Open research methods in computational social sciences and humanities: introducing R

Markus Kainu

2017-01-29 16:24:11

Introduction - Open Research Methods

The debate on open science in the context of Social Sciences and Humanities (SSH) has been predominantly focusing on open access to research publication and opening up the various types of digital research data (open research data). The openness of research methods has received a less attention.

I can think of two main reasons for that. On the one hand, research methods in SSH have predominantly been qualitative where software has played only a supporting role. Such research methods, let's take *discourse analysis*, have always been open, free to use and to modify and redistribute. On the other hand, the quantitative fields of SSH have mostly used statistics or survey and register data, or other, often closed, tailor-made data that custom proprietary data analysis tools such as SPSS, Stata or Excel are well suited for. However, the future of SHH looks somewhat different as the quantity and multiplicity of sources of digital data are challeging both traditional approaches in SSH, the purely qualitative approach and custom tools approach in quantitative analysis. The future that Gary King [2014] of Harvard describes as:

> An important driver of the change sweeping the field is the enormous quantities of highly informative data inundating almost every area we study. In the last half-century, the information base of social science research has primarily come from three sources: survey research, end-of-period government statistics, and one-off studies of particular people, places, or events. In the next half-century, these sources will still be used and improved, but the number and diversity of other sources of information are increasing exponentially and are already many orders of magnitude more informative than ever before.

In the data rich future of SSH research, as the role of software and computation becomes more central, the questions of licensing, ownership, modification and distribution of that software will become increasingly important. This chapter will introduce one viable option for analysing your data called R.

What is R?

R is one of the most popular platforms for data analysis and visualization currently available. R is distributed under the terms of the *GNU General Public License* so it is free and open source and it can be distributed under those conditions. R is available from Comprehensive R Archive Network (CRAN). The name R comes from the first names of two New Zealand statisticians Ross Ihaka and Robert Gentleman who created the language in late 1990's.

R can be regarded as an implementation of the S language which was developed at Bell Laboratories in 70's by Rick Becker, John Chambers and Allan Wilks [Venables et al., 2013]. R is an object-oriented programming language which means that unlike in SPSS or SAS that give you abundant information on particular model you implement, R only creates a *fit object* in memory that can be used in subsequent analysis. This structure of R directs the user to implement the data-analysis as stepwise process which becomes very useful later on when solving complex research problems using vast and messy data typical for emerging computational SSH research.

R user-interfaces

R runs in Windows, Mac OS X and GNU/Linux operating systems on local computer, but a different server implementations are becoming increasingly popular, such as R-Fiddle or rnotebook. The most basic user interface for R is *console*, which allows user to type in commands and outputs the results of the analysis. If the results is a plot a pop-up graphical windows is opened. There are several graphicsl user interfaces (GUI) in R that may be helpful in the beginning, like RCommander or Deducer. Perhaps the most productive way for using R is through a integrated development environment (IDE) that provide the user, in addition to console, several useful functionalities for controlling the whole research project. RStudio has gained a lot of popularity in last couple of years and is also my personal favourite IDE. It combines the console with script editor, plot browser, file browser and environment window. If the user uses plain text (latex or markdown) for typesetting the texts, RStudio has a tailored text editor and support for version control either in git or in subversion. In addition, RStudio has native support for html-based presentation graphics using reveal.js-framework. All these operations makes it possible to squeeze the whole research process within a single software environments from planning to publishing. Rstudio can also be run on a remote server through web browser. The RStudio company has another exciting open source tool for R called *shiny* that can be used for creating interactive web applications such as this experimental gadget of mine.

Structure of R-project

For someone new to R, the peculiar structure of the language creates a very steep learning curve. Same applies to learning how the whole project is organised.

Official name The R Project for Statistical Computing refers both to the centrally maintained core as well as R's distributed structure of contributed extensions, called *packages*. Packages in R are collections of functions and/or data that are packaged for conviniency. Installing a package broadens your the functionality of your R installation. Basic R installation consists of so called *base installation* that includes the core with some 25 additional *packages* for the most basic functionality. The core of the language is maintained by *R Development Core Team*, but the additional packages are developed and maintained by individual developers and research institutes. R users often create packages for themselves, but if one thinks the package could be useful for other users too, the packages can be distributed through repositories.

