# R-language for statistical computing and visualization in social sciences and humanities

Markus Kainu

January 23, 2014

## Contents



## 1 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 use of open research methods has received a lot less attention due to obvious reason.

Firstly, research methods in SSH have been predominantly qualitative meaning that the software has played a supporting role in the analytical process. Second, the quantitative fields of SSH have mostly used survey and other, often closed, tailor-made research data that is well suited for analysis of custom statistical tools as SPSS, Stata or Excel. However, the future 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.

As relevant data for SSH is being generated and published in multiple sites and as the data is often messy and big in size this development calls for flexible tools that can be tailored and modified for the particular data and research question. The custom proprietary tools that are well fitted for analysis of survey, register and statistical data

Proprietary software vendors cannot keep up with this development and therefore we have witnessed an unseen growth in userbase and developers of free and open source computational research methods. And it is not only that there are no proprietary software for these purposes, but the open research methods are also required becaused of

As Gary King (2014) describes it:

> 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.

# 2 What is R?

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, Smith, and Team 2013). Currently the development of R language is run by *R Development Core Team* that Chambers is a member of.

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 runs in Windows, Mac OS X and GNU/Linux operating systems on local computer, but a different server implementations are becoming increasingly popular.

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 guides the user to implement the data-analysis as stepwise process which is very useful when solving complex research problems using complex and messy data that is often the case in contemporary computational SSH research.

## 2.1 Structure of R-project

The R Project for Statistical Computing

**Organisation of the project**

- development vs. user help

**Development of the language**

R installation consists of so called *base installation* that includes the core with some 25 additional *packages* for the most basic functionality. As said the 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. 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. 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.

Comprehensive R Archive Network (CRAN) is the "official" repository for contributed packages and it currently hosts 5150 packages that can be used to extend R. In last couple of years various code hosting sites as GitHub have become increasingly important resources especially for collaborative development of new packages. Github hosts currently rougly 1500 packages for R. Another domain spesific project is Bioconductor that provides packages for *for the analysis and comprehension of high-throughput genomic data*. Such domain spesific projects are for example rOpenSci and rOpenGov that provide tools for open science and open government data, respectively.

## 2.2 R-environment

## 2.3 R-language as a programming language

(G)UI's, IDE Rstudio, git - contributed packages

## 2.4 Popularity of R

- structure
- visual
- open source, community, licensing, teaching

## 2.5 Popularity of R-language

- enterprise level services

> R has already won praise and plaudits from established media outlets such as the New York Times, Forbes, Intelligent Enterprise, InfoWorld and The Register. When you consider that R is a high-level computer programming language designed mostly for quants (the nickname for a subspecies of geeks who focus on quantitative analysis), the adoring media attention seems nothing short of astounding.

Joka tapauksessa D. Smith (2010 s.23) kirjoittaa että paska on aina paskaa.

# 3 Using R-language

## 3.1 Learning the language

### User support

- mailing lists - general vs. special interest groups blogs
- q & a site

As the internet has brought together the vast community around R and internet has become the main channel for delivering instructions for R. 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). MOOC work well for teaching programming and many courses in Coursera and EdX have become hugely popular attracting tens of thousands

of students each year. Courses of statistical programming have predominantly teached R-language on these platforms as freely licensed software is basicly the only viable alternative for teaching statiscal 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 statitics using R. *Discovering Statistics Using R* by A. Field, Miles, and Field (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. *Bayesian Data Analysis* by Gelman et al. (2013) and *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* by Snijders and Bosker (2011).

## 3.2 Applications of R

- Development of statistical methods
- new methods implemented first in R
- Applied statistics
- Bio/geo-sciences
  - Geograhical Information Systems
- Social sciences/economics
- Humanities - analysis of natural languages
- Business/enterprise analytics
  - insurance, big data, banking, industry
  - social media: facebook, google, twitter
- Data journalism
  - Guardian, New York Times, Chicago Herald Tribune

## 3.3 User examples

- Word clouds
- Networks maps
- Topic modelling
- Spatial visualisation
- Clustering

# 4 Summing it up

Some alternatives

The importance of (free & open) licensing in scientific work

# 5 References


Chang, Winston. 2012. *R Graphics Cookbook*. O'Reilly.

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

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

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

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

King, Gary. 2014. "Restructuring the Social Sciences: Reflections from Harvards Institute for Quantitative Social Science." *PS: Political Science and Politics* 47: 165–172. http://journals.cambridge.org/repo_A9100Nlq.

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

Smith, David. 2010. "R Is Hot - How Did a Statistical Programming Language Invented in New Zealand Become a Global Sensation?" Revolution Analytics.

Snijders, Tom A. B., and Roel Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Second Edition. Sage Publications Ltd.

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

Xie, Yihui. 2014. *Dynamic Documents with R and Knitr*. CRC Press.