Geographic Data Science

Gennady Andrienko and Natalia Andrienko ■ *Fraunhofer Institute IAIS and City, University of London*

Robert Weibel ■ *University of Zurich*

Data science is an emerging area of work concerned with the task of extracting useful information and gaining insight from large data collections. Methods that scale to big data in terms of volume, variety, velocity, and veracity are of particular interest in data science. Data science methods and approaches address all stages of transition from data to knowledge and action, including data acquisition, cleaning and processing, information extraction, integration and representation, data analysis, and knowledge extraction and explanation.

Visualization of data is essential for human understanding of the subject under analysis, analytical reasoning about it, and generating new knowledge. Interactive visual interfaces support the human cognitive processes by allowing analysts to look at a subject from different perspectives and at different scales and levels of detail, link diverse pieces of information, and direct and control the work of computational analytical tools. Therefore, visual analytics approaches play an important role in data science.

More specifically, geographic data science deals with data that incorporates spatial and, often, temporal elements. In 2010, Gennady Andrienko and his colleagues defined the research agenda for spatiotemporal visual analytics, pointing out the unique properties of space and time that necessitate specific approaches to analyzing data with spatial and temporal components.[1] Thus, spatial and temporal dependence (autocorrelation) enable interpolation and extrapolation, which can be used to fill gaps in incomplete data and derive plausible estimates beyond the areas and/or time periods represented in available data; the integration of information of different types and from different sources using references to common locations and/or time units; spatial and temporal inference; and many other operations.

The effects of the spatial and temporal dependences are not absolute, however. Geographic space consists of places with diverse properties, and spatial dependence is weakened by this heterogeneity and by natural or artificial barriers that often exist between places. In 2017, a strategic paper by Alan MacEachren called for geo-visual analytics approaches for defining and characterizing places based on multiple heterogeneous and interconnected data types and sources.[2]

Time can be considered a linearly ordered set of moments or intervals as well as a system of recurring time cycles: daily, weekly, annual, and domain-specific cycles. Therefore, temporal dependence is more complex than just a correlation between close time moments along a timeline because it also includes correlations between corresponding positions in different time cycles. For example, there may be more similarity between the mornings of different days than between the morning and noon time of the same day. As with barriers in space, temporal dependence could also be interrupted by various events. Thus, we must properly take into account both the existence of spatial and temporal dependences and the possibility of distortion or interruption of these dependences.

New Challenges

Traditionally, geographic analysis strongly relied on the use of visual (mostly cartographic) representations, but traditional approaches are now severely challenged by the volume, variety, complexity, dynamics, and other properties of the data requiring analysis. Advances in sensor and positioning technologies in recent years have facilitated an unprecedented growth in the collection of spatially and

temporally referenced data. Examples of big geographic data sources include aerial and terrestrial laser scanning, remote sensing imagery, weather data, data streams from geosensor networks, and tracks of various objects moving by land, sea, and air. The massive volumes of collected data contain complex, yet implicit spatial, temporal, and semantic interrelations that are waiting to be uncovered and made explicit.

One of the most challenging problems in geographic data science is the need to assess the data quality, suitability, and distribution of the data available for analysis.[3,4] Data-quality issues, structure, and feature relationships can often be revealed by appropriate visualizations. In spatiotemporal data, this might involve the detection of misaligned temporal resolutions, temporal regularity or irregularity, the presence of temporal gaps,

Visualization of data is essential for human understanding of the subject under analysis, analytical reasoning about it, and generating new knowledge.

varying spatial resolutions, the presence of spatial gaps, issues concerning the identities of moving objects, properties related to the data-collection method, positioning errors, and others. There is a pressing need for theoretical and methodological development in this direction. A related issue is the visualization of data uncertainty in space[5] and time.[6]

In This Issue

A common approach in geographic data science is the integration of multiple data sets characterized by different spatial and temporal references, at multiple scales and resolutions. Several articles in this special issue propose specific data transformations that allow us to address challenging practical problems. Generally, the articles selected for this special issue represent a mix of theoretical approaches and novel applications of geographic data science.

