R-language for statistical computing and visualization for social and humanitarian sciences

Markus Kainu

January 23, 2014

Abstract

In sit amet sapien mollis ligula dignissim vehicula. Fusce turpis dui, semper consequat erat eget, laoreet rutrum diam.

Contents

1 Open Research Methods in general

The debate on open science in the context of Social sciencies 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 role of software or computation has been minor in the analytical process. Second, the quantitative fields of SSH have mostly used tailor-made survey and other specific closed research data that was well suited for analysis of custom statistical tools as SPSS, Stata or Excel. However, the future looks different as the quantity of digital data and is's multiple sources are challeging both traditional approach in SSH, the purely qualitative analysis and closed data practises quantitative analysis.

There are now custom tools for open and often messy and big new digital data, but we flexible tools that can be tailored and modified for the particular question and data. From this demand we have witnessed an unseen growth in user and developers of free and open source computational research methods.

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.

This was sensible choice in the era of closed data

Open research methods are referred as of a wider open science discussion.

2 What is R? - Origin and characteristics

2.1 Origin

- R-language was iniated by two
- open source

2.2 R-language as a programming language

- object oriented, s-language bell laboratories (G)UI's, IDE Rstudio,
- contributed packages

2.3 R-language as a tools for statistical computing

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

2.4 Popularity of R-language

- enterprise level services

3 Who makes the R possible? - R-project

3.1 Organisation of the project

- development vs. user help

3.2 Development of the language

3.3 User support

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

3.4 Contributed packages

- CRAN - task views R-forge
- Github
- bioconductor

4 What is R used for? - R in action

Why R-language is popular

4.1 Development of statistical methods

- new methods implemented first in R

4.2 Applied statistics

4.2.1 Bio/geo-sciences

- Geograhical Information Systems

4.2.2 Social sciences/economics

4.2.3 Humanities - analysis of natural languages

4.3 Business/enterprise analytics

- insurance, big data, banking, industry
- social media: facebook, google, twitter

4.4 Data journalism

- Guardian, New York Times, Chicago Herald Tribune

5 How does it work? - Visualising data in humanities using R-language

5.1 Word clouds

5.2 Networks maps

5.3 Spatial visualisation

5.4 Clustering

5.5

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.