CyberGIS Capabilities for Scalable Spatial Data Synthesis

Anand Padmanabhan, Shaowen Wang, and Jing Gao
CyberGIS Center for Advanced Digital and Spatial Studies
University of Illinois at Urbana-Champaign

CyberGIS – geographic information science and systems (GIS) based on advanced cyberinfrastructure (CI) – has recently emerged as a vibrant interdisciplinary field (Wang 2010). It comprises a seamless integration of advanced CI, GIS, and spatial analysis and modeling capabilities to enable computing- and data-intensive research and education across a broad range of academic disciplines with significant societal impacts (Wright and Wang 2011). Previous works show scientific problems exhibiting high geophysical and spatial complexity require holistic approaches to integrating rich spatiotemporal data, analytics and models for complex problem-solving (Wang *et al.* 2013). CyberGIS has a compelling role to play to enable such holistic approaches particularly as the role and complexity of big data continues to challenge our ability to support collaborative, interactive, and scalable knowledge discovery. CyberGIS has already had significant impact in a number of domains (e.g., hydrology and water resources, complex coupled human-natural systems, emergency management, econometrics, geophysical sciences, and urban studies) through processing of complex and massive amounts of geospatial data and performing associated analysis, simulation, and visualization (e.g., Anselin and Rey 2012, Wang and Zhu, 2008).

CyberGIS software environment established by the ongoing NSF CyberGIS software initiative (www.cybergis.org) has three interrelated pillars: 1) CyberGIS Gateway for a large number users to simultaneously access online cyberGIS analytics and services, 2) CyberGIS Toolkit for distributing open source and scalable modules, and 3) GISolve middleware for integrating CI and cyberGIS capabilities (Liu *et al.* 2014, Wang *et al.* 2005, Wang *et al.* 2013, Wang and Liu 2009). CyberGIS Gateway provides transparent access to advanced CI resources and services including the NSF XSEDE (https://www.xsede.org/), the Open Science Grid (http://www.opensciencegrid.org/), and clouds. Through friendly user interfaces, the Gateway makes CyberGIS capabilities accessible for various research and education purposes. CyberGIS Toolkit, on the other hand, is targeted at advanced users and provides access to scalable geospatial software capabilities within advanced CI environments such as the NSF XSEDE. It is composed of a set of loosely coupled components to focus on exploiting high-end computing resources. GISolve is the leading spatial middleware that integrates advanced computing and information infrastructure with geographic information system capabilities for computationally intensive and collaborative geospatial problem solving. The purpose of GISolve is to establish user-friendly and spatially intelligent capabilities for performing computationally intensive geospatial data analysis and collaborative problem solving, and help non-technical users directly benefit from accessing advanced CI capabilities. It is openly accessible via a suite of open service application programming interfaces, and has been a key enabler for the CyberGIS Gateway.

CyberGIS software environment – collectively with thousands of users – provides tremendous opportunities to gain dynamic insight into complex phenomena from massive spatial data, collected from numerous sources that are increasingly used to instrument our natural, human and social systems at unprecedented scales. Though such big data streams play crucial roles in many scientific domains (e.g., ecology, geography and spatial sciences, geosciences, and social sciences, to name just a few) and promise to enable a wide range of decision-making practices with significant societal impacts, exploiting them successfully poses significant challenges. On one hand, spatial and location attributes serve as a common key to many types of data such as census and population, land use and cover, flood plain, and vegetation distribution. Oftentimes perceived as significant benefits, spatial data synthesis can be used to link disparate pieces of data that pertain to common spatial references and units. On the other hand, however, there are diverse spatial references and units for data collection and management and they are based on different representation models and assumptions. Furthermore, the quality, validity, and applicability of spatial data synthesis are dependent on scalable and timely discovery, correlation, and transformation of various sources of data. Hence, synthesis of various spatial data – a foundational process of various scientific problem-solving workflows – has become increasingly difficult and is not scalable to the significant size, complexity and diversity of spatial data.

To break through these challenges, it is important to establish a suite of cyberGIS capabilities (see for example the following two representative types) for scalable spatial data synthesis.

