Creating optimal conditions for reproducible data analysis in R with `fertile`

Audrey Bertin *
Program in Statistical and Data Sciences, Smith College
and
Benjamin S. Baumer
Program in Statistical and Data Sciences, Smith College

December 8, 2019

Abstract

The advancement of scientific knowledge increasingly depends on ensuring that data-driven research is reproducible: that two people with the same data obtain the same results. However, while the necessity of reproducibility is clear, there are significant behavioral and technical challenges that impede its widespread implementation, and no clear consensus on standards of what constitutes reproducibility in published research. We focus on a series of common mistakes programmers make while conducting data science projects in R, primarily through the RStudio integrated development environment. `fertile` is an R package that operates in two modes: proactively (to prevent reproducibility mistakes from happening in the first place), and retroactively (analyzing code that is already written for potential problems). Furthermore, `fertile` is designed to educate the user about why the mistakes are problematic, and how to fix them. We discuss experimental results from testing `fertile` in an introductory data science course.

Keywords: reproducibility, statistical software, workflow, collaboration

*The authors gratefully acknowledge contributions from Hadley Wickham, Jenny Bryan, Greg Wilson, Edgar Ruiz, and other members of the `tidyverse` team.

1 Extended Abstract

In the field of data science, an analyst has created *reproducible* work if: 1) their code and data files have been made publicly available, and 2) these files produce identical results when run by another analyst.

In the age of modern computer technology, where knowledge can be instantly shared worldwide, reproducibility is becoming incredibly important to the advancement of scientific knowledge.

The first reason for this is that reproducible research is transparent. When researchers provide the code and data used for their work in a well organized and reproducible format, readers are able to understand the steps taken to generate findings from raw data and determine whether or not they believe the results to be reliable or trustworthy. Without reproducibility, research results must be blindly trusted.

The second reason is that reproducibility allows for collaboration and extended research. The creators of reproducible research can easily receive feedback and recommendations on their work, and others interested in the research topic can see the code and easily apply the methods and ideas used in one project to their own future work with minimal effort.

However, while the necessity of reproducibility is clear, there are significant behavioral and technical challenges that impede its widespread implementation, and no clear consensus on standards of what constitutes reproducibility in published research.

In defining the steps toward creating reproducible analyses, different researchers place their emphasis on different areas, including file structure, documentation, the use of file paths in code, and accessibility. Within these categories, the recommendations for what steps to take in ensuring reproducibility also vary.

Much of the discussion around reproducibility is also generalized, written in a way applicable to users working with a variety of statistical software programs. Since all statistical software programs operate differently, generalized recommendations on reproducibility often cannot go in depth. While they provide some useful guidelines, they can often be relatively unhelpful, particularly to new analysts who are unsure how to apply such recommendations within the software programs they are using.

In order to be most effective, reproducibility recommendations must be tailored to

specific software programs. However, there has been very little work on reproducibility that is tailored to this level of specificity. In this paper, attempt to remedy that, focusing on reproducibility in the R programming language with a concentration on the RStudio integrated development environment.

R is an ideal candidate for reproducibility recommendations due to the language's popularity for statistical analyses and the ease with which analysts can download and begin using the software.

We propose an R package called `fertile`, which focuses on a series of common reproducibility-harming mistakes programmers make while conducting data science projects in R, warning users of their errors and providing recommendations for how to correct them.

`fertile` operates in two modes: proactively (to prevent reproducibility mistakes from happening in the first place), and retroactively (analyzing code that is already written for potential problems).

Retroactively, `fertile` is designed to be run on an R Project folder, analyzing the project structure and the analyst's use of file paths and considering randomness, among other areas of interest. `fertile` creates reproducibility reports, identifying mistakes that users have made and providing recommendations for remedies. For example, fertile might identify an analyst's use of randomness in code and recommend setting a seed.

