REPORT ON WORKSHOP: REPRODUCIBILITY AND REPLICABILITY IN GEOSPATIAL RESEARCH

March 2019

Background

A small and focused workshop on the topic of reproducibility and replicability in geospatial research (R&R) was held at Arizona State University (ASU) on February 11 and 12, 2019. It was organized under the auspices of the ASU Spatial Analysis Research Center (SPARC), with funding from Esri, SPARC, Dr Wenwen Li, and the ASU School of Geographic Sciences and Urban Planning (SGSUP). The workshop was organized by ASU Research Professor Michael Goodchild, ASU faculty members Stewart Fotheringham (Director of SPARC), Wenwen Li, and Peter Kedron, with the assistance of Nick Ray.

The workshop

In late August 2018 the organizers distributed the following invitation, using the mailing lists of the AAG's relevant specialty groups:

> "Replicability and reproducibility (R&R) have always been core requirements of scientific research. Recently cases of failure to replicate previously published findings in several areas of science have received widespread public attention. As research grows more complex, and increasingly reliant on data and software, it seems likely that concerns about replicability will grow rather than diminish. For example, different software packages may produce different results even when the same technique of spatial analysis is applied to the same data, or analysis results cannot be reproduced by the same software due to the lack of proper metadata or provenance documenting the spatial processing and parameters used. Moreover, there may be reasons why geospatial researchers need to be especially concerned about replicability; for example, when results from one geographic area fail to be replicated in other geographic areas. This small and focused workshop will address the following questions and related issues:
>
> 1) What forms of failure to replicate exist in the geospatial sciences? Can a formal framework be devised?
> 2) In what areas of geospatial research is the danger of non-replicability most severe?
> 3) What mechanisms can be used to avoid or minimize the danger of such failures?
> 4) Do we expect the results of model calibrations to be constant over space and if not, what are the implications for spatial analysis?
> 5) How should R&R be incorporated into the design and implementation of future spatial software?

6) How should students be made aware of these issues?

7) What follow-on activities might draw greater attention to these issues?"

A total of 26 participants attended the workshop: the four organizers, an additional six faculty members from ASU, two Esri staff (Dawn Wright and Kevin Butler), three special invitees (Shaowen Wang from the University of Illinois at Urbana-Champaign, Daniel Sui from the University of Arkansas, and Daniel Nüst from the University of Münster), and 11 selected from the 18 responses to the open call. The full list of participants is included below and is accessible on the meeting website https://sgsup.asu.edu/sparc/RRWorkshop. In addition an Open Science Framework project sharing the materials of the workshop is available, entitled **2019 Workshop "Replicability and Reproducibility in Geospatial Research" at SPARC**.

The Nüst presentation

After words of welcome from Trisalyn Nelson, Director of SGSUP, the workshop opened with a presentation by Daniel Nüst, one of the leaders of a comparable effort on R&R under the auspices of AGILE, the Association of Geographic Information Laboratories in Europe. He began with a general overview. A "replication crisis" has received widespread attention across the sciences, but perhaps nowhere as much as in psychology, where numerous attempts to replicate previous findings have failed (see, for example, Pexman and Lupker, 1995). He argued that scientists generally lack the relevant skills and tools to ensure that their findings are replicable, and that much academic literature amounts to little more than advertising of findings, rather than detailed reporting that would allow results to be reproduced and replicated. "Show me" should be more important than "trust me" in the culture of science.

Unfortunately terminology in R&R is far from standard. Daniel used "reproducible" to describe results that could be repeated using the same data and methods, and "replicable" to describe a higher level of scientific rigor in which results could be repeated using different samples of data and different software, and this distinction was maintained throughout the workshop. The geospatial case raises the question of whether results can also be repeated using different study areas; in other words, whether they are generalizable to other places.

Efforts to build a culture of open science, in which data, tools, methods, and software are all made accessible to everyone, are welcome. But openness in and of itself is not sufficient to ensure that results can be reproduced, let alone replicated. Stark (2018) proposed the term "preproducible" for research that is described in sufficient detail for someone else to repeat it, arguing that preproducibility should be an important criterion in the peer-review process.

