

Visual Inquiry of Spatio-Temporal Multivariate Patterns

Jin Chen

GeoVISTA Center and Department of Geography, Pennsylvania State University

ABSTRACT

While many large, multivariate datasets carry geographic and temporal references, our ability to analyze these datasets lags behind our ability to collect them because of the challenges posed by complexity and scalability issues. This research aims to develop a visual analytics approach that integrates visual, computational and cartographic methods and couples them with human knowledge and judgment to support the exploratory analysis of spatio-temporal, high dimensional, large datasets. By combining both human and machine strengths, the proposed research has a better chance to discover novel, relevant and potentially useful information than can be achieved by applying any of the methods in isolation. The effectiveness of the proposed approach will be tested via case analysis on a variety of systematically selected sample datasets.

KEYWORDS visual analytics, geovisualization, information visualization, exploratory data analysis, visual interaction, spatio-temporal and multivariate data, large and high dimensional data

1 INTRODUCTION

Datasets collected today become increasingly complex, carrying multi-scale geographic locations (i.e. state, county, zip code ...), time-varying information and high-dimensional multivariate data. There are increasingly urgent needs to analyze the datasets for addressing problems in various domains including epidemiology, environment science, homeland security and business analysis such as mobile computing, telecommunications and web traffic analysis. Existing data analysis approaches, from entirely computational to visually-led methods, are limited in ability to analyze the large complex datasets [1]. A combination of computational and visual approaches shows more promise [1, 2].

The research reported here focuses on developing a geo-visual analytic approach to extract complex patterns and relationships from large, multivariate spatio-temporal summary datasets (p , t and *attribute*, where t mean time and p mean a place such as states, counties or zip code area) and to support information synthesis based on domain knowledge and new information. Achieving this goal will require meeting both complexity and scalability challenges. Complexity refers here to multifaceted, intricate problems that require multi-step, often ill-defined, iterative analytical processes. Scalability refers here to methods that can scale up to large volume and high dimensional data, with multiple geographic and time scales (state \rightarrow county \rightarrow zip code; year \rightarrow month \rightarrow week), specifically how to adapt the methods to these datasets while retaining their effectiveness. The proposed approach is being implemented in a proof-of-concept software application—the *Visual Inquiry Toolkit*.

e-mail: jxc93@psu.edu

2 METHODOLOGY

Questions posed by spatio-temporal data analysis usually involve three components: *where*, *when*, and *what* (attribute/thematic objects); and analysis tasks usually are: *when + where \rightarrow what*; *when + what \rightarrow where*; *where + what \rightarrow when*. The tasks follow a scheme $A+B \rightarrow X$, where A and B denote known and X stands for unknown [3]. Andrienko et al [4] proposed a typology that links these tasks systematically to two “search levels” for the analysis tasks: (1) an elementary level task deals with individual objects (e.g.: what is the average temperature for state p in year t ?); (2) a general level task considers a set of objects (e.g. What industrial characteristics have been developed in the U.S. during the past decade? What geographical areas have similar or unusual characteristics, and what are they?). While an elementary question can be answered easily by an interactive data query, a general level task is hard to answer in the same way because: (1) exploratory goals are initially vague, we do not know what data components (what, when, where) to query; (2) these questions are too general to be stated as formal queries.

Analysts exploring a large multivariate spatio-temporal dataset typical start with general (vaguely specified) tasks and, thus need to first gain an overview of the entire dataset and identify interesting patterns [2]; then they need to focus on more specific patterns in detail views. Therefore, this research proposes to adopt and extend the well-known *overview + detail* strategy [5, 6] for the exploratory data analysis. To construct the overviews and the detail views, visualization methods must be employed to present outcomes, to identify complex patterns that are hard to detect by computational methods alone (by taking advantage of human cognition), and to provide an interface for navigating the data exploration processes. On the other hand, visualization methods and human power are limited in processing large volume and high dimensional data;

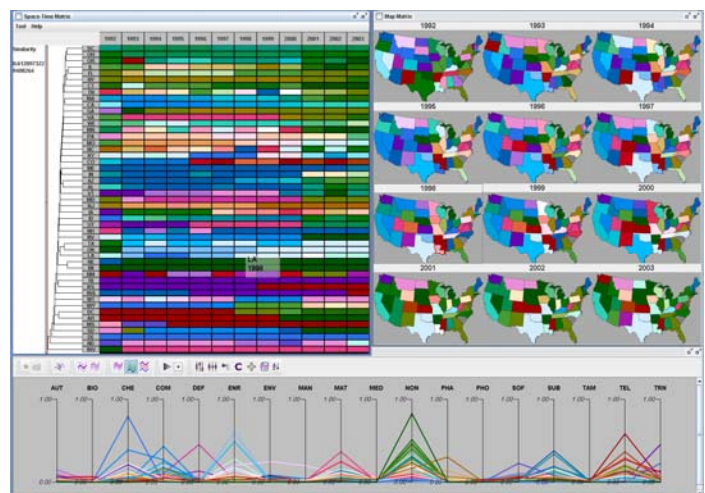


Figure 1 Visual Inquiry Toolkit – a software tool for visual analytics of spatio-temporal, multivariate patterns. In this case, IEEE Infovis 2005 contest dataset is analyzed. Multivariate patterns are displayed in a parallel coordinates plot. The distribution of the patterns is displayed in the space-time matrix (up left) and the small multiple maps (right).

hence data mining and computational methods (e.g. search, sort, aggregation, classification, clustering and so on) must be integrated together with visualization methods. For datasets carrying geographic reference, cartographic methods (e.g. maps) are necessary to identify complex spatial patterns and relationships. In addition, cartographic color schemes are often more effective in encoding patterns than other visual variables such as size, shape, orientation, texture and so on; especially when a number of patterns could potentially be detected and need to be distinguished. In summary, this research proposes an enhanced *overview+detail* exploratory data analysis strategy that couples visual, computational and cartographic methods together for effectively extracting complex patterns and relationships. This strategy emphasizes flexible interaction methods that enable human knowledge and judgment to navigate a multi-step, iterative process of data analysis and information synthesis.

