REPLICABILITY AND REPRODUCIBILITY IN GEOSPATIAL RESEARCH: A SPARC WORKSHOP

DIGITAL TERRAIN MODELING

John P. Wilson
Spatial Sciences Institute
University of Southern California

Great strides have been made in geomorphometry during the past 50 years spurred on by new sources of digital elevation data, the increasing use of theory to guide digital terrain modeling workflows, the specification of many new land surface parameters, the identification and extraction of landforms and other land surface objects, the improving characterization of error and uncertainty, and the development and sharing of new computer code to facilitate and support digital terrain modeling workflows. The rate of development has surged during the past 15 years, motivated by pressing environmental challenges at the topo-scale, and the computational resources that are now available to support digital terrain modeling applications.

The calculation and use of land surface parameters constitute the heart of geomorphometry as we know it today. There is now more than 100 primary and secondary land surface parameters in common use (Wilson, 2018). The majority are primary parameters derived from square-grid DEMs which measure site-specific, local or regional characteristics without additional input. The secondary parameters are derived from ≥2 primary parameters and additional inputs in some instances, and focus on water flow/soil redistribution or energy/heat regimes. Many of these land surface parameters incorporate flow direction and therefore make use of one or more of the 24 flow direction algorithms have been proposed during the past 33 years. Several of the newer algorithms combine square-grids and triangulated irregular networks (TINs) to avoid the shortcomings associated with square-grid DEMs and to take advantage of the additional discretization provided by TINs. An increasing number and variety of computer programs (i.e., the ArcGIS, GRASS, QGIS, SAGA, TauDEM, and Whitebox platforms) are available to calculate flow direction and flow accumulation as well.

Some applications have focused on the extraction and classification of landforms and land surface objects. Fuzzy classification methods have featured prominently in these applications because many of the landforms and land surface objects that are of interest have fuzzy boundaries. Several have attempted to automate and extend Hammond's (1964) map of repeating landform patterns for the conterminous U.S. to the globe (Karagulle et al., 2017), and Dragut and Eisank (2017) have borrowed concepts from remote sensing and data science to first segment DEMs and then classify the objects to avoid the problems of working directly with the DEM grid cells when extracting and classifying landforms and other land surface objects.

However, it is difficult to assess the efficacy of the aforementioned flow direction algorithms, and their impact on flow accumulation and the other land surface parameters which incorporate them. The most popular approaches have relied on four geometrical shapes for which the flow directions are known (e.g., Zhou and Liu, 2002) and/or comparisons of the performance of two or more of the aforementioned flow directions algorithms in specific landscapes (e.g., Pilesjö and Hasan, 2014; Wilson et al., 2007). A recent study by Buchanan et al. (2014) illustrates the enormity of the challenge here. This study calculated topographic wetness using >400 unique approaches that considered different DEM resolution, the vertical precision of the DEM, flow direction and slope algorithms, smoothing versus low-pass filtering, and the inclusion of relevant soil properties to compare the resulting topographic wetness maps with observed soil moisture in agricultural fields.

This state-of-affairs therefore demonstrates both of the outcomes noted in the call that: (1) different software tools may produce different results even when the technique of spatial analysis (i.e., the same flow direction approach in this instance) 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; and (2) there is every reason to believe that geospatial researchers will need to be especially concerned about replicability because it is highly likely that the results generated in one geographic area will not be able to be replicated in other geographic areas.

The rapid emergence of the web and all this entails (i.e., web portals to share geospatial datasets, provision of software as a service, etc.) coupled with advances in our knowledge and understanding of error and uncertainty and how these concepts can be used to clarify the 'fitness-for-use' of digital terrain modeling tools and data for specific applications bring both provenance and credibility to the fore. These elements have traditionally been handled by metadata in the geospatial sciences and it will be important for the geomorphometry community to adopt and use metadata to describe digital terrain modeling methods and datasets moving forward. However, these metadata may be necessary but not sufficient in their current form, and several recent approaches might be used to address at one of more of three of the questions and related issues which will be addressed in the upcoming SPARC workshop, as noted below.

The first pair of questions asked what forms of failure to replicate exist in the geospatial sciences and whether or not a formal framework could be devised. Qin et al. (2016) recently used case-based formalization and reasoning methods to acquire 'application-context' knowledge that may help to address these issues. They selected 124 cases of drainage network extraction (50 for evaluation and 74 for reasoning) from peer-reviewed journal articles and used these cases to determine the catchment area threshold for extracting drainage networks. This approach could be applied to many of the challenges currently encountered in digital terrain modeling workflows.

The answer to the fourth question, which asked whether or not we expect the results of model calibrations to be constant over space, is almost certainly negative for both flow direction and flow accumulation. The classification of landforms and other kinds of land surface objects and the use of these classes to test the performance of the 24 flow direction approaches would afford one way to address this challenge and manage the implications for spatial analysis. Wilson et al. (2007), for example, divided the landscape into a series of fuzzy landform classes and examined how the performance of six flow direction approaches varied across these classes. The recent work of Dragut and Eisank (2017) and Karagulle et al. (2017) provide more robust landform classification methods that could be used with this approach and the TerraEx application recently proposed by Netzel et al. (2014) provides a freely available, full service web application to locate landscapes that are similar to user-selected queries and doubles as a convenient portal to support the distribution of 3 arc-second DEMs and global maps of geomorphons and terrain relief.