Daniel then moved to a specifically geospatial perspective, and made four opening points. First, many geoservices are established and trusted, despite the frequent lack of full documentation. Second, geospatial data sets can be very large and computational requirements correspondingly heavy, making it very difficult for researchers to find the resources needed for replication. Third, much software used in geospatial research is proprietary and the user experience often amounts to little more than "point and click," giving the user little encouragement to look "under the hood" or "into the black box" of

the software. Default values of parameters are often accepted without full knowledge of their implications. Finally, maps are often used as the preferred method of publishing the results of geospatial research, yet often fall short of a full and open representation of those results; legends may be missing, and choices over parameters such as class intervals introduce an element of esthetics and subjectivity into what should be an objective reporting process. Konkol, Kray, and Pfeiffer (2019) have conducted extensive surveys of researchers in geoscience to explore the extent to which they understand the principles of R&R.

The final section of the presentation focused on approaches to science that might lead to greater replicability. Many technical approaches have been explored in recent years. *Containerization* attempts to structure software so that it can be readily packaged with the data, re-used across a wide range of platforms, and archived to be widely accessible. Docker is one popular example of software for containerization. Knoth and Nüst (2017) describe the use of Docker to containerize the software used in their study of object-based image analysis. A somewhat different approach focuses on a *research compendium*, which Gentleman and Temple Lang (2004) describe "both as a container for the different elements that make up the document and its computations (*i.e*. text, code, data, ...), and as a means for distributing, managing and updating the collection."

The presentation ended with a series of ideas that might help promote a research culture that is more sensitive to the need for replicability. Authors might be trained in the principles of preproducibility; awards might be given for stellar papers; special conference tracks might be organized for papers that represent best practices in R&R; and journal editors might provide reviewer guidelines that emphasize the principles of R&R.

The Kedron presentation

Peter Kedron began with four factors that together may have stimulated the "replication crisis" in science. Science is underpowered, forcing researchers to take shortcuts, reduce sample sizes, and make limiting assumptions; researchers are poorly incentivized, and forced to work in an environment that emphasizes the numerical aspects of publication rather than the quality and importance aspects; researchers may be unconsciously biased in the hunt for exciting new discoveries; and uncertainty underlies all data, all analyses, and thus all conclusions.

Unfortunately the definitions of reproducible and replicable that were given by Daniel Nüst and adopted during the workshop are exactly reversed in some disciplines—and political science and economics appear to make no distinction between the two terms. Peter urged adoption of Daniel's definitions, which he attributed to Claerbout, Donoho, and Peng (the *CDP* definitions). He elaborated them in the following table:

	reproducible	replicable
researcher	different	different
tools	same	similar

methods	same	similar
data	same	different
results	same	corroborating

To be reproducible the results produced by a different researcher using the same tools, methods, and data should be identical; on the other hand replicability implies that a different researcher using similar tools and methods, but different data, should obtain results that corroborate the original study.

The discipline of geography presents a somewhat distinct context for replicability, and hence an answer to the question "what can spatial science contribute to R&R?" The landmark Hartshorne-Schaefer debate of the 1950s focused on whether the results of geographic scholarship should be replicable over space—whether the results obtained in one study area should be corroborated in another. The nomothetic view advanced by Schaefer held that the discipline should indeed be focused on searching for such general principles, in emulation of sciences such as chemistry and physics for which geographic location (and time) has no relevance to discovery. The idiographic view promoted by Hartshorne implied that such generalization over space was unachievable for the phenomena studied by geographers: that all such phenomena were unique properties of places and should be studied on that assumption.

Today's discipline falls somewhere between these two positions, for two reasons. First, much of the geographic literature clearly implies the ability to generalize over space. Even though results may have been obtained from only one area, the conclusions may have been written without reference to that area. The Schaefer vision of a scientific geography persists strongly among many geographers, and there is clearly a need for greater attention to the question of geographic generalizability or replicability. Second, in recent years several techniques of spatial analysis have been developed that take a rather weaker view of generalization across space. Geographically weighted regression (GWR; Fotheringham, Brunsdon, and Charlton, 2002), for example, assumes that a single model applies everywhere in space, but allows the model's parameters to vary spatially. Such techniques of *place-based analysis* represent an intermediate position in the Hartshorne-Schaefer debate and the idiographic-nomothetic binary.

