Journal of Statistical Software

January 2016, Volume 69, Book Review 2. *doi: 10.18637/jss.v069.b02*

Reviewer: Jose M. Pavia
Universitat de Valencia

Humanities Data in R

Taylor Arnold, Lauren Tilton
Springer-Verlag, Cham, 2015.
ISBN 978-3-319-20701-8. 211 pp. USD 69.99 (H), USD 49.99 (E).
http://www.springer.com/9783319207018

Humanities Data in R. Exploring Networks, Geospatial Data, Images, and Text is one of the latest additions to the *Quantitative Methods in the Humanities and Social Sciences* series from Springer. This series tries to foster the dialogue between scholars of humanities and social sciences and quantitative scientists, more involved in applications of computational analysis, statistical models and computer-based programs. The book perfectly suits the series. It is a canonical example of the kind of ties that quantitative and qualitative methods are setting up through the interdisciplinary collaboration between statisticians and scholars in the humanities. Indeed, the authorship of *Humanities Data in R* is a vivid example of the fruitful bridges that can be established among researches with different backgrounds. The authors, Taylor Arnold (Senior Scientist at AT&T Labs Research and Lecturer of Statistics at Yale University) and Lauren Tilton (a doctoral candidate in American Studies at Yale University), are both co-directors of Photogrammar (http://photogrammar.yale.edu/), "a web-based platform for organizing, searching, and visualizing the 170,000 photographs from 1935 to 1945 created by the United States Farm Security Administration and Office of War Information (FSA-OWI)."

As a social scientist with a statistical and methodological background, I was interested in seeing what this book has to offer. The book is not a detailed manual for using the various R packages introduced nor does it provide an exhaustive coverage of the conceptual and technical foundations of the topics that it presents. It just offers a practical introduction to quickly explore and visualize four basic types of data structures commonly encountered in digital humanities: networks, geospatial data, images and text. The result is a really valuable "single text that brings together several disparate methodologies", none of which taught in introductory textbooks in statistics, that humanities scholars (including archaeologists, historians, anthropologists, political scientists, philologists, classicists, and new media specialists) may integrate into their own work with a minimal computational training. Indeed, the book is especially devoted to this readership. It assumes no prior exposure to R, uses instructive examples to provide a user-friendly guide of the key concepts in the digital humanities, and offers valuable instances of code snippets that might help scholars in the humanities and social researches to quickly produce novel applications with their own data.

Despite the above comments, the pace at which the text advances will likely represent a substantial challenge for R beginners without external help. Indeed, as the authors state in the preface, the book is aimed at "students in a one- or two semester introductory course on digital methods in the humanities" and also at "intermediate users looking for a self-study text to solidify and extend their basic working knowledge of both computational methods and R". In my opinion, the book offers an opportunity for both humanities scholars and statisticians. On the one hand, less quantitatively minded social scientists can discover what statistics and statistical software can offer to them. On the other hand, statisticians and computer scientists can find new fields in which to apply their knowledge. They all can look for new ways of collaborative work, opening new bridges between quantitative and qualitative methods.

The book is divided into two parts, each with five chapters, and an appendix, with three additional chapters. The early five chapters move readers from installing R to exploratory data analysis. After this, each topic covered (networks, geospatial data, image data, natural language processing and text analysis) has its own chapter. For easy assimilation and comprehension, each chapter is grounded in examples that can be completely reproduced by the reader using the data and code available on the book website (http://humanitiesdata.org/). Indeed, some errata that remain in the code available in the book are corrected in the R-files downloadable from the website. For those R beginners interested in reproducing the examples as they advance reading, it is advisable to first visit Chapter 11 (the first chapter of the appendix). There, they will find instructions for installing the R packages and also the third-party libraries needed for reproducing the examples.

Chapter 1, with only four pages (including references), offers a brief introduction to the structure of the book and succinctly explains (i) how to install R and **RStudio**, (ii) how to access the accompanying materials of the book, and (iii) how to get help in R. Chapter 2 provides a straightforward and very basic introduction to the R language and how to interact with it. It introduces some basic notions and concepts of R. The ideas of objects and classes are shown, and basic computations on vectors, matrices, and data frames displayed. Different forms of subsetting and some basic functions to work out with character vectors are also presented. The chapter also contains basic instructions on how to load external data into R and on how to save R objects in the computer's file system.

