Computational ecosystems for social science*

Markus Kainu [†1], Joona Lehtomäki [‡2], Juuso Parkkinen [§3], Juha Yrjölä [¶4], Måns Magnusson [‖5], and Leo Lahti [**6]

[1]Aleksanteri Institute, University of Helsinki, Finland
[2]Department of Biosciences, University of Helsinki, Finland
[3]Reaktor Innovations Oy, Finland
[4]Kansan Muisti ry, Finland
[5]Linköping University, Sweden
[6]Department of Veterinary Bioscience, University of Helsinki, Finland

May 29, 2015

Keywords: social science; elections; open government data; statistical programming; machine learning

Title & authors

[TODO lisätäänkö Mikko Tolonen tekijälistalle? Lisäksi pitää kysyä tahtooko Juha olla nyt mukana kun ei olekaan vaalijuttua + kansanmuistia posterissa]

Background

The recent explosion in open data availability has created novel opportunities for research, journalism and citizen science. High-quality machine readable data streams are increasingly available on political decision making, historical processes, welfare, traffic, and other aspects of society. There is a great need for analytical tools to take advantage of these new data streams in computational social science, digital humanities, and related fields.

Open source data analytics

Efficient data analysis relies on customized, reproducible analysis workflows that are best developed jointly by the user community. Availability of ready-made algorithms for standard data analysis tasks allows an individual researcher to avoid reinventing the wheel, leaving more time to solve the specific research problems. Solutions have emerged in data intensive research fields, such as bioinformatics and particle physics, based on open source statistical programming languages. In computational social sciences and digital humanities, analogous statistical software libraries are now emerging and have a huge potential to contribute to transforming the field. However, these resources are currently highly scattered and come in various formats, hindering wider adoption. Specific web-based tools are available, but more flexible computational tools are urgently needed for fully powered data processing and analysis.

*Extended abstract submitted for International Conference on Computational Social Science in June 8-11, 2015 in Helsinki, Finland
†markuskainu@gmail.com
‡joona.lehtomaki@helsinki.fi
§juuso.parkkinen@iki.fi
¶juha.yrjola@iki.fi
‖mans.magnusson@gmail.com
**leo.lahti@iki.fi

1 Example: Eurostat tools

2 Social coding

The ecosystem enables rapid development of scalable and interoperable software and provides tools to expand the quantitative methods base. The advantages of the open development model include:

- Open source: We use GitHub for shared version control. All contributions are openly licensed. This guarantees that the tools are freely available and the international scientific community remains the owner of the research software.
- Reproducible documentation: High-quality documentation is critical for package usability. We provide online tutorials with fully reproducible documentation on how to access and analyse specific data sources, and to report the statistical results.
- Transparent research: The programmatic approach makes it possible to publish the data analysis steps from raw data to the final results in full detail. To exemplify this, we publish reproducible case studies based on open data and algorithms in the rOpenGov blog.
- Standardization: A community-driven approach helps to pool scarce research resources and develop common standards for data analysis. Joint development ensures that the applicability of the tools extends beyond individual data sets and is compatible with other tools. Whereas different research projects can utilize the same standard algorithms to access and preprocess the data, the source code can be flexibly adapted to different tasks.

Further resources

The rOpenGov tools are distributed as R packages, including for instance:

- bibliographica: Bibliographic data analysis
- estc: British Library English Short Title Catalogue analytics
- eurostat: Eurostat open data analysis
- fennica: Finnish national bibliography analytics
- finpar: Finnish parliamentary data
- gisfin: Finnish geographic location information
- helsinki: Helsinki open data tools
- pxweb: R interface to PX-Web data (Statistics Finland & Sweden etc.)
- rustfare: Russian open data
- sorvi: Finnish open government data

[TODO: NÄITÄ VOISI LISTATA LISÄÄKIN VEPPISIVULTA JA IHAN RYHMITELLÄ JOHONKIN BOKSEIHIN?]


```
p <- ggplot(mtcars, aes(x=mpg,y=qsec,color=cyl))
p <- p + geom_point()
p <- p + geom_smooth(method="lm")
print(p)
```