The Lightning Talks

After lunch and some welcome remarks from Dean Libby Wentz, a series of 11 lightning talks provided additional examples and allowed many of the participants to express themselves somewhat more formally. Each talk was limited to eight minutes and four slides. Some of the highlights of this session are described next.

Mark Gahegan began the session by arguing that the process of dissemination of scholarship had changed little in 300 years. Yet the nature of research had changed almost beyond recognition, especially in disciplines such as geography, suggesting a need to rethink the process of scholarly dissemination. He argued the case for the "executable" journal which would allow someone not only to read about the research but to reproduce it. He pointed to the Physiome Project (physiomeproject.org) as an example. He also

noted that many other words beginning with "re" were relevant to the R&R discussion (including relevance), and the group helped him to list eight.

Song Gao gave examples of common aspects of geospatial research that contributed to lack of replicability: kernel density estimation for which parameters were not reported; class intervals that were not documented; undocumented parameters used in GWR analysis; and unspecified provenance of data.

Jason Tullis recalled the work of David Lanter in the early 1990s (*e.g.*, Lanter, 1991), and the tools he developed to document data provenance. He noted that Esri's ModelBuilder is in many ways an outgrowth of that early work, in contributing to a greater degree of replicability.

Shaowen Wang described his CyberGIS project which brings GIS into the world of advanced cyberinfrastructure and high-performance computing. He noted the value of Jupyter Notebooks in their role of allowing researchers to track and document the stages of research.

John Wilson used the example of terrain analysis to illustrate the ways in which the many different algorithms for such simple tasks as viewshed analysis or hydrographic simulation contribute to a lack of replicability.

Dawn Wright wondered if concern for R&R could be regarded as a gamechanger in geospatial research. Would it help to make such research more resilient (another "re" word)? Would it help the community to address the threats that are emerging in response to the "replication crisis"?

Group discussions

The participants broke into groups to pursue these issues in more detail, and then each group reported back in a plenary session. The following points emerged:

- What institutions might make a difference in moving geospatial science into a new, more reproducible era? The major funding organizations—NSF and NIH—might require greater commitment to R&R by PIs. Universities might add R&R to the criteria used in faculty advancement. The learned societies, such as the American Association of Geographers, might also take a lead in advocating R&R among their members, and establishing awards for best practice.
- Some aspects of scientific practice are fundamentally in conflict with the objectives of R&R. They include the ownership of intellectual property, which might make it difficult for developers of software to make their methods fully open; and privacy, especially in research involving human subjects. In these and other ways the interests of individuals may conflict with the interests of the scientific community as a whole.
- Any push for R&R will require a degree of transformation of scientific practice, and a willingness to escape the legacy of the past. It would be very difficult to advance R&R with respect to prior research.
- Various aspects of practice present challenges to the objectives of R&R. They

include the existence of multiple versions of software and of data; advances in hardware that are backward incompatible; and the inadequacy of today's approaches to metadata and data provenance, which leave many important aspects undocumented and unspecified.

The Sui presentation

The second day opened with a presentation by Daniel Sui. He began by citing the work of Victoria Stodden, who has defined three distinct issues in reproducibility: empirical, statistical, and computational. By contrast, he suggested that published work in GIScience can be theoretical and conceptual and beyond the domain of R&R; technical and computational, and fall into Stodden's computational category; and concerned with applications of GIS, and thus straddle all three categories.

Daniel chose the arena of geodesign to illustrate the meaning of R&R. Like other uses of GIS, geodesign is a blending of the scientific and the esthetic. Its esthetic aspects are clearly outside the domain of R&R, echoing a point made earlier by Daniel Nüst about the use of cartography to present results. Replication in the esthetic world raises interesting issues, which Daniel Sui illustrated by reference to the "copycat cities" that are found in many parts of China. He commented that copycat replication is often regarded as a sound business strategy.

Story maps are a popular use of geospatial technology, as they allow the results of GIS analysis to be presented in compelling ways. But they raise their own issues of R&R because in simplifying results into a story they omit the essential details of preproducibility. Jonathan Phillips (2012) has argued that all stories in earth science follow one of eight possible plots.

A very early discussion of the special nature of replication in geography was provided by Wayne Davies (1968) in special reference to central place theory, a focus of much research in geography at that time. The paper raised many of the issues that have surfaced again in this workshop.