The next three chapters, Chapters 3-5, are devoted to introducing tools of exploratory data analysis (EDA), with a special emphasis on graphical methods. In Chapter 3 data from the US American Community Survey, corresponding to the state of Oregon, are used to illustrate some basic methods for univariate data. How to calculate tables and quantiles, how to plot a histogram or how to transform a continuous variable into a categorical one (a process known as binning) are some of the topics covered in this chapter. Control flow using *for loops* is also introduced.

Chapter 4 reuses the data from Chapter 3 and extends the analysis by introducing tools of EDA for multiple variables. In particular, using scatter plots as basic building blocks, the techniques for exploring the relationship between various continuous variables are shown. In the process, the authors present additional methods for producing more aesthetic graphics and for their customization by, for instance, varying color, opacity, shape or sizes of points.

Chapter 5 goes deeper into the process of increasing the aesthetic quality, professionality, and usability of R graphics. This time some data from the French 2012 Presidential Election are used to illustrate the different procedures. Methods to save the graphics as external files, to build (sequential, divergent and categorical) color palettes and to include legends are presented. Random number generators are also introduced.

The second half of the book, Chapters 6–10, is the core of the book. Here, the authors introduce some key areas of investigation for humanistic data. Each chapter deals with a different type of analysis (with the exception of Chapters 9 and 10, both dedicated to text data) and shows through examples how it can be applied to the kind of information usually available in humanities and social sciences. As the territory to cover is enormously large, the text just hints at possibilities, and each chapter closes with suggestions for further extensions and some exercises to absorb the topics presented.

Chapter 6 (14 pages) introduces the concept of a network (also known as a graph) and shows how to plot and customize it using the R package **igraph**. Although several other R packages also provide these functionalities, Arnold and Tilton ground their selection in the fact that the **igraph** library is relatively simple to use and in the circumstance that there are also versions of the package in other languages, such as Python, Ruby, and C++ which facilitates collaboration. To introduce readers to the concept of networks and their construction in R, the authors adopt a really didactic tone showing how easily a family tree graph of the British royal family can be constructed. Once the basic issues have been understood, the rest of the chapter handles a citation network built from United States Supreme Court opinions. Through this example, the authors introduce useful concepts, like centrality and community, and show how to plot them. The chapter offers a general overview of more advanced topics of graph drawing and (making intensive use of some features, such as subsetting or color palettes, introduced in previous chapters) displays how to paint in the graph the most influential vertices (nodes) or the clusters (communities) detected.

Chapter 7 deals with geospatial data. With only 16 pages devoted to this vast issue (Bivand, Pebesma, and Gomez-Rubio 2013), the authors are able to show some of the functionalities that geospatial data can offer to social science scholars. Using a data set of major border crossings into West Berlin between 1961 and 1989, the authors show how to flag our points of interest in a Google-like map using the `osmap` function of the **snippets** library. Following this, the chapter turns to consider some of the types of data, problems and issues of presentation that arise more frequently when dealing with spatial data. In this part some of the more known packages for handling vectorized and points geospatial data are introduced: **sp**, **maptools**, **rgdal**, and **rgeos**. Using the shapefile of the geospatial vector data defining the shape of all the US states, the authors show how to plot an empty map, how to convert between coordinate system projections, and how to paint our data in a map. The rest of the chapter is devoted to the process of merging both tabular data with geospatial data and geospatial data with tabular data as a way to create enriched datasets for further analysis. This time the data used to illustrate the possibilities come from the above mentioned Photogrammar platform. They use the approximate latitude and longitude where each photograph is believed to have been taken to join county census data to the Photogrammar dataset and, conversely, to count the number of photographs that were taken inside each county.