- *Interactive cyberGIS analytics* is important to many science applications as GIS-based scientific problem solving is normally conducted through map-based interactive user interfaces. Because such analytics must be highly responsive to user requests (e.g., real-time or near real-time) and often deal with unstructured spatial data sources (e.g., social media), they pose challenges in data-driven analytics and visualization;
- *Spatial decision-making* is widely used in many scientific domains (e.g., urban infrastructure planning, environmental sustainability, political redistricting, and energy supply chain optimization) and often requires what-if scenarios to be analyzed synthesizing different datasets in a near-real time fashion together with associated uncertainty analysis. Consideration of spatial uncertainty propagation (i.e. how uncertainty in model inputs contributes to uncertainty in optimization results) and estimation of relative influence of inputs' variation on results need to be estimated (e.g., using Monte Carlo simulation) and communicated to users (e.g., using online visualization).

Challenges and Opportunities of Spatial Data Synthesis

Spatial data synthesis represents two distinct but interrelated challenges: data aggregation and data integration. Data aggregation refers to the problem of bringing together various data streams at scale, while data integration concerns the challenges of harmonizing diverse (in format, type, spatial reference, spatial units, etc.) spatial datasets. Though the data integration and aggregation challenges are fundamental problems in their own right, they often need to be addressed holistically and simultaneously in scientific workflows. We illustrate several data synthesis challenges through a scientific use case described in the following section, but these challenges are common across many related science domains.

Conventional GIS approaches and software, cannot effectively synthesize big spatial data. Existing spatial data repositories are often "information silos" (Andrade *et al.* 2011) with no easy ways to synthesize data across them. With massive and dynamic spatial data now available from multitude of sources and sensors this problem is increasingly exacerbated. It is natural for scientists to want to define experiments customized to answer specific questions they are working on, using carefully selected sets of spatial data coming from a variety of sources. However the level of dynamicity and the variety of such data sources translates into undesirable uncertainty of the analysis results generated and leads to unpredictable requirements for computational support. Due to the dynamic nature of many data sources, and the user-driven nature of data synthesis, this process requires an always-on, potentially highly available, computational support as the produced data are filtered, reduced, correlated, processed, and stored. With the dramatic upward trend in the quantity and varied quality of spatial data that are forecast to continue, conducting scientific study dependent on big spatial data will become unrealistic due to significant gaps of existing spatial data synthesis capabilities. Furthermore, the quality, validity, and applicability of such spatial data synthesis are dependent on scalable and timely discovery, correlation, and transformation of various sources of data.

Scientific Use Case

In a typical CyberGIS workflow, input data are needed from multiple domains and sources. Consider a use case that focuses on data-driven cyberGIS-enabled approaches to understanding urban sustainability. This would require characterizing urban land change and modeling coupled human and environmental systems. A National Research Council (NRC) report titled "Advancing Land Change Modeling: Opportunities and Research Requirements" (NRC 2014), identified better aligning modeling schemes with the goals of their applications, better integrating Land Change Models (LCMs) with available data and models, improving and disseminating model evaluation and uncertainty assessment as key areas need immediate attention for further development. To resolve these issues from the data perspective, spatial data synthesis capabilities that offer additional information for facilitating scalable data exploration and knowledge discovery, and incorporate uncertainty assessment as part of data products are needed.

Approaching this study would require accessing large volumes of spatial data, such as high-resolution LiDAR data from the US Geological Survey (USGS) that would need to be used for extraction of land change features. The 3D Elevation Program (3DEP) initiated by USGS, as a result of the National Enhanced Elevation Assessment (NEEA) study, estimates to generate $1.2 billion to $13 billion in new benefits annually with nationwide high-resolution elevation data (Snyder 2012). The expected data volumes are of the magnitude of 8.2-9.4 petabytes, considering just the LiDAR point cloud, intensity signals, and bare Earth elevation model, for covering the US. While the science study is expected to greatly benefit from the high-resolution topographic data, the size and complexity of such data present unprecedented challenges for synthesis with other related data. Additionally, complementary data sources such as population, climate, and land cover over multiple spatial and temporal scales will also need to be integrated.