The Esri keynote

Dawn Wright and Kevin Butler provided an overview of the efforts Esri is making with respect to R&R. In the laboratory sciences the workbench is the hallmark of science, and the basis of claims to R&R. Dawn argued that in the geospatial sciences an experiment can be formalized as a workflow, and executed using a software workbench. In ArcGIS this can be accomplished using the ModelBuilder workbench. It allows the user to specify data sets and the operations to be performed on them, and in principle would allow an experiment to be reproduced, or replicated using the same procedure on an alternative data set. A ModelBuilder process can be exported in xml or as a Python script. Publishing a ModelBuilder process thus goes some way to achieving the objectives of R&R. Dawn's slides are available at esriurl.com/workbench.

Unfortunately new releases of software make this more difficult. Moreover it would be difficult to replace a ModelBuilder process with another vendor's representation of the same process, because what appears to be an identical GIS function may differ in many respects, a point illustrated earlier by John Wilson and Song Gao.

Dawn reviewed forthcoming developments from Esri. Hosted Python notebooks

integrate Open Science libraries with all types of data, the ArcGIS API for Python, and analytic servers, providing "a workbench for R&R within the world of open science." Containers are coming in ArcGIS 10.7.1 and ArcGIS Pro 2.6.

Dawn ended her presentation with her view of the hallmarks of R&R for geospatial research:

- Workflows must be shared alongside data, wherever possible;
- Workflows must be further amplified with use cases;
- Workflows and their associated use cases must be valued as much as journal articles and data sets;
- As a best practice, software producers should explicitly document as many aspects of their implementation of an algorithm as possible;
- When random-number generators are used in simulation, software producers should allow users to set fixed seeds, thus allowing exact reproduction of results.

Kevin provided a real-time demonstration of many of the Esri tools relevant to R&R. He argued that developers of technology must see it as their responsibility to address R&R issues, especially in capturing workflows and data provenance. He raised a very important point: since all geospatial data are subject to uncertainties of various kinds, must a result produced without attention to uncertainty be irreproducible by definition?

The Fotheringham summary

The formal part of the workshop program ended with some summary comments by Stewart Fotheringham. He opened by noting the confusion over the key terms reproducibility and replicability, as detailed by Peter Kedron, and argued that progress in this field will be necessarily impaired unless a more rigorous lexicon can be devised and accepted.

He noted the complexity of the technological solutions to R&R, as detailed by Daniel Nüst, and argued that such complexity would inevitably discourage interest by the average researcher. As with information technology in general, any solutions to the R&R dilemma must be easy to understand and use if they are to be widely adopted. Results obtained by different studies should be easy to compare. The issue may resolve to a simple matter of costs and benefits: what are the costs to the average researcher of addressing R&R using technical solutions, and what are the corresponding benefits? Adoption will not occur, except among specialists, unless benefits clearly exceed costs.

The privacy issue raises important concerns about R&R, especially with geospatial research involving humans. Data enclaves, such as the census data centers, offer a partial solution by rigorous control over access, and virtual data enclaves that achieve similar objectives using firewalls have promise.

Finally, Stewart returned to the question "what is special about spatial?" While many of the issues raised during the workshop apply to any research, the spatial case creates its own version of replicability, in the question of whether corroborating results can be obtained from distinct geographic areas. The concept of weak generalizability represents a distinct geospatial contribution to the R&R problem.

Final plenary

Several possible follow-on activities were suggested and discussed. A forum is being proposed to the *Annals of the American Association of Geographers*, to focus on some of the R&R issues that arise across the discipline of geography. The journal editors present, representing the *Annals*, *Transactions in GIS*, and *Urban Remote Sensing*, were urged to increase the attention to R&R in the information they provide to authors and reviewers. A short *Perspective* was suggested for *PNAS*, and a *Foresight* piece in *IJGIS*. Many of the materials generated before and during the workshop will be made available on the SPARC website and through the Open Science Framework.

The workshop adjourned after thanks were voted to the sponsors, and to Nick Ray for outstanding staff support.