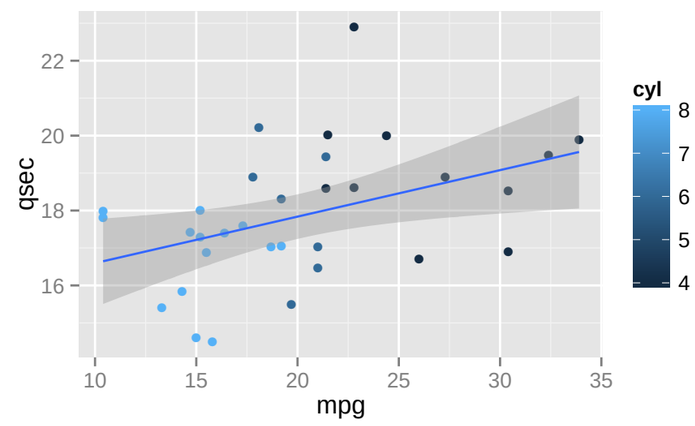

Figure 1: This is figure caption.

rOpenGov developer community and ecosystem

rOpenGov is a statistical ecosystem focused on open source data analysis algorithms relevant to computational social sciences and digital humanities. We provide a discussion forum and flexible algorithms for reproducible data analysis in these fields. We are a community of independent package developers from various countries, and we build on experiences learned from similar initiatives in other fields, such as Bioconductor and rOpenSci.

rOpenGov is based on the R statistical programming language, which has a versatile computational ecosystem with rich statistical modeling and state-of-the-art visualization capabilities. We are actively monitoring developments in other languages, such as Python and Julia. The wide scope of the R language is essential for addressing the diversity of analysis tasks. We complement the prevailing R ecosystem with custom tools for computational social sciences and digital humanities. The packages are distributed through Github (ropengov.github.io). The project is maintained by a core team and a number of independent contributors (see the rOpenGov site for the up-to-date author list).


```
p1 <- p + geom_smooth(method="lm")
p2 <- p + geom_smooth(method="glm")
p3 <- p + geom_smooth(method="loess")
p4 <- p + geom_smooth(method="gam")
grid.arrange(p1,p2,p3,p4, ncol = 2)
```

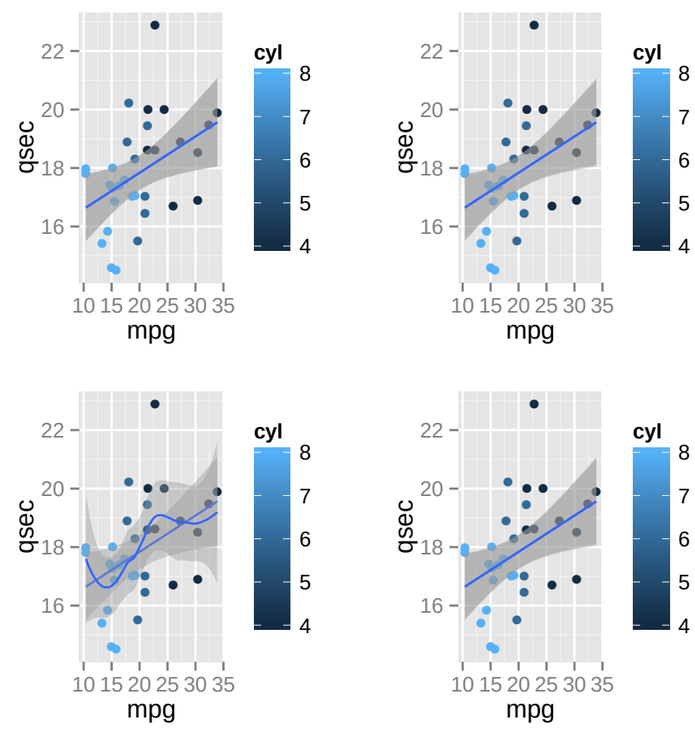

Figure 1: This is figure caption, too!

Contact & Contribute

- Web & Blog: ropengov.github.io
- IRC: ropengov@Freenode

References [ONKOHAN TARPEEN POSTERISSA? JÄTTÄISIN EHKÄ POIS]

[1] S. Kasberger (2012). Grazwahl: Data Analysis and Visualizations of the communal elections in Graz.R package [2] S. Fortunato and C. Castellano (2012). Physics peeks into the ballot box. Physics Today 65:74 [3] G. King, J. Pan and M. E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review, 107(02), 326–343 [4] M. L. Jockers (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press. [5] S. Chou, W. Li and R. Sridharan, Democratizing Data Science. [6] D. Lazer, et al. (2009). Computational Social Science 323, 721–723