Chapter 80

Information Visualization Techniques for Big Data:
Analytics Using Heterogeneous Data in Spatiotemporal Domains

William H. Hsu
Kansas State University, USA

ABSTRACT

This chapter presents challenges and recommended practices for visualizing data about phenomena that are observed or simulated across space and time. Some data may be collected for the express purpose of answering questions through quantitative analysis and simulation, especially about future occurrences or continuations of the phenomena – that is, prediction. In this case, analytical computations may serve two purposes: to prepare the data for presentation and to answer questions by producing information, especially an informative model, that can also be visualized. These purposes may have significant overlap. Thus, the focus of the chapter is about analytical techniques for visual display of quantitative data and information that scale up to large data sets. It begins by surveying trends in educational and scientific use of visualization and reviewing taxonomies of data to be visualized. Next, it reviews aspects of spatiotemporal data that pose challenges, such as heterogeneity and scale, along with techniques for dealing specifically with geospatial data and text. An exploration of concrete applications then follows. Finally, tenets of information visualization design, put forward by Tufte and other experts on data representation and presentation, are considered in the context of analytical applications for heterogeneous data in spatiotemporal domains.

1. TRENDS IN DATA VISUALIZATION

1.1. Learning and Analytics Tasks

This section provides a brief history of information visualization for educational and scientific applications, followed by a survey of challenges and tools encountered in visualizing data.

1.1.1. Brief History of Prediction

Information visualization is the study of (interactive) visual representations of abstract data to reinforce human cognition. ("Information visualization", Wikipedia, 2013) Abstract data include both numerical data such as geospatial locations or other physical measurements, and non-numerical data such as text. However, information visualization differs from scientific visualization: Munzner (2008) advises using the term "infovis (for information visualization) when the spatial representation is chosen", and "scivis (scientific visualization) when the spatial representation is given". According to Friendly (2009), scientific visualization is primarily concerned with the "visualization of three-dimensional phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic (time) component".

Input data for visualization includes observational data, collected for the express purpose of answering questions through quantitative analysis, and simulated data, which is generated using a mathematical model. One particular type of simulated data consists of future occurrences or continuations of the phenomena – that is, prediction. Modeling of phenomena for the purpose of forecasting predates computational realization of the methods used, including econometrics (Frisch, 1929) and statistical hypothesis testing (Neyman & Pearson, 1933; Fisher, 1935). Some of the earliest methods for nonlinear time series prediction were extrapolation, interpolation, and smoothing methods derived by: Wiener (1949); Brown (1956), Holt (1957), and Winters (1960); and Box and Jenkins (1970). These contributions comprise fundamental representation and estimation methods that underlie spectral analysis approaches to signal identification, including autoregressive moving average (ARMA) process models.

Although specifically geared towards time series and geospatial data, the visualization approaches covered in this chapter are generally applicable to a variety of data sets and to the behavior and output of many type of machine learning algorithms. Hall et al. (2009) give a much more detailed catalogue of the models and algorithms implemented in the Waikato Environment for Knowledge Analysis (WEKA), to which we refer the interested reader. Predictive visualization, the aspect of information visualization that especially focuses on the continuation of time series beyond historical observations, often poses questions of evaluation using previous unseen data. Watson and Wixom (2009) describe architectures for this type of analytical modeling, among others, in the domain of business intelligence (BI). Business intelligence comprises theories, methodologies, and technologies that serve to transform raw data into information for business decision making. Similar uses of prediction and visualization can be found in most fields where sequences and time series are observed as signals. This includes neuroscience, where such measurements are fundamental, giving rise to the work of scientists such as Elger and Lehnertz (1998).

1.1.2. Challenges of Heterogeneity in Big Data

The term heterogeneous data refers to variables that are fundamentally diverse in character, particularly their source and means of acquisition. One of the key challenges to working with heterogeneous data is that multiple dimensions and a very high volume of data may result from differences in data provenance (origin and preprocessing history). This issue gives rise to the problem of designing visual representations that can consistently support the display of such data. Heer, Kong, and Agrawala (2009) present adjustable parameters such as layering and chart sizing, and discuss the perceptual effects of introducing such degrees of freedom. Monmonier (1990) discusses methods from statistics for coping with the additional technical challenge of working with spatial data over time.

Related Content

Optimizing School Bus Stop Placement in Howard County, Maryland: A GIS-Based Heuristic Approach
Michael Galdiand Paporn Thebpanya (2016). *International Journal of Applied Geospatial Research (pp. 30-44).*
www.irma-international.org/article/optimizing-school-bus-stop-placement-in-howard-county-maryland/143075

Critical Incident Management and Geographically–Based Systems
David W. Webband David R. Hoffpauir (2010). *International Journal of Applied Geospatial Research (pp. 69-75).*
www.irma-international.org/article/critical-incident-management-geographically-based/45131

From Data-Centered to Activity-Centered Geospatial Visualizations
Olga Bucheland Kamran Sedig (2016). *Geospatial Research: Concepts, Methodologies, Tools, and Applications (pp. 246-268).*
www.irma-international.org/chapter/from-data-centered-to-activity-centered-geospatial-visualizations/149496

Classification in GIS Using Support Vector Machines
Alina Lazarand Bradley A. Shellito (2009). *Handbook of Research on Geoinformatics (pp. 106-112).*
www.irma-international.org/chapter/classification-gis-using-support-vector/20393

Built Environment and Driving Outcomes: The Case for an Integrated GIS/GPS Approach
Xiaoguang Wang, Lidia Kostyniukand Michelle Barnes (2014). *International Journal of Applied Geospatial Research (pp. 11-29).*
www.irma-international.org/article/built-environment-and-driving-outcomes/111098