Chapter 8 (17 pages) shows methods for loading, manipulating, and saving image files using R. In a really simple way, the authors first present (using a portrait of van Gogh) the structure of the objects containing image data in R and explain how to manipulate them. The packages introduced in this chapter are **jpeg**, as an example of library to load digital images, and **abind**. The rest of the chapter is dedicated to show how these methods can be applied to separate into two groups a collection of photographs. In particular, they consider as a running example a corpus of 311 outdoor photographs available in the Digital Video Multimedia Lab at Columbia University and show some procedures to discriminate between those photos taken during the day from those taken at night. In the process, Arnold and Tilton illustrate how to easily change, using the `rgb2hsv` function, the representation of an image in the (red, green, blue) coordinate system to the most orthogonal (hue, saturation, value) cylindrical coordinate system as a mechanism to facilitate the separation process. As an alternative, the text shows how a similar goal, usually accompanied by a dimension reduction, can also be reached using the well-known procedure of principal component analysis. As a procedure to clustering, the chapter chooses k-means. The chapter finishes with a really interesting code snippet for displaying a collection of images in a scatter plot, with the dots replaced with small thumbnail versions of the images themselves.

The next two chapters are devoted to text analysis; probably the type of data more traditionally related to humanities. These two chapters are by far the most complex chapters of the book, partly because textual data are highly unstructured. This issue implies a large amount of pre-processing before more high-level procedures can be applied. Chapter 9, the longest chapter of the book with 25 pages, focuses on pre-processing. It demonstrates several procedures of low-level natural language processing (NLP) to clean and organize the data. In particular, the methods explained in this chapter are based on the package **coreNLP** developed by the authors to call the methods available in the java library Stanford **CoreNLP**. The topics covered in this chapter include tokenization and sentence splitting, lemmatization and part of speech tagging, named entity recognition, and coreference detection. As in the rest of the book, real data are used to illustrate the different issues. The story "A Scandal in Bohemia" by Sir Arthur Conan Doyle is chosen this time as raw data to show the processing sequence. Finally, as an appetizer of high-level analyses, the methods are applied to the corpus of short stories by Sir Arthur Conan Doyle featuring his famous detective Sherlock Holmes as a way to classify stories based on the number of times each main character appears and on the moment in the story where it appears.

In Chapter 10 (20 pages), the true power of textual analysis becomes clear. In this chapter, Arnold and Tilton show, adding to the **coreNLP** package the capabilities of the **mallet** package, how to apply various high-level analyses to pre-processed data. First, using a corpus of 179 Wikipedia articles regarding philosophers from the sixteenth to the twentieth centuries, they illustrate several methods for extracting meaning from each document in the context of the collection. They determine the relative importance of the most frequently occurring lemmas (meaningful words, terms) to a given philosopher as a way to (i) distinguish a given article from the body of the other documents and to (ii) construct topic models, from which it is possible to explore, visualize, interpret, group, catalog, formulate hypotheses and discover new features of the collection of documents. As a second instance, and with the aim of exemplifying how to characterize writing styles (a discipline called *stylometrics*), a collection of 26 novels from 4 authors (Mark Twain, Charles Dickens, Nathaniel Hawthorne, and Sir Arthur Conan Doyle) is subsequently analyzed by looking at speech bigrams and word frequencies.

The book closes with an Appendix divided into three chapters. Chapter 11 is dedicated to software issues, such as installing the R packages. Chapter 12 poses 100 short programming questions related to Chapters 2 to 5. And Chapter 13 provides example solutions to the question proposed in Chapter 12.

In summary, *Humanities Data in R* offers a really accessible bridge between digital humanities and its implementation in R. Hence, if you are considering leading a course on digital humanities or starting to work with digital humanistic data, this is your textbook. The book is well-written, the information is arranged in a logical manner and it is easy reading. It will provide you with a broader idea of the forms data may take and will introduce you to new ways to explore and visualize information. This book is also a bridge between quantitative and qualitative methods and suggests new opportunities for collaborative work between humanities scholars and statisticians.

References

Bivand RS, Pebesma E, Gomez-Rubio V (2013). *Applied Spatial Data Analysis with R.* 2nd edition. Springer-Verlag, Nueva York. URL http://www.asdar-book.org/.

Reviewer:

Jose M. Pavía
Universitat de Valencia
Department of Applied Economics
Valencia, Spain 46022
E-mail: pavia@uv.es
URL: http://www.epo-uv.es/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2016, Volume 69, Book Review 2 *Published:* 2016-01-29
doi:10.18637/jss.v069.b02