Making one's work reproducible

Christophe Pouzat

December 16, 2017

Contents

1 Introduction

1.1 What is reproducible research?

> 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.

Thoughts of Jon Claerbout "distilled" by Buckheit and Donoho (1995).

The preparation of manuscripts and reports in neuroscience often involves a lot of data analysis as well as a careful design and realization of figures and tables, in addition to the time spent on the bench doing experiments. The data analysis part can require the setting of parameters by the analyst and it often leads to the development of dedicated scripts and routines. Before reaching the analysis stage *per se* the data frequently undergo a preprocessing which is rarely extensively documented in the methods section of the paper. When the article includes numerical simulations, key elements of the analysis, like the time step used for conductance based neuronal models, are often omitted in the description. As readers or referees of articles / manuscripts we are therefore often led to ask questions like:

- What would happen to the analysis (or simulation) results if a given parameter had another value?
- What would be the effect of applying my preprocessing to the data instead of the one used by the authors?

- What would a given figure look like with a log scale ordinate instead of the linear scale use by the authors?
- What would be the result of applying that same analysis to my own data set ?

We can of course all think of a dozen of similar questions. The problem is to find a way to address them. Clearly the classical journal article format cannot do the job. Editors cannot publish two versions of each figure to satisfy different readers. Many intricate analysis and modeling methods would require too long a description to fit the usual bounds of the printed paper. This is reasonable for we all have a lot of different things to do and we cannot afford to systematically look at every piece of work as thoroughly as suggested above. Many people (Claerbout and Karrenbach, 1992; Buckheit and Donoho, 1995; Rossini and Leisch, 2003; Baggerly, 2010; Diggle and Zeger, 2010; Stein, 2010) feel nevertheless uncomfortable with the present way of diffusing scientific information as a canonical (printed) journal article. We suggest what is needed are more systematic and more explicit ways to describe how the analysis (or modeling) was done.

These issues are not specific to published material. Any scientist after a few years of activity is very likely to have experienced a situation similar to the one we now sketch. A project is ended after an intensive work requiring repeated daily long sequences of sometimes "tricky" analysis. After six months or one year we get to do again very similar analysis for a related project; but the nightmare scenario starts since we forgot:

- The numerical filter settings we used.
- The detection threshold we used.
- The way to get initial guesses for our nonlinear fitting software to converge reliably.

In other words, given enough time, we often struggle to exactly reproduce *our own* work. The same mechanisms lead to know-how being lost from a laboratory when a student or a postdoc leaves: the few parameters having to be carefully set for a successful analysis were not documented as such and there is nowhere to find their typical range. This leads to an important time loss which could ultimately culminate in a project abandonment.

We are afraid that similar considerations sound all too familiar to most of our readers. It turns out that the problems described above are not specific to our scientific domain, and seem instead to be rather common at least in the following domains: economics (Dewald et al., 1986; Anderson and Dewald, 1994; McCullough et al., 2006; McCullough, 2006), geophysics (Claerbout and Karrenbach, 1992; Schwab et al., 2000), signal processing (Vandewalle et al., 2009; Donoho et al., 2009), statistics (Buckheit and Donoho, 1995; Rossini, 2001; Leisch, 2002), biostatistics (Gentleman and Temple Lang, 2007; Diggle and Zeger, 2010), econometrics (Koenker and Zeileis, 2007), epidemiology (Peng and Dominici, 2008) and climatology (Stein, 2010; McShane and Wyner, 2010) where the debate on analysis reproducibility has reached a particularly acrimonious stage. The good news about this is that researchers have already worked out solutions to our mundane problem. In the next section we review some of the already available tools for reproducible research, which include data sharing and software solutions for mixing code, text and figures.

References

Richard G. Anderson and William G. Dewald. Replication and scientific standards in economics a decade later: The impact of the jmcb project. Working Paper 1994-007C, Federal Reserve Bank of St. Louis, 1994. URL http://research.stlouisfed.org/wp/more/1994-007/. Available at: http://research.stlouisfed.org/wp/more/1994-007/.

