[Re] Drivers of evapotranspiration from boreal wildfires

Ben Bond-Lamberty

04 May 2020

Abstract

Computational reproducibility is a difficult challenge across science. I attempted to use R 3.6.1 to reproduce linear model fits, done originally using v2.6.0 for a 2009 paper on the drivers of large-scale forest evapotranspiration after wildfire. Model outputs were largely identical, aside from minor formatting changes, except for one–out of 12 total–regression in which the median residual value changed very slightly (in the sixth decimal place). I suggest that this essentially successful reproducibility is due to the relative simplicity of the script, its use of only base R functions, and R's historically conservative approach to breaking changes.

Introduction

Significant changes in forest fires have occurred in recent decades in the global boreal (high latitude) forest, but the effects of changing climate and disturbance on evapotranspiration (ET) and forest water cycling more generally are not well understood. In a previous study (Bond-Lamberty et al. 2007) we had explored the ecological and carbon-cycle consequences of changes in the fire regime using an ecosystem model run at high resolution across a large (1000 km x 1000 km) area of the central Canadian boreal forest. In a subsequent paper (Bond-Lamberty et al. 2009), we used the same model outputs to examine the *hydrological* implications of these wildfire regime shifts (Nolan et al. 2014).

How reproducible are the results reported in Bond-Lamberty et al. (2009)? The article manuscript was written in 2007-2008, initially submitted in April 2008, and a final, revised version was submitted in October of that year. The system timestamp on the script output file is 2008-03-16. The methods section reports that R 2.6.0 (R Core Team 2007), which was released on 2007-10-03, was used for all analyses. The source code was not publicly archived, but retained in the lead author's personal records; it has no license. The original hardware would have been an Apple laptop (probably a 2006-2007 MacBook Pro). As noted above, the code file was archived by the lead author, and so was easy for him (but no one else) to find. The code has no comments or instructions.

Methods

Retrieval of the software

The source code being reproduced here is short and was written to analyze relationships between various potential driving factors and three output variables of interest: ET, canopy evaporation, and canopy transpiration, all annual flux totals over the 1948-2005 model run period. The code simply reads in a comma-separated data file holding the modeling outputs, prints a summary of the data, and then fits and prints 12 separate regressions (three output variables times four possible independent variables). These fit statistics were reported on p. 1247 of Bond-Lamberty et al. (2009).

Execution

Because the code uses only base R functions, it has no dependencies other than a standard R installation. Both in 2007-2008 and 2020, the default Mac R installer provided by CRAN was used to install R (i.e. it was not built from source). The code was re-run, without modification, on a 2018 MacBook Pro under R version 3.6.1 (2019-07-05) (R Core Team 2019); platform: x86_64-apple-darwin15.6.0 (64-bit); running under macOS

Mojave 10.14.6. Its output was then compared with the printed output from the 2007-2008 R 2.6.0 code, which had been recorded via R's `sink()` function and archived with the code.

Results and discussion

Setting aside minor spacing and text capitalization changes, only one numerical change occurred, albeit in the sixth decimal place. In the 2008 R 2.6.0 output of a linear regression between canopy evaporation (dependent variable) and precipitation (independent), the residuals were given as:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.028858 -0.009186  0.002088  0.010753  0.019997
```

In the 2020 R 3.6.0 output, this output line (specifically, the "Median" value) was:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.028858 -0.009186  0.002089  0.010753  0.019997
```

In addition, in two cases p-values were reported with a different number of significant digits in the 2020 output: once gaining a digit ("0.005451" in 2020 versus "0.00545" in 2008) and once losing one ("0.000103" and 0.0001030" respectively). No other numerical differences occurred.

Whether these minor differences were due to the different R version, or different underlying operating system and/or libraries, is uncertain. R version 2.6.0 (https://cran.r-project.org/bin/macosx/old/R-2.6.0.dmg) is unfortunately not installable under modern versions of macOS. Instead I downloaded R 2.6.0 for Windows (`i386-pc-mingw32`) and ran the code on it; the printed results were identical to the 2007-2008 output file. This suggests that the minor differences noted above between the R 2.6.0 and 3.6.0 outputs were due to changes in R itself, not differences in the underlying systems.