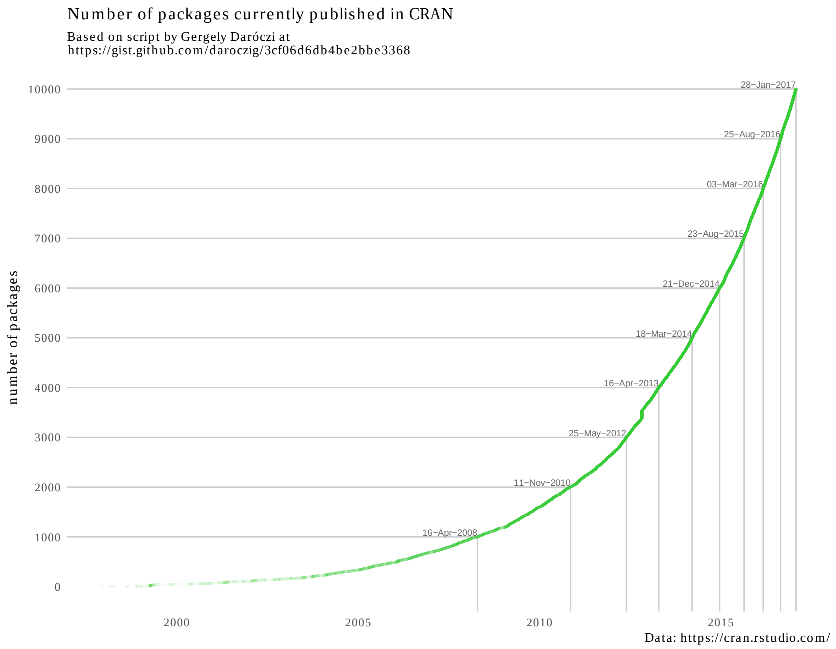
CRAN is the "official" repository for contributed packages and currently hosts 5150 packages that can be used to extend R. In the last couple of years various code hosting sites such as GitHub have become increasingly important resources especially for collaborative development of new packages. Github hosts currently roughly 1500 packages for R. Bioconductor is another separate package repository, but can be regarded as *domain spesific* for it hosts packages for *the analysis and comprehension of high-throughput genomic data*. Other such domain spesific projects are for example rOpenSci and emerging rOpenGov that provide tools for open science and open government data, respectively.

```
library(XML)
library(tidyverse)
library(extrafont)
loadfonts()
if (!file.exists("./data/packages.RDS")) {
    current <- readHTMLTable(readLines("https://cran.rstudio.com/src/contrib"),
        which = 1, stringsAsFactors = FALSE)
    names(current)[c(1:3)] <- c("col1", "name",
        "date")
    pkgs <- current %>% filter(Size != "-", grepl("tar.gz$",
        name)) %>% mutate(name = sub("^([a-zA-Z0-9\\.]*).*",
        "\\1", name)) %>% select(name, date)

    packages <- data_frame()
    for (i in 1:nrow(pkgs)) {
        path <- paste0("https://cran.rstudio.com/src/contrib/Archive/",
            pkgs$name[i])
        if (httr::GET(path)$status == 200) {
            tbl <- readHTMLTable(readLines(path),
                which = 1, stringsAsFactors = FALSE)
            names(tbl)[2:3] <- c("name", "date")
            new_row <- tbl[3, c("name", "date")]
            new_row$archived <- TRUE
        } else {
            new_row <- data_frame(name = pkgs$name[i],
                date = pkgs$date[i], archived = FALSE)
        }
        packages <- bind_rows(packages, new_row)
    }
    saveRDS(packages, file = "./data/packages.RDS")
} else packages <- readRDS("./data/packages.RDS")
packages <- packages %>% mutate(date = as.POSIXct(date,
    format = "%d-%b-%Y %H:%M")) %>% arrange(date) %>%
    mutate(rank = 1:nrow(.))

vdat <- packages %>% filter(rank %in% seq(0, 10000,
    by = 1000))

p <- ggplot(data = packages, aes(x = date, y = rank))
p <- p + geom_segment(data = vdat, aes(xend = as.POSIXct(-Inf,
    origin = "1970-01-01"), yend = rank), color = "grey80")
p <- p + geom_segment(data = vdat, aes(xend = date,
    yend = -Inf), color = "grey80")
p <- p + geom_point(alpha = 0.1, shape = 1, size = 0.5,
    color = "limegreen")
p <- p + geom_text(data = vdat, aes(x = date,
    y = rank, label = format(date, format = "%d-%b-%Y")),
    hjust = 1, vjust = -0.2, size = 2.5, color = "dim grey")
p <- p + theme_minimal() + theme(text = element_text(family = "Open Sans"))
p <- p + theme(panel.grid.major = element_blank(),
    panel.grid.minor = element_blank())
p <- p + scale_y_continuous(breaks = seq(0, 10000,
    by = 1000))
p <- p + labs(title = "Number of packages currently published in CRAN",
    subtitle = "Based on script by Gergely Daróczi at
https://gist.github.com/daroczig/3cf06d6db4be2bbe3368",
    caption = "Data: https://cran.rstudio.com/",
    y = "number of packages", x = NULL)
p
```



Learning the language

As the internet has brought together the vast community around R, the internet has become the main channel for delivering instructions for R. The official *Introduction to R* by Venables et al. [2013] is an important document to master when getting into the language. Besides this general introduction R-project has also a domain spesific structure where you can start learning from so called task views. For SSH researcher the social sciences and natural language processing task views are good places to begin with.

