, and because I find work with **RStudio** painful on small screens, I reverted to my previous work flow with the R GUI or the **TinnR** editor and stored scripts. Nevertheless, working within the **RStudio** environment makes many tasks much more comfortable. Unfortunately, the book contributes nothing to support successful incorporation of **RStudio** into an efficient work flow; rather, it relies on **RStudio** being more or less self-explanatory (which is true in a way). I was disappointed about this lack of thorough treatment, as I am convinced that there would have been a potentially huge benefit from thoughtful explanations on implementing an efficient work flow including **RStudio**.

Another aspect of modern R: I had no experience with the package **markdown** and had hoped to learn how to use it. The book explains its use using a sample markdown (`Rmd`) file that can be automatically created from within **RStudio**; in doing so one has to decide whether one wants to create HTML, PDF or Word output, and the book explains creation of an HTML file. On the computer in our university lab, the book's example works exactly as described. It took me quite a while to figure out which crucial step I was missing when trying to run the same example on my laptop (same OS and locale settings as in the lab): I did not store the automatically created `Rmd` file before using the "Knit html" button; apparently, this prevented the output from showing; after storing the file, everything worked as intended also on the laptop. It seems that, the more interactive a system becomes, the more the details of the necessary steps may depend on details of particular computer systems. (By the way, the book's figure that shows the produced output is incomplete: it does not show the plot, maybe for similar reasons.) While the book's instructions for creating HTML files through **markdown** are helpful, they did not suffice to make me succeed in creating a pdf file through **markdown** despite a reasonable amount of effort.

The "Special topics" chapter is a diverse collection of topics, ranging from processing by group over simulation-based power calculations or reproducible analysis and output (e.g., **markdown**) to "Advanced statistical methods". Of course, with about 300 pages, the book cannot cover all kinds of statistical methods; thus, the "Advanced statistical methods" section assumes that the reader is already familiar with the methods and wants to look up how to implement them in R. Also, completeness is not intended and cannot be expected; for example, the "missing data" section (11.4.4) presents multiple imputation with the **mice** package for chained equations and does not mention any further possibilities, like the `aregImpute` function from the **Hmisc** package, which would be an interesting competitor.

The "Case studies" chapter is again a diverse collection, covering data management and related tasks, reading variable format file data, plotting maps, data scraping, text mining, interactive visualization, manipulating bigger datasets, and constrained optimization (of the knapsack problem). I picked the "Plotting maps" and the "Interactive visualization" sections for closer looks; both deal with the creation of maps, which I have recently studied a bit closer when giving a class on exploratory data analysis. The book covers creation of maps not only in the "Case studies" section but also in the graphics chapter, where function `ggmap` of package

ggmap is used for creating a choropleth map. The first case study in Section 12.3 shows visualization of the boundaries of Massachusetts counties (data from Nick Horton's website, given in ASCII format, read in Section 12.2) via the base `plot` function, labeling counties with the base `text` function (Section 12.3.1). This is the only mapping case study that was already part of the first edition. The second mapping case study maps bike ride data from Nick Horton's website onto the background of a map downloaded from Google Maps via the `get_map` function of **ggmap**; plotting is again done with the `ggmap` function (Section 12.3.2). The third mapping case study creates a choropleth map of the number of murders per 100,000 inhabitants from the well-known `USArrests` data that ship with R (Section 12.3.3). Here, function `map_data` from **ggplot2** is used for extracting the US state boundaries map from package **maps**, the extracted dataframe (that contains x and y coordinates for polygons for each state, on top of each other) is merged with the `USArrests` data, and a choropleth map of the resulting file is created with function `ggplot`. This third example – implemented differently – is also used for the case study on interactive visualization with **shiny** (see below). It is helpful to see several possibilities of obtaining maps implemented. I was a bit disappointed that the mapping examples did not touch upon data structures for maps; in times of open data, maps from official bodies often come in a widespread proprietary format, e.g., as Esri shape files; for example, the Massachusetts county boundaries are also available in this form. It would have been nice to inform readers about possibilities to read these into R spatial objects, after unzipping them, by using, e.g., functionality from package **maptools**, **rgdal** or the recent package **tmap**. Of course, there is no way to cover the detail of all such functionality in a brief book chapter, but a few pointers would have been very welcome. After reading such data, the resulting spatial objects (e.g., of S4 class `SpatialPolygonsDataFrame`) can, e.g., be processed with tools from package **sp**, or transformed into data frames of the nature used in the **ggplot2** universe by using the `fortify` function.

