Reproducible Geocomputation: an open or shut case?

Mark Gahegan*[1]

[1]Centre for eResearch, the University of Auckland, New Zealand
*Email: m.gahegan@auckland.ac.nz

Abstract

This submission tackles the issue of reusability and replication of experiments and other forms of spatial analysis from the perspective of eScience or eResearch—a community that has been deliberately grappling with this issue for many years—and applies it to our geocomputational and GIScience methods and models.

Most of the effort that routinely goes into our analysis and modelling efforts is not documented in any consistent or accessible way. It might be in part referable from scripts and code we use, and from the 'methods' section of papers that we write. But by and large much of the background, context and process of research remain unrecorded. This makes reproducibility hard (or intractable) and reusability nigh on impossible. Funding agencies worldwide are beginning to insist that we do better, and this pressure is now beginning to be felt in geography and the spatial sciences as well as in psychology and biology (where the science 'reproducibility crisis' began).

We describe several approaches to improve reusability and reproducibility for geocomputational work, and point to best practice from other disciplines as appropriate. (The presentation will elaborate on each of these headings and give examples relevant to geocomputation and spatial analysis.

Keywords: Reproducibility, geocomputation, GIScience, Open Science, Computational Workflows, Virtual Laboratories, Computational Notebooks

1. Heading

There is a crisis in science around reproducibility (Baker, 2015). It started in psychology, spread to biology and medicine and is now affecting most disciplines. The painful scrutiny shows that our research practices are, by and large, inadequate. So change is needed. Geocomputation and GIScience are not immune; indeed, we are as guilty as the next discipline. But we can help lead the way in approaches to resolve the crisis.

Most of the effort that routinely goes into our analysis and modelling efforts is not documented in any consistent or accessible way. It might be in part referable from scripts and code we use, and from the 'methods' section of papers that we write. But by and large much of the background, context and process of research remain unrecorded. This makes reproducibility hard (or intractable) and reusability nigh on impossible. Funding agencies worldwide are beginning to insist that we do better, and this pressure is now beginning to be felt in geography and the spatial sciences as well as in psychology and biology (where the science 'reproducibility crisis' began).

2. Approaches to Reproducible Computation

We describe several approaches to improve reusability and reproducibility for geocomputational work, and point to best practice from other disciplines as appropriate. (The presentation will elaborate on each of these headings and give examples relevant to geocomputation and spatial analysis.)

2.1. Seeing firsthand what was done (virtual witnessing)

At the most basic level, journals such as the Journal of Visual Experiments (JoVE, 2019) which share video recordings of experimental procedures, can be used to capture a live account of the process of investigation, that can be peer-reviewed and shared. It does not guarantee repeatability, but it gives an insight into the mechanics of conducting research that is often missing from more traditional publications.

2.2. Replicating the computational environment used

A further step towards repeatability is offered by virtualised computational infrastructure such as *Docker* (Docker, 2019) which offer a 'containerised' approach to supporting research. The container in this case is a place to store a computational image—a stack of software that might include an operating system, various databases, and application programs. This stack is created by serializing a working application running on a virtual machine. It has many advantages, (for example, it overcomes versioning and software integration issues for the new user) but chief amongst them for our purposes here is that the image can be moved to a completely separate virtual machine, in a different organisation or even country, where it can be opened, 're-imaged' and run in 2-3 minutes. It will behave exactly the same as the original software did, thus it provides a very convenient way to 'wrap-up' and share a complete software environment with new users. It is a mechanistic way to achieve some basic repeatability/refutability, and is mature enough now to be used reliably as part of a peer review process.

2.3. Creating a Library of reusable software environments

Perhaps the best example of the use of containerization for research is the *Nectar Research Cloud* (Nectar, 2019), developed and used in Australia for the last 4-5 years (and also used by the Centre for eResearch in Auckland). As well as making it easy for researchers to create and share experiments, it also contains a huge library of existing research software images that can be easily discovered and quickly restarted. For example, one can spin up a Hadoop cluster to conduct spatial data replication experiments in just 2 minutes. Nectar has greatly increased the amount of sharing amongst Australian researchers and has been shown to enable replicability and reproducibility (e.g. Kanwal et al, 2015) in terms of the software used (and avoid all the complexities and dependencies that typically plague software-reuse).