As simple as the code and this reproducibility exercise was, there are some interesting notes that can be drawn from it. Reproducibility has been a well-supported (see e.g. https://cran.r-project.org/web/views/ReproducibleResearch.html) core value of the R Core Team and larger communnity, and the base version of the software has been highly stable over the last twenty years since version 1.0. Specifically with regard to this exercise, the core of the `lm` (linear model) source code, found at https://svn.r-project.org/R/trunk/src/library/stats/R/lm.R, has changed little over the last decade. Refinements and extensions to R's linear modeling capabilities have instead come from contributed packages (e.g. Grömping 2006).

In contrast, many of the most highly used R packages, for example those of the popular tidyverse ecosystem (Wickham et al. 2019) have historically changed their behavior and syntax much more frequently. This allows for faster evolution and cleaner, consistent syntax (not among base R's strong points) but poses greater challenges for stable and reproducible software. For example, the popular `dplyr` package lists (see https://github.com/tidyverse/dplyr/blob/master/NEWS.md) breaking changes introduced at versions 0.5, 0.7, 0.7.5, 0.8, 0.8.1, and 1.0, over a timespan of four years; it seems unlikely that code taking advantage of its powerful features and speed would be reproducible without being tweaked or rewritten. This highlights some of the challenges surrounding computational reproducibility (Thornton et al. 2005; Lowndes et al. 2017), in particular balancing reproducibility with other potentially important criteria such as performance or confidentiality (Stodden 2014; Peng 2011).

Conclusions

There were many potential limitations in reproducing even this very short analysis script over ten years later: the code contained no documentation about R or package versions, nor information about the hardware information it was originally run on. The analysis date could only be reconstructed from the output file timestamp (luckily unaltered). Saving this output file, however, at least allowed for a robust check on the reproducibility of the analysis, script, and underlying R software after 12 years. The essentially successful result is due to the relative simplicity of the script, its use of only base R functions, and R's conservative approach to breaking changes.

References

Bond-Lamberty, Ben, Scott D Peckham, Douglas E Ahl, and Stith T Gower. 2007. "Fire as the Dominant Driver of Central Canadian Boreal Forest Carbon Balance." *Nature* 450 (7166): 89–92.

Bond-Lamberty, Ben, Scott D Peckham, Stith T Gower, and Brent E Ewers. 2009. "Effects of Fire on Regional Evapotranspiration in the Central Canadian Boreal Forest." *Glob. Chang. Biol.* 15 (5): 1242–54.

Grömping, Ulrike. 2006. "Relative Importance for Linear Regression in R: The Package Relaimpo." *J. Stat. Softw.* 17 (1): 1–27.

Lowndes, Julia S Stewart, Benjamin D Best, Courtney Scarborough, Jamie C Afflerbach, Melanie R Frazier, Casey C O'Hara, Ning Jiang, and Benjamin S Halpern. 2017. "Our Path to Better Science in Less Time Using Open Data Science Tools." *Nat Ecol Evol* 1 (6): 160.

Nolan, Rachael H, Patrick N J Lane, Richard G Benyon, Ross A Bradstock, and Patrick J Mitchell. 2014. "Changes in Evapotranspiration Following Wildfire in Resprouting Eucalypt Forests." *Ecohydrol.* 6 (January). Wiley Online Library.

Peng, R D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–7.

R Core Team. 2007. "R: A Language and Environment for Statistical Computing, Version 2.6.0." Vienna, Austria: R Foundation for Statistical Computing.

———. 2019. "R: A Language and Environment for Statistical Computing, Version 3.6.1." Vienna, Austria: R Foundation for Statistical Computing.

Stodden, Victoria. 2014. "Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency." *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 1. Cambridge University Press: 112–35.

Thornton, Peter E, Robert B Cook, Bobby H Braswell, Beverly E Law, Wilfred M Post, Herman H Shugart, B Timothy Rhyne, and Leslie A Hook. 2005. "Archiving Numerical Models of Biogeochemical Dynamics." *EOS* 86 (44): 431–32.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43). joss.theoj.org: 1686.