Replicability and Reproducibility in High-Performance and Cloud Geocomputaions

Alexandre Sorokine (SorokinA@ornl.gov), Robert Stewart (stewartrn@ornl.gov); Oak Ridge National Laboratory

Importance of reproducibility and replicability (R&R) of geocomputations will grow with the increasing demand for evidence-based policy and decision-making in such areas as climate change prediction impact assessment, urban infrastructure development, energy, and many others. Geocomputations in high-performance (HPC) and cloud environments typically involve the use of large datasets and unique computing platforms that present significant challenges for achieving effective R&R of such studies.

In geocomputations the eventual goal of R&R can be viewed as the ability to correctly apply proposed algorithm or workflow to a new data and run it on a system other than the one used in the original study. This involves (1) access to the original source code under a permissive license for inspection and understanding, (2) availability of a compatible hardware platform on the side of the reviewer, (3) ability to reproduce the software environments including compilers and third party libraries, (4) access to the data that was used to reach the original results and make original claims, and (5) the ability to use the proposed workflow or algorithm to achieve meaningful results using another data source and/or geographic area.

Not all of these requirements are easily achievable at the current state of technology development. It is relatively easy to make the source code available to a reviewer or other scientists by using standard open source software development practices and version control systems. However, the source code may not be made open if the research is done by a commercial entity or falls under export control. Availability of the compatible hardware is trivial for standard systems like desktops yet in HPC innovative studies often require unique or rare equipment. Such computing capabilities typically cannot be replicated on commonly available systems. Examples of such platforms are leadership supercomputers like TITAN and SUMMIT at the Oak Ridge National Laboratory, specialized visualization facilities, or custom hardware accelerators including quantum computers. Among other R&R challenges in the HPC domain is the limited bit-wise repeatability of parallel computations and studies like global climate models that require so much computing power that repeating them is too costly.

Reproducing software environments used to be a very challenging problem. Research software often uses very fragile environments dependent on rare compiler versions and libraries. Recent progress in containerization (docker, singularity, etc.) has mitigated this problem but effective containerization on specialized hardware like still remains an issue.

There is a number of R&R challenges specific to geocomputations. Many of them are related to the data as most of the studies in this field use very diverse and very large geodatasets. This is an especially difficult problem for cloud-based analytics where researchers have to work with multiple data vendors and providers. Not all data providers and data hosting sites have sufficient data curation and data preservation strategies. Often the data cannot be redistributed or made public due to licensing restrictions, security or privacy concerns. Datasets sometimes completely disappear from the Internet or receive unannounced updates to the data or APIs that makes their reuse hard or impossible. The problem is exacerbated by the typically large size of the data requiring significant efforts to make and maintain their full updated copy on the researchers' side. In addition, as of today there are no common and effective tools for data version control similar to the ones that exist for the source code (e.g., git, fossil).

We have examined several strategies for data-level R&R in high-performance and cloud-based geocomputations. One of such strategies involves providing a user with a managed data store and cloud-side analytical and processing capabilities. This architecture was implemented in ORNL World Spatiotemporal Analytics and Mapping Project (WSTAMP, [1]). In such system a user has access to a large curated dataset that can be used with the cloud-based analytical functionalities. This strategy is very effective at saving researchers' time and effort spent on data copying, preparation, and processing. It simplifies R&R by keeping all the data managed by the provider and making it readily available for shared analytical workflows and procedure. Similar approach is used by Google Earth Engine. However, this strategy limits researchers' ability to use the data from other sources as it has to be converted and put under the management of the provider. To address this problem currently we are working now on another approach that is being implemented within Energy-Water Nexus Knowledge Discovery Framework (EWN-KDF, [2]) in which the users will be able to incorporate datasets available through online services in their cloud-based workflows. The software behind EWN-KDF will transparently create reusable archival copies of the geodata subsets used in the workflows and make them available for R&R.

Based on our experience in numerous projects we conclude that R&R is critical for the future of geocomputing as a viable and reliable scientific discipline. However, its practical implementation presents a number of challenges and imposes significant burden on the researchers. R&R should not be only encouraged but built-in into the future geoprocessing systems and workflows.

Acknowledgements
This manuscript has been authored by employees of UT-Battelle, under contract DE-AC05-00OR22725 with the US Department of Energy. The authors would also like to acknowledge the financial and intellectual support for this research by the Integrated Assessment Research Program of the US Department of Energy's Office of Science, Biological and Environmental Research and by the Department of Energy Office of Policy. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] Stewart, Robert, Jesse Piburn, Alexandre Sorokine, Aaron Myers, Jessica Moehl, and Devin White. 2015. “World Spatiotemporal Analytics and Mapping Project (WSTAMP): Discovering, Exploring, and Mapping Spatiotemporal Patterns across the World’s Largest Open Source Data Sets.” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2 (4): 95.

[2] Bhaduri, Budhendra L., Ian Foster, Varun Chandola, Bob Chen, Jibonananda Sanyal, Melissa Allen, and Ryan McManamay. 2017. “Energy-Water Nexus Knowledge Discovery Framework.” In *AGU Fall Meeting Abstracts*.