"Typology of Uncertainty in Static Geolocated Graphs for Visualization" by Tatiana von Landesberger, Sebastian Bremm, and Marcel Wunderlich addresses the important problem of visualizing uncertain geographic information, specifically considering geolocated graphs. The authors consider major types of graph uncertainty (node and edge uncertainty) and their interplay (uncertainty of node-edge relationships and structural implications of uncertainty).

The article "ANALYTiC: An Active Learning System for Trajectory Classification" by Amílcar Soares Júnior, Chiara Renso, and Stan Matwin follows the active learning paradigm for semantic labeling of trajectories. Starting with a manually labeled set of trajectories and their features, the proposed method identifies examples that are hard to label automatically and presents them visually to the user. User input drives the following semantic labeling process.

In the article "Impact of Spatial Scales on the Intercomparison of Climate Scenarios," Wei Luo, Michael Steptoe, Zheng Chang, Robert Link, Leon Clarke, and Ross Maciejewski propose a coordinated multiple views interface for exploring differences between outputs of different climate modeling scenarios. A map interface allows the user to select specific areas of interest. Hierarchical clustering highlights differences between scenarios with respect to attribute groups.

"Urban Space Explorer: A Visual Analytics System for Understanding Urban Planning" by Alireza Karduni, Isaac Cho, Ginette Wessel, William Ribarsky, Eric Sauda, and Wenwen Dou proposes a coordinated multiple views interface for exploring topics in social media posts in relation to the population distribution and the locations of points of interest. Assuming that social media contributors travel on the street network, the proposed system infers the transportation modes, estimates travel times, and assesses traffic flows. Potential users (urban planners) found the multiscale analysis of places to be a useful feature. They expressed interest in the analysis of multiple related geographic layers, such as weather or car accidents.

Finally, in the article "Name Profiler Toolkit," Feng Wang, Brett Hansen, Ryan Simmons, and Ross Maciejewski describe an approach that supports the exploration of spatial distributions of first and last names, their co-occurrence, and relationships with additional data, such as wealth distribution. The power of the system is in fusing data from multiple sources, thus helping users to investigate interesting place-name-age-income relationships.

We expect these articles to help showcase some of the new strands of research in geographic data science. Further, we hope this special issue will stimulate widespread interest among scientists and practitioners in exploring this emerging and exciting field.

References

1. G. Andrienko et al., "Space, Time and Visual Analytics," *Int'l J. Geographical Information Science*, vol. 24, no. 10, 2010, pp. 1577–1600.
2. A.M. MacEachren, "Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier," *Spatial Data Handling in Big Data Era*, Springer, 2017, pp. 139–155.
3. G. Andrienko, N. Andrienko, and G. Fuchs, "Understanding Movement Data Quality," *J. Location Based Services*, vol. 10, no. 1, 2016, pp. 31–46.
4. T. Gschwandtner et al., "A Taxonomy of Dirty Time-Oriented Data," *Proc. Int'l Conf. Availability, Reliability, and Security*, 2012, pp. 58–72.
5. C. Kinkeldey et al., "Evaluating the Effect of Visually Represented Geodata Uncertainty on Decision-Making: Systematic Review, Lessons Learned, and Recommendations," *Cartography and Geographic Information Science*, vol. 44, no. 1, 2017; doi:10.1080/15230406.2015.1089792.
6. T. Gschwandtner et al., "Visual Encodings of Temporal Uncertainty: A Comparative User Study," *IEEE Trans. Visualization and Computer Graphics*, vol. 22, no. 1, 2016, pp. 539–548.

Gennady Andrienko *is a lead scientist at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and a professor at City, University of London. His research interests include visual analytics and geovisualization. He has a PhD in computer science and artificial intelligence from Moscow Lomonosov State University. Contact him at gennady.andrienko@iais.fraunhofer.de.*

Natalia Andrienko *is a lead scientist at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and a professor at City, University of London. Her research interests include visual analytics and geovisualization. She has a PhD in computer science and artificial intelligence from Moscow Lomonosov State University. Contact her at natalia.andrienko@iais.fraunhofer.de.*

Robert Weibel *is a professor of geographic information science (GIScience) and head of the Geographic Information Systems Unit at the University of Zurich. His research interests include computational cartography, computational movement analysis, and spatial analysis for linguistics. Weibel has a PhD in GIScience from the University of Zurich. Contact him at robert.weibel@geo.uzh.ch.*