Multi-User and Collaborative Framework for Reproducibility and Replicability

Bandana Kar, Ph. D.
Research Scientist, Oak Ridge National Laboratory, Oak Ridge, TN
Phone: 865-576-3717; Email: karb@ornl.gov

Assessing the accuracy and correctness of scientific research is fundamental to science. With increase in data volume, availability of new technologies to gather data with help of public in real-time, and the rise of data science as a new paradigm, there is no shortage of scientific research. Contrary to this trend, there is an increased mistrust in science associated with the reproducibility and replicability of research (Baker 2016; Peng 2015). Fanelli (2018), however, suggested that the *reproducibility crisis* is not significant the way it has been portrayed. Whether the crisis exists or not, it is essential to understand the difference between reproducibility and replicability to undertake steps to increase trust in geospatial research.

Reproducibility is the ability of a researcher to produce the same result using the same data sets, methods, workflow and can draw the same conclusions if the study is repeated multiple times at different locations (Ostermann and Granell 2017; Leek and Peng 2015). Replicability means an independent study targeting the same scientific question will produce consistent results and conclusion (Peng 2015). Being a researcher with interest in geospatial analytics for emergency management that relies on conventional geospatial datasets as well as citizen derived data, I am interested in understanding implications of location privacy and data quality on reproducibility and replicability of geospatial research. Below I have provided my responses to the questions of replicability and reproducibility (R & R) of geospatial research from the perspective of data quality as they pertain to emergency management.

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

The failure to replicate geospatial research results from (i) Data, (ii) Methods, and (iii) Scientific Question at hand. Even with large volume of data at our disposal, specific data sets used for a research and metadata stating the provenance of the data are not always available. In case of emergency management related studies that require integration of citizen generated data with geospatial data sets, there is no standard or mechanism in place to ensure data interoperability and integration. It is also a challenge to access proprietary data sets. To address privacy, citizen generated data also tend to lack location information needed for replication. Licensing issues of proprietary software lead to using open source software that may not be similar to the methods used in proprietary software. Due to availability of large volume of data, significant number of studies are undertaken in emergency management. Rarely there is an understanding of the problem at hand. Very often what the decision-makers need to make the study usable are unknown to the researcher. The varying perspectives about the same problem magnifies the R & R issue.

To address this issue, a formal framework could be developed. In 2018, the Open Geospatial Consortium (OGC) conducted a Disasters Concept Development Study Workshop to develop standards focusing on data integration, automation and modeling in disaster related research. Such standards could be developed for other areas where geospatial data sets and analytics are extensively used. Nonetheless, transparency on the part of the researcher is crucial to improving R & R issues.

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

Geospatial data sets, techniques and technologies are predominantly used for emergency management activities. Given the necessity for real-time and/or near real-time response following a disaster event, agencies are relying on crowdsourced and citizen science derived data sets. More and more researchers, practitioners, citizens and industries are also getting involved in emergency management efforts. To protect their research and product portfolio, there is an inherent lack of transparency about data and methodology on the part of involved parties. For instance, OneConcern, a for-profit company implements AI for damage assessment in real-time. To maintain profit, the company is less likely to disclose data and methods.

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

Assessing the quality and provenance of data sets, and having access to metadata, methods and software are essential to ensuring R & R. Establishing standards will also ensure everyone involved in emergency

management research will follow the same protocol and guideline to document data sets, methodologies, and integrate and analyze data sets. A collaborative multi-user cyber-infrastructure framework similar to NSF's EarthCube could be established to allow community access to data sets, models and methodologies. Industry standards should be set in place to ensure private-public collaboration to address R & R issues.

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

Because of spatial dynamics and scaling issues associated with spatial analyses, and the issues discussed above contributing to R & R, the results of model calibration over space cannot be expected to be constant. From a spatial analysis perspective, this could exacerbate replicability problem and increase mistrust in scientific outcomes due to variable uncertainties. It will also lead to the development of different algorithms to account for spatial dynamics and scaling, thereby contributing to the reproducibility problem.

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

The design and implementation of future spatial software should enable incorporation of open source data sets and methodologies with existing software. The design should also follow a multi-user collaborative framework that could be deployed on the web to allow multiple users to access same data and methods. Newer version of a software should provide access to methods that were available in older versions or provide documentation how best to replicate unavailable methods.

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

The increasing pressure to publish and rise of inter-disciplinary research have contributed to some extent to R & R problems. Given the multi-disciplinary nature of geospatial data, courses offered in GIScience both in the academic sector and industries do not cover the issue of data quality and provenance, ethics and privacy, and reproducibility and replicability of research. To increase student awareness of R & R (i) courses should be offered on these topics, (ii) mentors and advisors should ensure their students understand causes and consequences of R & R and know the solutions to address the problem, and (iii) workshops should be offered by academic, public and private sector institutions on these topics.

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

The findings of this workshop should be presented to the broader research and practitioner communities as a white paper or a peer-reviewed publication, and through professional organizations like Association of American Geographers, American Geophysical Union, and American Society for Photogrammetry and Remote Sensing, etc. Participants of this workshop should come together in developing curriculum focusing on R & R that could be offered across academic institutions. There should be follow-on workshops on the topic in different venues to bring scientists involved in geospatial research from different disciplines.

References

1. Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533(7604):452-454.
2. Fanelli, D. 2018. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences* 115(11):2628-2631.
3. Leek, J. T., and Peng, R. D. 2015.Opinion: reproducible research can still be wrong: adopting a prevention approach: Fig 1. *Proceedings of the National Academy of Sciences* 112(6):1645-1646.
4. Ostermann F. O, and Granell, C. 2017.Advancing science with VGI: reproducibility and replicability of recent studies using VGI. *Transactions in GIS* 21(2):224-237.
5. Peng, R. 2015. The reproducibility crisis in science. *The Royal Statistical Society* 30 – 32.