The book implicitly left me under the impression that maps cannot be created with traditional R tools in reasonable quality with reasonable effort. I include the following example in order to show how a choropleth map for the murder data can be created, starting from a shape file (downloaded 2015-10-12 at `http://www2.census.gov/geo/tiger/GENZ2014/shp/cb_2014_us_state_500k.zip`; status 2014, resolution 1:500,000), using base graphics with the help of package **sp**. The shape file, unzipped into sub directory “`USStates2014`" of the working directory, is read into a `SpatialPolygonsDataFrame` object by function `readOGR` from package **rgdal**; the `merge` method from package **sp** merges this object with the `USArrests` data frame, whose “Murder" variable has been categorized into quintiles (Alaska and Hawaii are omitted; option `by.y = 0` for merging on row names requires at least **sp** version 1.2-1). Beware that merging always requires great care; this is of course true for all versions of matching geometric shapes with corresponding data rows, not only for the **sp** implementation. The `spplot` function (based on **lattice** graphics) could be used for plotting; more basic, the code below uses the `plot` function provided by package **sp**, which has the charm that the map can be enhanced by the numerous possibilities that are offered in package **graphics**. Function `brewer.pal` creates a palette of five suitable blue values. The `col` option defines the fill colors, the `text` statement adds state labels, and the `legend` statement adds a legend to the bottom of the figure. Figure 1 shows the resulting map; improvements are of course possible, e.g. regarding the placement of labels for small states.

```
R> library("rgdal")
R> library("RColorBrewer")
```