Proactively, the package works similarly, identifying mistakes as they are made by the user and outputting an educational warning message identifying, and providing a solution for, the mistake. In its proactive warning system, `fertile` focuses primarily on the use of file paths. As users execute code, `fertile` looks for file paths passed to functions, identifying when absolute paths are provided or when paths point to a location outside of the project directory.

`fertile` is designed to educate the user about why the mistakes are problematic, and how to fix them. It is written for a wide audience, simple enough to be used by students in an introductory data science course.

2 Introduction

3 The Importance of Ensuring Reproducibility

One reason for this is that reproducible research encourages transparency. Providing one's audience with well organized, reproducible code and results allows readers to understand the steps taken to generate findings from raw data and determine for themselves whether or not they believe the results to be reliable or trustworthy.

Additionally, reproducibility encourages collaboration and extended research, allows others to easily apply the methods used in one project to their own work with minimal effort. Bray et al. (2014)

Fertile is designed to make reproducibility simple, providing fast and easy methods to test an R project for reproducibility. The package is intended to be usable by introductory data science students in their first semester of R.

–> make sure this info is included somewhere:

From "Five Concrete Reasons Your Students Should Be Learning to Analyze Data in the Reproducible Paradigm" (Bray et al. (2014))

- Helps with transparency: make clear data cleaning steps between raw data and final data
- Helps with collaboration: easier to share code when it takes very few steps to run on a different computer

The other big ideas on the importance/use of this package (not from sources):

- Make it easier for professors to grade students' code
- Should be usable by intro level data science students
- Possible use by reviewers of journal articles (and by those writing the articles)

4 What Defines Reproducibility in Data Science?

5 Creating Comprehensive Reproducibility Reports With "fertile"

Main points of different sources, as well as info about how they might be used for the paper.

Here are some sources we might use for motivation behind the project:

5.0.1 The Reproducibility Crisis

Big idea: most scientific fields are facing a reproducibility crisis and poor statistical use is considered one of the important reasons behind this.

Not sure how useful this is due to the fact that it does not necessarily focus on the same kind of reproducibility we are looking at, which is code reproducibility rather than experimental reproducibility.

Baker (2016)

5.0.2 Why is reproducibility important?

From Popper in `The Logic of Scientific Discovery`: "non-reproducible single occurrences are of no significance to science."

Popper (2005)

5.0.3 What does reproducibility mean in data science?

"The ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results."

Goodman et al. (2016)

From "The Practice of Reproducibile Research": Can all of the figures/calculations related to the result be reproduced in a single button press or at least with a reasonably small effort?

Kitzes et al. (2017)

5.0.4 What makes a data science project reproducibile?

Another way to think about this is: what features does fertile have that match what different sources think needs to be checked to ensure reproducibility?

Some big ideas from the OpenSci discussion page:

- code should use tidy style
- project should have short vignette files – most written code should be in an R/ directory
- code and data should be stored in separate folders
- a run environment and dependencies should be specified
- there should be a README file
- there should be a data dictionary
- one master script
- code should only use relative paths
- code should be well commented and all variables should be defined

Big ideas from "Packaging data analytical work reproducibly using R"

- research should be organized like an R package!
- clear separation of data, method, and output
- specify the computational environment that was used for the original analysis (typically in a plain text file)
- there should be a README describing the project and where to get started
- script files with reusable functions should go in an R/ directory
- raw data files should be kept in a data/ directory
- analysis scripts and report files should go in an analysis/ directory
- for simpe projects, scripts should be given ascending names. For more complicated ones, the use of some sort of makefile is recommended
- there should also be a DESCRIPTION file with information about the authors, project license, and software dependencies

Marwick et al. (2018)

From "The Practice of Reproducible Research"

- Are the data openly accessible? If hosted online, is the web address reliable long-term?
- Are they in a commonly used and well-documented file format? Avoid spreadsheets and instead use plain text data if possible!
- Is the raw data available? Is sufficient metadata provided?
- Are dependencies described properly?
- Is full history of source code available through a public version history
- Is there a README?
- Are functions documented?
- Is there narrative documentation explaining how the different pieces work together?
- Are there usage examples?