Discussions and announcement on R happen mainly through R official mailing lists that have their own lists for development and user help. R help is the main list for general help and receives tens of mails per day. Most of the individual packages have their own mailing list for development where anyone can join if wanting to contribute to the packages.

The official mailing lists have recently been challenged by so called Question & Answer -sites like Stack Overflow in delivering solutions for one-off user questions. Stack Overflow has currently almost 47000 questions tagged with R. In comparison to proprietary software, there are 2014 questions tagged with SAS, 616 with Stata and 362 SPSS. These figures are used as one indicator of the increasing popularity of R. Besides the Question & Answer sites, there are hundreds of blogs discussing specific analytical problem using R and feeds from the blogs are aggregated in R-bloggers-website.

Another, more formal channel for distributing and communicating R have become the so called massive open online courses (MOOC). MOOCs seem to work well for teaching programming and many courses in Coursera and EdX have become hugely popular attracting tens of thousands of students each year. The free licensing of R has made it primary language on these coursesas it is basicly the only viable alternative for teaching statistical programming for massive crowds.

Aside with vibrant internet community more and more books are being published on R. Books can be put in three categories. First are the general introductions to statistics using R. *Discovering Statistics Using R* by Field et al. [2012] and *R in Action: Data Analysis and Graphics With R* by Robert Kabacoff [2013] are popular examples of that category. Second there are more and more books addressing how to solve some specific analytical problems using R. A prime examples of books in this category are *Complex Surveys: A Guide to Analysis Using R* by Thomas Lumley [2011], *Text Analysis with R for Students of Literature* by Matthew Jockers [forthcoming], *R Graphics Cookbook* by Chang [2012] and *Dynamic documents with R and knitr* by Yihui Xie [2014]. A third category are the books that focus on specific theoretical issue in statistics and use R as a primary language to demonstrate this. Such books are for instance *Bayesian Data Analysis* by Gelman et al. [2013] or *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* by Snijders and Bosker [2011].

Use of R language

Throughout it's existence the main use of R has been implementation of new statistical methods. This is still the case and implementations of new statistical methods are usually first available in R. However, various fields of applied statistics have become more active as researchers across disciplines have started to migrate into R. *Bioconductor* was already mentioned as an example of a domain spesific initiative to apply R in their analysis, for genome data in this case. Natural sciences in general have been early adopters and for example in Geographical Information Systems (GIS) the R has started rival proprietary GIS-software. In the case of GIS in R it is possible to combine traditional statistical methods and programming with spatial data and statistics in one environment. In SSH this is useful as there are a lot of spatial data available and researcher may want to cluster the data thematically, but also visualize it as maps. As for humanities, Matthew Jockers [forthcoming] book is one of the first attemps to foster use of R. In digital humanities blogoshpere there are few others besides Jockers blog that are worth reading, namely W. Caleb McDaniel from Rice University and Quantifying Memory blog by Rolf Fredheim from Cambridge.

Addition to academic applications, R has become a major player in business analytics. This is largely due to R's capabilities in visualisation and analysing so called *big data*, but also die to companies like Revolution Analytics that have started proving consultation and creating tailored application for enterprise needs. Annual R/Finance 2014-conference gives nice overview of adoption of R in banking and insurance sectors. For example Google uses R in-house and also provides packages as r-google-analytics or rgooglevis. One emerging fiels is so called data journalism where major players as New York Times or Guardian use R in the data-driven stories such as this.

Conclusions

R is certainly not the only alternative for proprietary data analysis software or for analysis of complex digital data. For example, Python is another viable option especially for someone looking for more general purpose language that also masters data analysis. Whether python is going to displace R has recently been debated in data science blogosphere. The data analysis is becoming mainstream on many fields, not just in academic research, but R is still remaining hard to learn and very much researcn oriented. Programmers rather want to extend the language they already know than learn a new one and python is a lot more common than R. For very intensive computation Julia is becoming popular open source option, too. It is still in early phase of development, but is already viable option if processing time is important.