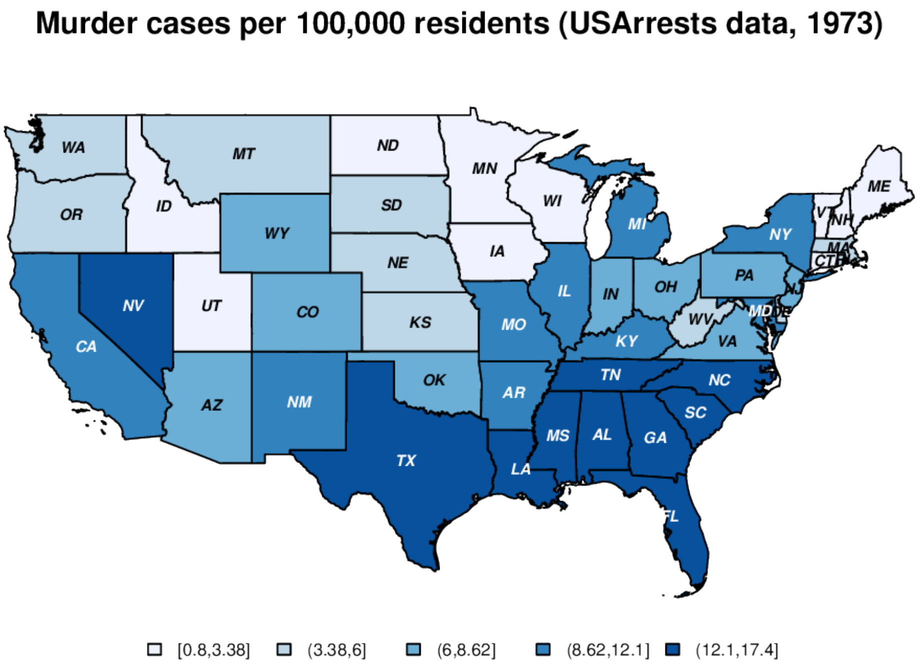

Figure 1: Map created with function `plot` from package **sp** and base graphics enhancements.

```
R> USStates <- readOGR("USStates2014", layer = "cb_2014_us_state_500k")
R> USArrests$murder_cat <- cut(USArrests$Murder,
+    quantile(USArrests$Murder, (0:5)/5), include.lowest = TRUE)
R> mapdat <- merge(USStates, USArrests[-c(2, 11), ],
+    by.x = "NAME", by.y = 0, all.x = FALSE)
R> blues <- brewer.pal(5, "Blues")
R> plot(mapdat, col = blues[mapdat$murder_cat])
R> title("Murder cases per 100,000 residents", line = 0)
R> text(coordinates(mapdat), as.character(mapdat$STUSPS), cex = 0.6,
+    col = c(rep("black", 3), "white", "white")[mapdat$murder_cat], font = 4)
R> legend("bottom", fill = blues, legend = levels(mapdat$murder_cat),
+     cex = 0.7, horiz = TRUE, bty = "n")
```

The case study on interactive visualization implements a choropleth map of the murder data in **shiny**, allowing to vary the number of levels and to toggle between showing and omitting the state labels. This **shiny** app is incorporated in a markdown file which can be created with **RStudio**; the alternative classic way of creating a **shiny** app is also reported (provide a "`ui.R`" and a "`server.R`" file, and possibly a file "`app.R`"). Package **choroplethr** is used for implementing the map; that package handles maps of the countries of the world or various levels of US administrative regions only; for this limited set of maps, it takes care of merging the data to the regions by built-in region-specific functions. Unfortunately, the book's code for this example does not work any more, since the package has undergone changes that invalidated previously valid code. Here, the book's website is helpful: it provides modified code for both versions of the **shiny** app. The **shiny** example is quite instructive, even though it would have suggested itself to broaden the scope of control elements presented, by using a `numericInput` or `sliderInput` control instead of the less suitable `selectInput` control for the number of levels. As readers can find a lot of material on the internet on how to do **shiny** apps, the entire scope of controls will be easily accessible in spite of this omission. It is less easy to find instructions on providing web access to **shiny** apps, locally or over the internet; some pointers regarding this topic would have been helpful.

In a few places, the book could cater better for international readers. For example, it would not have hurt to mention the function `read.csv2` along with `read.csv` for the European format with decimal comma and semicolon as the separator (Section 1.1.4), and it would have helped to mention that the time format used in `as.POSIXlt` is highly dependent on system and locale – I only got it to work after using `Sys.setlocale("LC_TIME", "US")` on my Windows machine (section on data scraping). Nevertheless, the book caters very well to users' needs by making it exceptionally easy to look up ways to solve a task and by providing a very useful website. Also, it is a nice feature that the authors ensure for reproducibility of important examples by providing downloadable snap-shots of websites needed for them (not directly accessible via the book's website, but through links provided in the book's text), and by providing updated code in case a software update causes the book's code to fail.

In summary, the second edition of the book preserves the many good points of the first, and makes some improvements to the structure, e.g., on the graphical compendium. It also contains added material on more recent possibilities, which would, however, benefit from a little more depth. I think that owners of the first edition need not upgrade to the second. For those who do not own the first edition, the second edition is a good buy, if the goal is to have a reference book which allows to quickly find a way of accomplishing a task at hand in R, be it with or without **RStudio**. R beginners who want to become serious R programmers should additionally use more traditional introductory material, especially for building proficiency regarding the many possibilities offered by base R.

References

Geissler PH (2011). "Reviews of Books and Teaching Materials – Using R for Data Management, Statistical Analysis, and Graphics." *The American Statistician*, **65**(4), 296. doi:10.1198/tast.2011.br654.

Kleinman K, Horton NJ (2014). *SAS and R: Data Management, Statistical Analysis, and Graphics*. 2nd edition. CRC Press, Boca Raton.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin
Department II

D-13353 Berlin
E-mail: groemping@bht-berlin.de
URL: http://prof.beuth-hochschule.de/groemping/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2015, Volume 68, Book Review 4 *Published:* 2015-11-23
doi:10.18637/jss.v068.b04