A Provenance-Based Model for Facilitating Replicability and Reproducibility in Spatial Analysis

(Position Paper)

Song Gao

Department of Geography, University of Wisconsin, Madison, USA

Email: song.gao@wisc.edu

Replicability and reproducibility are important in research community across different teams and institutions for solving scientific problems. According to the Association for Computing Machinery (ACM) definitions, "*replicability* means that an independent group can obtain the same result using the author's own artifacts, while *reproducibility* means that an independent group can obtain the same result using artifacts which they develop completely independently." The "artifacts" can be software, computer scripts used to run experiments, input datasets, or analysis tools, and so on. In geographic information systems (GIS), there exist hundreds of spatial analysis tools that can help reveal spatial relationships, patterns, and trends (Longley et al. 2015). However, those spatial analysis tools in a software system might be developed by referring different original methodology papers that may have different assumptions or at least need a calibration or parameter setting process rather than just using the default one. For instance, GIS users may need to understand that there are many data classification methods such as natural breaks, quantile, equal interval, and geometrical interval when they want to create a choropleth map for an attribute (de Smith et al. 2007). Users may get varying choropleth maps by using the same data layer but with different data classification methods. When another user interprets the output maps, the user may get different insights or even contradictory results with the same attribute data. This might also be the case for "heat maps" when users choose different kernel configurations for generating spatial kernel density maps. Another example is using the geographically weighted regression (GWR) analysis (Brunsdon et al. 1996), a user may need to understand the spatial stationarity assumption (Fotheringham 2009) and different types of weighting scheme (i.e., fixed or adaptive) or kernels can be configured during the analysis for exploring the spatial variability of relationships between the dependent variable and independent variables.

Therefore, the inclusion of parameter settings in the provenance information (e.g., metadata, or output XML file) from which the spatial analysis results and maps are generated or derived, could be one of the key elements for facilitating replicability and reproducibility in geospatial studies. More specifically, a PROV Ontology (PROV-O, 2013) that defines the OWL Web Ontology Language encoding of the provenance data model has been developed from the W3C community. Research needs to be further conducted to investigate whether the PROV-O could be a potential candidate for ontology engineering and GIS software development to improve the replicability and reproducibility in spatial analysis.

In summary, it should be a collaborative effort among researchers, instructors, students, engineers, software developers, and practitioners for facilitating the replicability and reproducibility in geospatial studies. A provenance-based model in spatial analysis might yield a first-step towards this direction in practice.

References:

Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: https://www.acm.org/publications/policies/artifact-review-badging

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. Geographical analysis, 28(4), 281-298.

de Smith, M. J., Goodchild, M. F., & Longley, P. (2007). Geospatial analysis: a comprehensive guide to principles, techniques and software tools. Troubador Publishing Ltd.

Fotheringham, A. S. (2009). “The problem of spatial autocorrelation” and local spatial statistics. Geographical analysis, 41(4), 398-403.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (4th Edition) (2015). Geographic information systems and science. John Wiley & Sons.

PROV-O (2013): The PROV Ontology: https://www.w3.org/TR/prov-o/