Multiuser GIS and Practical Workflow Replicability

Jason A. Tullis, Professor of Geography
Department of Geosciences and Center for
Advanced Spatial Technologies, 321 JBHT
J. William Fulbright College of Arts & Sciences
University of Arkansas, Fayetteville, AR 72701
Phone 479.575.8784 Email jatullis@uark.edu

Replicability and interchange of workflows and provenance information in GIScience and remote sensing research is rarely efficient, even within highly collaborative settings dedicated to internal transparency. On the plus side, there is tremendous excitement about GIS supercomputing efforts that have produced advances in speed, reproducibility, and shared access to data. However, distributed replicability and workflow interchange problems are now more pressing given such advances as geospatial unmanned aircraft systems (UAS)-based discovery, where research data transfer requirements are eclipsing common bandwidths.

I have been fascinated by the history of the geospatial replicability crisis, and by the search for a greater technical understanding of the problem. In "Geoprocessing, Workflows, and Provenance" (*Remote Sensing Handbook*; Tullis et al., 2015) we review the history of the poorly understood concept of geospatial provenance (or lineage), and conclude that international standards aside, research is critically needed to demonstrate and elucidate *practical benefits* of provenance (including but not limited to improved replicability). David P. Lanter, who obtained an early (if not the first) GIS-related U.S. patent (Lanter, 1993) for his replicability-focused project *Geolineus*, is a leading co-author on the *Remote Sensing Handbook* section. In Tullis (manuscript), I identify common points of failure to replicate within the ArcGIS platform, and demonstrate how minor modifications to this software enable multiuser workflow replicability, interchange and provenance curation. This was achieved for hundreds of participating students, faculty, and staff using an enterprise geodatabase, local area network (LAN), and less than 500 lines of Python code.

Below I am pleased to partially address some of the excellent geospatial replicability and reproducibility questions posed by the workshop organizing committee. My responses are bracketed by a belief that the replicability crisis across all IT-intensive inquiry is largely based on what motivates people toward collaborative interaction or a lack thereof. (Technically speaking, the problem can mostly be solved; the question of whether enough geospatial scientists, software engineers, executive boards, etc. are sufficiently informed and want the problem solved it is certainly worth asking.)

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

In Tullis (manuscript) I identify a common multiuser GIS paradox, where extensive software engineering has gone into enterprise interchange, curation, and versioning of *data* with less concern for the multiuser requirements of *geoprocessing*, *workflows*, and management of *provenance* information. The latter are typically relegated as single user activities. There are notable exceptions, e.g., Google Earth Engine (Gorelick et al., 2017) which uses Git-based repositories! Unfortunately, the single user workstation paradigm still carries significant weight and in some circles is gaining momentum to address local processing needs (e.g., for UAS platforms that collect many GiB of data per minute). Any attempt to develop a formal framework should convey a careful understanding of geospatial provenance, which due to the unfortunate application of international standards (e.g., ISO, 2009) is often misunderstood as metadata when in fact it is a form of contextual metadata (Tullis et al., 2015; Gil et al., 2010). A formal framework should also incorporate the psychology of replicability. (Failure to replicate in GIScience may stem in part from a lack of interest in doing so. If someone feels methodologies should be closely guarded, then replicability is thwarted.)

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

As a lead author for the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report cycle, I see firsthand the importance of GIScience and remote sensing in a complex international endeavor. Unfortunately, without better replicability capacity, widespread understanding of and *trust* in complex processes

are thwarted. The danger is greatest where public trust in national and international institutions is critical to address problems of food security, natural and technological hazards, and misuse of human and natural resources, to name a few. *Provenance* is sometimes conflated with *trust* (Gil et al., 2010), so lack of provenance may similarly be conflated with lack of trust.

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

Careful presentation and visualization of provenance information can help both experts and more impressively non-experts to appreciate geospatial data value, quality, and potential (Del Rio and da Silva, 2007). Practical application of *multiuser* provenance information in the everyday GIS software paradigm can go a long way towards this end. Major changes in software architecture may not be needed (Tullis, manuscript), and careful integration of international Web standards such as PROV (Moreau and Missier, 2013) may first require geospatial scientists to see the practical value of such capacity (and thus influence software engineering).

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

The multiuser (and not single user) paradigm should prevail. The whole point of replicability is not just that we can replicate our own work (which is certainly a huge benefit given individual human nature and digital entropy.) The ability to access a colleague's full geospatial provenance (including all intermediate data) is more likely if the multiuser paradigm ensures that no file collisions (or broken paths) occur (Tullis, manuscript). On the technical side, a major challenge is software versioning over time. Open software and the possible use of curated virtual machines are two components that should inform the workshop.

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

Some traditional GIScience and remote sensing curriculum is a bit dated even for industry teams that use dated versions of the software. We need to be using the multiuser paradigm in the geospatial classroom, with less time spent troubleshooting learning management systems (LMS), and we should teach students about provenance and replicability. I have proposed to teach a new course next semester titled *Enterprise and Multiuser GIS*, with a significant component on provenance and replicability.

References cited

Del Rio, Nicholas, and Paulo Pinheiro da Silva. 2007. "Probe-It! Visualization Support for Provenance." In *Advances in Visual Computing*, 732–741. Lake Tahoe, NV: Springer.

Gil, Yolanda, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, and Paulo Pinheiro da Silva, eds. 2010. "The Foundations for Provenance on the Web." *Foundations and Trends in Web Science* 2 (2–3): 99–241.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*, Big Remotely Sensed Data: tools, applications and experiences, 202 (December): 18–27.

ISO. 2009. "ISO 19115-2:2009(E) Geographic Information - Metadata - Part 2: Extensions for Imagery and Gridded Data." International Organization for Standardization.

Lanter, David P. 1993. Method and Means for Lineage Tracing of a Spatial Information Processing and Database System. United States Patent and Trademark Office 5193185, filed May 15, 1989, and issued March 9, 1993.

Moreau, Luc, and Paolo Missier, eds. 2013. "PROV-DM: The PROV Data Model."

Tullis, J.A., J.D. Cothren, D.P. Lanter, X. Shi, W.F. Limp, R.F. Linck, S.G. Young, and T. Alsumaiti. 2015. "Geoprocessing, Workflows, and Provenance." In *Remotely Sensed Data Characterization, Classification, and Accuracies*, edited by P. Thenkabail, 1:401–21. Remote Sensing Handbook. Boca Raton, FL: CRC Press.