Keith Baggerly. Disclose all data in publications. *Nature*, 467(7314):401–401, September 2010. ISSN 0028-0836. URL http://dx.doi.org/10.1038/467401b.

Jonathan B. Buckheit and David L. Donoho. *Wavelets and Statistics*, chapter Wavelab and Reproducible Research. Springer, 1995. Preprint available at: http://www-stat.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf.

Jon Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *Proceedings of the 62nd Annual Meeting of the Society of Exploration Geophysics*, pages 601–604, 1992. URL http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92. Available at: http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92.

William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. Replication in empirical economics: The journal of money, credit, and banking project. *American Economic Review*, 76(4):587–603, 1986. ISSN 00028282. URL http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=4497372&site=ehost-live.

Peter J. Diggle and Scott L. Zeger. Editorial. *Biostatistics*, 11(3):375–375, July 2010. URL http://biostatistics.oxfordjournals.org/content/11/3/375.short.

David L. Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. Reproducible research in computational harmonic analysis. *Computing in Science and Engineering*, 11:8–18, 2009. ISSN 1521-9615. doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2009.15. Preprint available at: http://www-stat.stanford.edu/~donoho/Reports/2008/15YrsReproResch-20080426.pdf.

Robert Gentleman and Duncan Temple Lang. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16(1):1–23, 2007. doi: 10.1198/106186007X178663. URL http://pubs.amstat.org/doi/abs/10.1198/106186007X178663.

Roger Koenker and Achim Zeileis. Reproducible econometric research. a critical review of the state of the art. Research Report Series / Department of Statistics and Mathematics 60, Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna., 2007. URL http://epub.wu.ac.at/638/. Available at: http://epub.wu.ac.at/638/.

Friedrich Leisch. Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL http://www.stat.uni-muenchen.de/~leisch/Sweave. Available at: http://www.statistik.uni-muenchen.de/~leisch/Sweave/.

B. D. McCullough. Section editor's introduction. *Journal of Economic and Social Measurement*, 31(1-2):103–105, 2006. URL http://www.pages.drexel.edu/~bdm25/publications.html. Available at: http://www.pages.drexel.edu/~bdm25/publications.html.

B. D. McCullough, Kerry Anne McGeary, and Teresa Harrison. Lessons from the jmcb archive. *Journal of Money, Credit and Banking*, 38(4):1093–1107, 2006. URL http://www.pages.drexel.edu/~bdm25/publications.html. Available at: http://www.pages.drexel.edu/~bdm25/publications.html.

Blakeley B. McShane and Abraham J. Wyner. A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? To be published in The Annals of Applied Statistics., 2010. URL http://www.e-publications.org/ims/submission/index.php/AOAS/user/submissionFile/6695?confirm=63ebfddf.

Roger D. Peng and Francesca Dominici. *Statistical Methods for Environmental Epidemiology with R*. Use R! Springer, 2008.

A. J. Rossini. Literate Statistical Analysis. In Kurt Hornik and Friedrich Leisch, editors, *Proceedings of the 2nd International Workshop on Distributed Statistical Computing, Vienna, Austria*, 2001. URL http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/. ISSN 1609-395X.

Anthony Rossini and Friedrich Leisch. Literate Statistical Practice. UW Biostatistics Working Paper Series 194, University of Washington, 2003.

M. Schwab, N. Karrenbach, and J. Claerbout. Making scientific computations reproducible. *Computing in Science & Engineering*, 6:61– 67, 2000. URL http://sep.stanford.edu/doku.php?id=sep:research:reproducible. Preprinty available at: http://sep.stanford.edu/lib/exe/fetch.php?media=sep:research:reproducible:cip.ps.

Michael L. Stein. Editorial. Available at: http://www.e-publications.org/ims/submission/index.php/AOAS/user/submissionFile/8887?confirm=6adde642, 2010.

Patrick Vandewalle, Jelena Kovacevic, and Martin Vetterli. Reproducible research in signal processing - what, why, and how. *IEEE Signal Processing Magazine*, 26(3):37–47, May 2009. URL http://rr.epfl.ch/17/. Available at: http://rr.epfl.ch/17/.