2.4. Virtual laboratories—encapsulating data, methods and processing

Of course, having access to an identical software configuration does not guarantee reproducibility or replicability, though it removes a traditionally difficult burden. But to fully replicate an analysis, the same data is also needed. A virtual Laboratory extends the idea of a Research Cloud by also including the data and macros that are used as inputs and control / conditioning elements in an analysis. The resulting environment provides a completely self-contained environment where many analytic activities become reliably repeatable, reproducible and refutable (apart from any non-deterministic methods that use randomization). An excellent example is the *Biodiversity and Climate Change Virtual Laboratory* (BCCVL, 2019) which supports some very sophisticated geospatial modelling, and visualisation, but in a controlled environment that essentially wraps together all of the tools, data, methods and scripts used in analysis so they can be shared within a community. BCCVL has become a vital resource for the biodiversity research community in Australasia. Similar Virtual Laboratories have been created for several other research communities, including Genomics, Marine and geophysics (see https://nectar.org.au/labs-and-tools/ for the full selection). Virtual Laboratories need sophisticated interfaces to allow new methods and datasets to be contributed, so that they can grow to encompass new analytical methods and new data opportunities. But to do so, the methods and data need careful curation, and in the case of methods, they must fit within specifically-designed templates in terms of how the connect together. This is a current research challenge.

2.5. Computational Workflows—flexibility for complex tasks

Where a community does not have an agreed set of methods or data, or indeed is actively developing new methods that do not easily fit into the templates used in a Virtual Laboratory, a more generic form of repeatability can be obtained using a Computational Workflow such as *Galaxy* (2019). Workflows completely describe all the analytical steps taken in an experiment or procedure, as a directed graph. They are more flexible than Virtual laboratories, in that they can create complex workflows with loops and hierarchies of analytical methods, but are also more complex to use. GeoVISTA *Studio* (Takatsuka & Gahegan, 2002) is an early example of a workflow environment for geographical analysis and visualisation.

2.6. 'Executable' Journals

Perhaps the holy grail of repeatability is a journal article that is itself an executable experiment—that describes an analysis in words, formulae and code, but also allows the analysis to be repeated by the reader. A good example is the *Physiome* journal (Physiome, 2019) that evaluates submissions "to determine their **reproducibility**, **reusability**, and **discoverability**. At a minimum, accepted submissions are guaranteed to be in an executable state that reproduces the modelling predictions in the primary paper, and are archived for permanent access by the community." The journal uses shared method libraries, common workflow descriptions and packaged data to come good on its ambitious claims.

3. Summary

All of these methods, by increasing levels of sophistication, record what was done in precise ways that can survive the process of sharing and enable researchers to reproduce the findings in a separate computational environment. However, none of them describe *why* specific choices were made by their originator, which remains an ongoing challenge (Gahegan & Adams, 2014).

4. References

Baker, M. 2015. Over half of psychology studies fail reproducibility test. *Nature News*.

BCCVL. 2019. The Biodiversity and Climate Change Virtual Laboratory: http://www.bccvl.org.au/

Docker. 2019. The Docker Container Platform. https://www.docker.com/

Gahegan, M. and Adams, B. 2014. Re-Envisioning Data Description Using Peirce's Pragmatics. Vienna Austria: 8th International Conference, GIScience 2014, 24-26 Sep 2014. (LNCS 8728): 142-158.

Galaxy. 2019. The Galaxy workflow engine. https://galaxyproject.org/learn/advanced-workflow/

JoVE. 2019. Journal of Visual Experiments. https://www.jove.com/

Kanwal, S, Lonie, A, Sinnott R O and Anderson, C. 2015. Challenges of Large-Scale Biomedical Workflows on the Cloud -- A Case Study on the Need for Reproducibility of Results. 2015 IEEE 28th International Symposium on Computer-Based Medical Systems (CBMS) (2015), Sao Carlos, Brazil, June 22-25, 2015, pp: 220-225,
Bookmark: http://doi.ieeecomputersociety.org/10.1109/CBMS.2015.28

Nectar. 2019. The Nectar Research Cloud. https://nectar.org.au/research-cloud/

Physiome. 2019. The Physiome journal. https://journal.physiomeproject.org/about.html

Takatsuka, M. and Gahegan, M. 2002. GeoVISTA *Studio*: a codeless visual programming environment for geoscientific data analysis and visualization. Computers and Geosciences 28(10):1131-1144 2002.