CyberGIS-enabled spatial data synthesis is expected to enable a large number of users from multiple scientific communities for solving scientific problems with broad and significant impacts. The National Data Service (NDS) provides an open platform on which novel cyberGIS capabilities for scalable spatial data synthesis can be created and evolved by leveraging distributed data resources and services to lower the barriers of storing, searching, accessing, and publishing enormous amounts of geospatial data. Meanwhile, a number of geospatial science communities are expected to contribute rich geospatial data and services to NDS through innovative cyberGIS capabilities for scalable spatial data synthesis.

References

Anselin, L. and S.J. Rey, *Spatial Econometrics in an Age of CyberGIScience*. International Journal of Geographical Information Science, 2012. **26**(12): p. 2211-2226.

de Andrade, F. G., C. d. S. Baptista, and F. L. Leite, *Using Federated Catalogs to Improve Semantic Integration among Spatial Data Infrastructures*, Transactions in GIS, vol. 15, pp. 707-722, OCT 2011 2011.

de Man, W. H. E., *Thinking Outside the Disciplinary Box in Coping with Dilemmas in Geoinformation Management for Public Policy*, Transactions in GIS, vol. 17, pp. 452-462, JUN 2013.

Diaz, L., A. Broering, D. McInerney, G. Liberta, and T. Foerster, *Publishing sensor observations into Geospatial Information Infrastructures: A use case in fire danger assessment*, Environmental Modelling & Software, vol. 48, pp. 65-80, OCT 2013.

Gong, P. *Integrated Analysis of Spatial Data from Multiple Sources: An overview*, Canadian Journal of Remote Sensing, 1994. 20: 349-359.

Liu, Y.Y., A. Padmanabhan, and S. Wang, *CyberGIS Gateway for Enabling Data-Rich Geospatial Research and Education*. Concurrency and Computation: Practice and Experience, 2014. http://dx.doi.org/10.1002/cpe.3256.

National Research Council. *Advancing Land Change Modeling: Opportunities and Research Requirements*. Washington, DC: The National Academies Press, 2014.

Schade S., and P. Smits, *Why Linked Data Should Not Lead to Next Generation SDI,* in 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), ed New York: IEEE, 2012, pp. 2894-2897.

Snyder, G.I., *National Enhanced Elevation Assessment at a glance: U.S. Geological Survey Fact Sheet 2012-3088*, US Department of the Interior, US Geological Survey 2012.

Stewart R. N., and S. T. Purucker, *An Environmental Decision Support System for Spatial Assessment and Selective Remediation*. Environmental Modelling & Software, 2011. 26: p. 751-760.

Wang, S., *A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis.* Annals of the Association of American Geographers, 2010. **100**(3): p. 535 **-** 557.

Wang, S., L. Anselin, B. Bhaduri, C. Crosby, M.F. Goodchild, Y. Liu, and T.L. Nyerges, *CyberGIS Software: A Synthetic Review and Integration Roadmap.* International Journal of Geographical Information Science, 2013. **27**(11): p. 2122-2145.

Wang, S., M.P. Armstrong, J. Ni, and Y. Liu. *GISolve: A Grid-Based Problem Solving Environment for Computationally Intensive Geographic Information Analysis*. in *Proceedings of the 14th International Symposium on High Performance Distributed Computing (HPDC-14) – Challenges of Large Applications in Distributed Environments (CLADE) Workshop, IEEE Press*. 2005.

Wang, S. and Y. Liu, *Teragrid Giscience Gateway: Bridging Cyberinfrastructure and GIScience.* International Journal of Geographical Information Science, 2009. **23**(5): p. 631 **-** 656.

Wang, S. and X.-G. Zhu, *Coupling Cyberinfrastructure and Geographic Information Systems to Empower Ecological and Environmental Research.* BioScience, 2008. **58**(2): p. 94-95.

Wright, D.J. and S. Wang, *The Emergence of Spatial Cyberinfrastructure.* Proceedings of the National Academy of Sciences, 2011. **108**(14): p. 5488-5491.