Folder setup might look like:

1. Raw Data
    - Data
    - README
2. Clean Data
    - Data
3. Results
    - Results file
4. Src
    - Analysis script

- Script to clean data

Kitzes et al. (2017)

From R OpenSci's Reproducibility Guide:

`http://ropensci.github.io/reproducibility-guide/`

- Is it clear where to begin?
- Can you determinine which files were used as input to create output files?
- Is there documentation about every result?
- Are exact versions of external applications noted?
- If using randomness, are seeds noted?
- Have you specified a license or noted licenses if you used other people's content?
- Are files easy to find?
- Is it clear what the most recent file is?
- Are there any folders that could be deleted?
- Is analysis output done hierarchically?
- Are there lots of manual data manipulation steps?

From "A Guide to Reproducible Code in Ecology and Evolution" (these ideas are pretty universal, though):

A basic project structure:

- The data folder contains all input data (and metadata) used in the analysis.
- The doc folder contains the manuscript
- The figs directory contains figures generated by the analysis
- The output folder contains any type of intermediate or output files (e.g. simulation outputs, models, processed datasets, etc.). You might separate this and also have a cleaned-data folder.
- The R directory contains R scripts with function definitions.
- The reports folder contains RMarkdown files that document the analysis or report on results

- Consistent, ordered naming of scripts
- Use portable paths
- Write unit tests (only for advanced coding)
- Show the packages you used
- Record dependencies and versions of outside things you use

Cooper et al. (2017)

6 Karl Broman's Suggestions

- Encapsulate everything within one directory
- Separate raw data from derived data
- Separate data from code
- Use relative paths
- Choose filenames correctly
- Write README files

`http://kbroman.org/steps2rr/pages/organize.html`

7 Why focus on R?

- R is the most popular language for statistical programming and is specifically designed for statistics
- R is great for reprodubility because the code is readable by users, and RMarkdown is a great way to show/explain processes
- R is easy to install and begin using

–> from `https://openresearchsoftware.metajnl.com/articles/10.5334/jors.bu/print/`

8 Similar R Packages

8.0.1 rrtools

`https://github.com/benmarwick/rrtools`

- Creates a basic R package named after your research topic
- Generates a license file
- Connects to GitHub and creates a repository
- Generates a README
- Generates a reproducible directory structure
- Creates a dockerfile
- Creates a minimal travis file
- Sets up testthat

9 Msc sources to look at:

9.0.1 Victoria Stodden's "Implementing Reproducible Research" book

`https://books.google.com/books?hl=en&lr`=&id=JcmSAwAAQBAJ&oi=fnd&pg=PP1&dq=Victoria

References

Baker, M. (2016), '1,500 scientists lift the lid on reproducibility', *Nature News* **533**(7604), 452.

Bray, A., Çetinkaya-Rundel, M. & Stangl, D. (2014), 'Five concrete reasons your students should be learning to analyze data in the reproducible paradigm', *RPubs retrieved from http://rpubs. com/mine/21454* .

Cooper, N., Hsing, P.-Y., Croucher, M., Graham, L., James, T., Krystalli, A. & Michonneau, F. (2017), 'A guide to reproducible code in ecology and evolution', *British Ecological Society* .

Goodman, S. N., Fanelli, D. & Ioannidis, J. P. (2016), 'What does research reproducibility mean?', *Science translational medicine* **8**(341), 341ps12–341ps12.

Kitzes, J., Turek, D. & Deniz, F. (2017), *The practice of reproducible research: case studies and lessons from the data-intensive sciences*, Univ of California Press.

Marwick, B., Boettiger, C. & Mullen, L. (2018), 'Packaging data analytical work reproducibly using R (and friends)', *The American Statistician* **72**(1), 80–88.
URL: *https://peerj.com/preprints/3192.pdf*

Popper, K. (2005), *The logic of scientific discovery*, Routledge.