R for reproducible geographical research

Edzer Pebesma
(joint work with Daniel Nüst* and Roger Bivand**)
This is ongoing work, and most of the core ideas are not original.

- What is reproducible research?
- What is R? Why R?
- How can R be used for reproducible geographical research?
- Challenges
- Outlook
Why is reproducible research a good thing?

- the credibility of science is at stake when research is not reproducible
- we cannot repeat observation, but we can repeat the procedures that led us from observations to research findings and conclusions
- even in cases where data cannot be shared, sharing procedures will increase credibility
- even in case of errors, being able to trace them back to the source (data? script? software?) increases credibility
Why is reproducible research a good thing?

- the credibility of science is at stake when research is not reproducible
- we cannot repeat observation, but we can repeat the procedures that led us from observations to research findings and conclusions
- even in cases where data cannot be shared, sharing procedures will increase credibility
- even in case of errors, being able to trace them back to the source (data? script? software?) increases credibility

then, why don’t we do this? Why is reproducibility not compulsory?
Claerbout’s Principle

Claerbout’s Principle\textsuperscript{1}:

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

\textsuperscript{1}J. de Leeuw, Reproducible research: the bottom line, statistics Program, University of California, Los Angeles, CA, USA (2001), quoting: J. Buckheit, D. Donoho, WaveLab and reproducible research, statistics Department, Stanford University, CA, USA (1995).
Versioning of research?

- once accepted, papers are rarely revised
- data and analysis scripts are, over time, typically improved
- to keep the link between submitted paper and the research to reproduce it, data and scripts should be equally frozen, and be part of the submission procedure, and downloadable with the paper
- the author can provide (documented) updates of procedures.
R is a free, open source environment for statistical computation and graphics, running on all operating systems.

R is developed and maintained by about 20 PhD/professors in academia.

R has an unknown number of users.

R can be extended by add-on packages.

Around 4000 of such packages are part of R, and are developed and maintained by a similar number of developers.

Last year, R entered the top-20 most used programming languages.

Increasingly, R is the platform of choice for teaching and research, both in academia and industry.

---

\(^2\)Forbes (2011) estimated it to be 2 Million
Before 2005

R

- spatstat
- maptools
- splancs
- spdep
- geoR
- gstat
After 2005

R

sp

spatstat

maptools

sp

splancs

spdep

geoR

gstat
2011: over 100 spatial packages on CRAN

- maptools
- spatstat
- spdep
- geoR
- gstat
- rgdal
- rgeos
- raster
- +100
How can R be used for reproducible geographical research?

Stage I: share data and scripts
- share and discuss research ideas, as a basic attitude, by scripts that need three mouse clicks maximally to reproduce (this is basically what r-help and r-sig-geo require)
- provide, in your paper, reference to the software used
- write, in your paper, *that* you are willing to share (data and) scripts needed to reproduce the research
- share these scripts through a web site (*)
- provide the URL in your paper (*)
- submit data and procedures (scripts) as supplementary material

(*) this puts responsibility on the author’s side
How can R be used for reproducible geographical research?

Stage II: the **executable paper**

- inspired by Donald Knuth’s *literate programming*, R has since long had Sweave, for literate analysis.
- **Sweave** documents mix text (the journal paper), and R scripts.
- **Sweave** runs R, and automatically merges (“weaves”) the output (text, figures) needed to generate the final document (e.g. pdf, html)
- R code is *replaced* or *augmented* with the output of running R
- many R books are being written in this system, including “Applied Spatial Data Analysis with R” (Bivand, Pebesma, Gomez-Rubio)
- **Sweave** papers are *executable* papers
Challenges

- when submitting an executable paper, how will the journal publisher react when we submit 15 years of Landsat imagery for the Amazon basin?
- if historic data is improved over time, can we still access the version of the data on which a particular outcome was based?
- how does versioning of data carry through in the linked data cloud?
- scripts work with particular versions of software (R, add-on packages). Where do we find working instances of the software as it was used to execute the paper in 2002?
- What are the responsibilities of journal publisher, and what are those of the author?
Outlook

- we need to convince journals that reproducible papers are (i) useful, (ii) reproducible, and (iii) better than non-reproducible ones.
- we ought to start rejecting papers without accompanying readable and reproducible procedures
- we probably need more disasters, climate-gates, etc., before this will fly
- we need to adopt open science, and reproducibility, as the default case, not as an exception.
Further reading