Conclusions

To conclude, the following draws on the workshop discussions to address the seven original questions:

1) *What forms of failure to replicate exist in the geospatial sciences? Can a formal framework be devised?*

In the geospatial sciences it is failure to replicate across space (and time) that is most in need of a formal framework. A clear terminology is helpful, and the concept of weak generalizability clearly has profound implications for the geospatial sciences.

2) *In what areas of geospatial research is the danger of non-replicability most severe?*

If we think of research in the geospatial sciences as ranging from the theoretical and conceptual to the technical and statistical, then the danger is clearly most severe in the statistical, in applications of geospatial technology in areas such as remote sensing and the social sciences where it is common practice to test ideas in a few selected areas.

3) *What mechanisms can be used to avoid or minimize the danger of such failures?*

Technical research can do much to add to the tools available for support of R&R, and improved documentation and openness can clearly help.

4) *Do we expect the results of model calibrations to be constant over space and if not, what are the implications for spatial analysis?*

The principle of spatial heterogeneity implies that calibrations will not be constant over space. Moreover under-specification will lead to varying calibrations because missing variables will themselves be spatially heterogeneous. Thus there is a strong argument for weak generalizability, and the further development of place-based techniques.

5) *How should R&R be incorporated into the design and implementation of future spatial software?*

Many answers to this question were discussed in the workshop, and are summarized in parts of this report. Better documentation is perhaps the most pressing need.

6) *How should students be made aware of these issues?*

Education surfaced at several points in the discussion, but no simple answer to the question was proposed. Instead the answer to the next question appears to be the most promising path to follow.

7) *What follow-on activities might draw greater attention to these issues?*

Several follow-on activities were discussed. The editorial review process can be strengthened, and papers published in appropriate journals will certainly help. All of this essentially amounts to raising awareness throughout the geospatial sciences community.

List of participants

Vanessa Bastos	Arizona State University
Ling Bian	State University of New York at Buffalo
Kevin Butler	Esri
Dylan Connor	Arizona State University
Stewart Fotheringham	Arizona State University
Amy Frazier	Arizona State University
Mark Gahegan	University of Auckland
Song Gao	University of Wisconsin
Michael Goodchild	Arizona State University
Yingjie Hu	State University of New York at Buffalo
Bandana Kar	Oak Ridge National Laboratory
Peter Kedron	Arizona State University
Wenwen Li	Arizona State University
Alan Murray	University of California Santa Barbara
Soe Myint	Arizona State University
Trisalyn Nelson	Arizona State University
Daniel Nüst	University of Münster
Vaclav Petras	North Carolina State University
Nathan Piekielek	Pennsylvania State University
Alex Sorokine	Oak Ridge National Laboratory
Daniel Sui	University of Arkansas
Daoqin Tong	Arizona State University
Matt Toro	Arizona State University
Jason Tullis	University of Arkansas
Shaowen Wang	University of Illinois at Urbana-Champaign

John Wilson University of Southern California

Dawn Wright Esri

References

Davies, W.K.D., 1968. The need for replication in human geography: some central place examples. *Tijdschrift voor Economische en Sociale Geographie* 59(3): 145. DOI: 10.1111/j.1467-9663.1968.tb01703

Fotheringham, A.S., C. Brunsdon, and M. Charlton, 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ: Wiley.

Knoth, C. and D. Nüst, 2017. Reproducibility and practical adoption of GEOBIA with open-source software in Docker containers. *Remote Sensing* 9(3): 290. DOI: 10.3390/rs9030290

Konkol, M., C. Kray and M. Pfeiffer, 2019. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study, *International Journal of Geographical Information Science* 33(2): 408-429. DOI: 10.1080/13658816.2018.1508687

Lanter, D.P., 1991, Design of a lineage-based meta-database for GIS. *Cartography and Geographic Information Systems* 18(4): 255-261.

Pexman, P.M. and S.J. Lupker, 1995. Effects of memory load in a word-naming task: Five failures to replicate. Memory and Cognition 23: 581-95. DOI: 10.3758/BF03197260.

Phillips, J., 2012. Storytelling in Earth sciences: the eight basic plots. Earth Science Reviews 115(3):153-162. DOI: 10.1016/j.earscirev.2012.09.005

Stark, P.B., 2018. Before reproducibility must come preproducibility. Nature 557: 613. DOI: 10.1038/d41586-018-05256-0