But for scientific work I would emphasize the *licensing* of the software more than the name of the particular technology. It is well possible that the recent buzz around digital data in SSH marks only the beginning of a data intensive research tradition. For someone wanting to success in that game it will be equally important to develop the subtantial understanding of the research topics as well as technological understanding of the new emerging tools. R is a prime example of this development where academics have taken a major role in software development and created tools that suits better for their research problems than proprietary software.

This development will go on and therefore it is advisable for someone who is interested in learning these techniques to carefully look at the licensing before investing time and effort in learning the technology. Free and open source tools are great in this respect as once you can pick up the skills to use the techonolgy, you soon will find that it needs to be improved for your purposes. In free and open source technology you can learn how the code works, write improvements and then publish them for the wider research community for use and for further development. In addition, free licensing also allows you to teach the technology, apply it in any purpose, including commercial, and to distribute it. Open source research software is not always the easiest and quickest way to get job done, but in the long run they are often worth the time invested.

In addition, the openness of the computational research methods is important from the *reproducibility* of your research. Along with demands for open access of research publications there are tendencies that more and more journals in computational sciences will require both the data and algorithms behind the results to published together with the article. As SSH scholars are moving towards computational analysis this issue of reproducibility should also be taken into account. R is a great tool that fullfills all these conditions, but are several other out there, too. After all, it is not necessary for all to become software developers, but to have basic understanding and to pair with developers who know more.

Steve Lohr [2013] interviewed some leading digital humanists in New York Times article *Literary History, Seen Through Big Data's Lens* on the future of SSH and posed a question whether these emerging computational technologies will undermine the role qualitative research in the field. Matthew Jocker, whose book *Macroanalysis: Digital Methods and Literary History* [2013] was central in the article, emphasized that finding the right questions and flaws in the analysis still requires deep, both qualitative and quantitative, understanding of the field:

> But we're at a moment now when there is much greater acceptance of these methods than in the past. There will come a time when this kind of analysis is just part of the tool kit in the humanities, as in every other discipline.

And that:

> Quantitative tools in the humanities and the social sciences, as in other fields, are most powerful when they are controlled by an intelligent human. Experts with deep knowledge of a subject are needed to ask the right questions and to recognize the shortcomings of statistical models.

The quest for new kind of collaboration between scholars and fields of research is also emphasized by professor Gary King [2014]. He claims that the analysis of large digital data requires skills that can't be found from traditional fields of social sciences.

> Through collaboration across fields, however, we can begin to address the interdisciplinary substantive knowledge needed, along with the engineering, computational, ethical, and informatics challenges before us.

In addition, King [2014] assumes that this collaboration will eventually blur the dichotomy between qualitative and quantitative analysis, and he portraits a future where both traditions have merged into social sciences where the important research problems are solved in collaboration.

> Instead of quantitative researchers trying to build fully automated methods and qualitative researchers trying to make due with traditional human-only methods, both now are heading toward, using, or developing computer-assisted methods that empower both groups. This development has the potential to end the divide, to get us working together to solve common problems, and to greatly strengthen the research output of social science as a whole.

This may well be true for humanities as well if we dare to take upo the challenge. (Or some other sentimental ending...)

References

Winston Chang. *R graphics cookbook*. O'Reilly, 2012.

Andy Field, Jeremy Miles, and Zoë Field. *Discovering Statistics Using R*. SAGE, March 2012. ISBN 9781446258460.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, November 2013. ISBN 9781439840955.

Matthew L Jockers. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Matthew L. Jockers. *Text Analysis with R for Students of Literature*. Quantitative Methods in the Humanities and Social Sciences. Springer, forthcoming.

Robert Kabacoff. *R in Action: Data Analysis and Graphics With R*. MANNING PUBN, November 2013. ISBN 9781617291388.

Gary King. Restructuring the social sciences: Reflections from harvard's institute for quantitative social science. *PS: Political Science and Politics*, 47: 165–172, 2014. URL http://journals.cambridge.org/repo_A9100Nlq.

Steve Lohr. Literary history, seen through big data's lens. *The New York Times*, January 2013. ISSN 0362-4331. URL http://www.nytimes.com/2013/01/27/technology/literary-history-seen-through-big-datas-lens.html.

Thomas Lumley. *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, September 2011. ISBN 9781118210932.

Tom A. B. Snijders and Roel Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications Ltd, second edition edition, December 2011. ISBN 184920201X.

William N. Venables, David M. Smith, and R. Development Core Team. *An introduction to R*. Network Theory, 2013. URL http://www.math.vu.nl/sto/onderwijs/statlearn/R-Binder.pdf.

Yihui Xie. *Dynamic documents with R and knitr*. CRC Press, 2014. ISBN 9781482203530 1482203537.