The fifth question, which asked how replicability and reproducibility should be incorporated into the design and implementation of future spatial (i.e., digital terrain modeling) software, could be pursued using the best practices for documenting and sharing research from data to software to provenance recently proposed by Gil and colleagues (e.g., Garijo et al., 2013; Gil et al., 2016; Stodden et al., 2016).

And finally, the rapid advances in computational power and changing models of computing (i.e. cloud computing, cyberinfrastructure, interoperability, software-as-a-service) may offer new opportunities to develop new analytical approaches, clarify and strengthen the role of theory, and expand the geographic extent and heft of digital terrain modeling tools. For example, Survilla et al. (2016) recently developed a scalable high performance topographic flow direction algorithm that eliminated the bottleneck caused by flow direction, one of the most computationally intensive functions in the current implementation of TauDEM (Tarboton, 1997). This essentially local operation was transformed into a global operation to route flow across flat regions, by first identifying the flat areas and then using this information to reduce the number of sequential and parallel iterations needed to calculate flow direction. Qin et al. (2017), on the other hand, proposed an efficient solution to calculate the theoretically-based differential equation for specific catchment area proposed by Hutchinson and Gallant (2011) from gridded DEMs for small- and moderate-sized catchments that would avoid most if not all of the uncertainties associated with the calculation of the 24 flow direction approaches proposed thus far altogether.

References Cited

Buchanan, B.P., Fleming, M., Schneider, R.L., Richards, B.K., Archibald, J., Qiu, Z., Walter, M.T. (2014). Evaluating topographic wetness indices across central New York agricultural landscapes. *Hydrology & Earth Systems Science, 18,* 3279-3299.

Drăguţ, L., Eisank, C. (2011). Object representations at multiple scales from digital elevation models. *Geomorphology, 129,* 183-189.

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P.E., Gil, Y. (2013). Quantifying reproducibility in computational biology: The case of the tuberculosis drugome. *PLoS ONE 8*(11): e80278.

Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., Karlstrom, L., Lee, H., Mills, H.J., Oh, J.-H., Pierce, S.A., Pope, A., Tzeng, M.W., Villamizar, S.R., Yu, X. (2016). Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth & Space Science, 3*(10), 388-415.

Hammond, E.H. (1964). Analysis of properties in land form geography: An application to broad-scale land form mapping. *Annals of the Association of American Geographers, 54,* 11-19.

Hutchinson, M.F., Gallant, J.C. (2011). A differential equation for specific catchment area. *Water Resources Research, 47,* 95335.

Karagulle, D., Frye, C., Sayre, R, Breyer, S., Aniello, P., Vaughan, R., Wright, D. (2017). Modeling global Hammond landform regions from 250 m elevation data. *Transactions in GIS, 21*(5), 1040-1060.

Netzel, P., Jasiewicz, J., Stepinski, T.F. (2016). TerraEx: A geoweb app for world-wide content-based search and distribution of elevation and landforms data. In *Proceedings of the 9th International Conference on Geographic Information Science.* Montreal, Quebec.

Pilesjö, P., Hasan, A. (2014). A triangular form-based multiple flow algorithm to estimate overland flow distribution and accumulation on a digital elevation model. *Transactions in GIS, 18,* 108-124.

Qin, C.-Z., Ai, B.-B., Zhu, A.-Z., & Liu, J.-Z. (2017). An efficient method for applying a differential equation to deriving the spatial distribution of specific catchment area from gridded digital elevation models. *Computers & Geosciences, 100,* 94-102.

Qin, C.-Z., Wu, X.-W, Jiang, J.-C., Zhu, A.-X. (2016). Case-based formalization and reasoning method for knowledge in digital terrain analysis: Application to extracting drainage networks. *Hydrolrology & Earth System Sciences, 20,* 3379-3392.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240-1241.

Survila, K., Yildirim, A.A., Li, T., Liu, Y.Y., Tarboton, D G., Wang, S. (2016). A scalable high-performance topographic flow direction algorithm for hydrological information analysis. In *Proceedings of the 5th Annual Extreme Science & Engineering Discovery Environment Conference.* Miami, Florida.

Tarboton, D.G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research, 33,* 309-319.

Wilson, J.P. (2018). *Environmental applications of digital terrain modeling.* Wiley, Oxford, UK.

Wilson, J.P., Lam, C.S., Deng, Y.X. (2007). Comparison of performance of flow routing algorithms used in geographic information systems. *Hydrological Processes, 21,* 1026-1044.

Zhou, Q., Liu, X. (2002). Error assessment of grid-based flow routing algorithms used in hydrological models. *International Journal of Geographical Information Science, 16,* 819-842.