Specifically, in an initial proof-of-concept prototype, a self-organizing map (SOM) [7] is employed to cluster multivariate data. The clusters are visually encoded with a 2D diverging-diverging cartographic color scheme in which similarity of color signifies similarity of cluster [8], thus color is used to uniquely encode a multivariate pattern. The detail of each multivariate pattern is visualized as a string in a Parallel Coordinates Plot (PCP). The distribution of the multivariate patterns across places over time is displayed (as distinct color regions) in a space-time matrix (Figure 1, left). A map matrix (Figure 1, right) is used to easily uncover how geographic patterns change across space over time. The space-time matrix and map matrix complement each other to construct a holistic overview of the entire data with all space, time, attribute components.

The exploratory data analysis process follows three steps of the so-called “information-seeking mantra” [5, 6]: (1) *overview* (2) *Focus and filter* (3) *Details on demand*. In addition, the research adds one more step: information synthesis. Initially, overviews guide an analyst to focus on the most interesting patterns. The analyst can investigate the detail patterns in the PCP and apply domain knowledge to determine what patterns are novel, relevant and important. These interesting patterns can be exported to a *Pattern Basket* for further investigation and synthesis.

To effectively identify complex and implicit patterns and relationships, a multi-step, iterative analysis process is necessary. To support the process, some mechanisms are needed to save and retrieve intermediate, partial results. However, human memory capacity is limited and unable hold all the findings during the analysis process. Analysts need to expand memory capacity by offloading new findings to external memory. The *Pattern Basket* (Figure 2, middle, right) is the proposal addressing the need, where interesting patterns can be preserved. Moreover, the basket provides an interface where information can be synthesized based on domain knowledge and new information available later.

In summary, the proposed combination of space-time matrix, PCP, map matrix and the pattern baskets will provide a platform for a multi-step, incremental and spiral process of pattern searching, analyzing and synthesizing. As part of the research, the methods and the toolkits implemented thus far will be generalized and customized to support analysis of a wider array of datasets (*p*, *t* and *attribute* which either are or can be encoded to numeric values). A key component of the proposed generalization of the methods is to support analyzing geographic detail data (*x*, *y*, *t* and *attribute*).

3 PLAN AND PROGRESS

The research so far has identified the general analytical problems and strategy as previously mentioned. An initial prototype

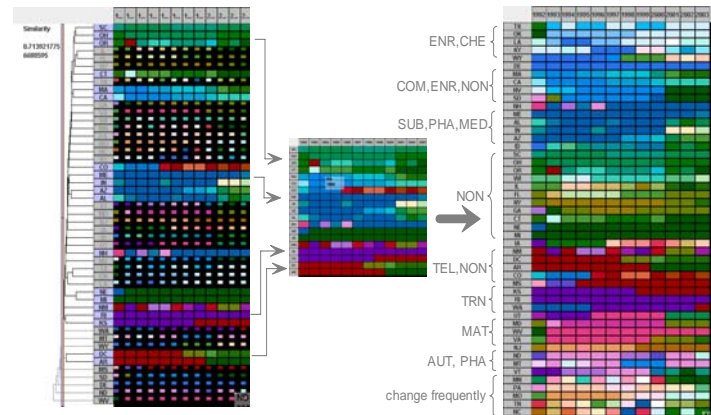


Figure 2 Novel, important patterns (left) are exported to a Pattern Basket (middle) – a variation of the space-time matrix. With incremental pattern searches, a refined overview (right) is constructed where rows are thematically arranged. Each abbreviation (e.g. NON, TEL...) denote a multivariate pattern detected in the IEEE Infovis 2005 contest dataset.

implementation of the strategy – the *visual inquiry toolkit* – has been applied to InfoVis 2005 contest data set. The analysis case provides initial evidence of the potential of the overall visual-computational, overview+details -- focus+filter -- details on demand -- synthesis approach. The proof-of-concept toolkit needs to be compared with other approaches later. The dissertation research will also focus on assessing both the generalizability of the approach and its scalability. The effectiveness of the research will be tested via case analysis on a variety of systematically selected sample datasets.

ACKNOWLEDGMENTS

The research is supported by grant CA95949 from the National Cancer Institute and by the Penn State Regional Visualization and Analytics Center, which is part of a National Consortium led by the National Visualization and Analytics Center at the Pacific Northwest National Laboratory. Special thanks go to Alan MacEachren and Diansheng Guo for their valuable advice.

REFERENCES

- [1] M. Gahegan and B. Brodaric, "Computational and Visual Support for Geographic Knowledge Construction: Filling in the Gaps Between Exploration and Explanation," *Advances in Spatial Data Handling, Proceedings of the 10th International Symposium on Spatial Data Handling*, 2002.
- [2] D. A. Keim, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, pp. 1-8, 2002.
- [3] D. J. Peuquet, "It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems," *Annals of the Association of American Geographers*, vol. 84, pp. 441-461, 1994.
- [4] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, pp. 503-541, 2003.
- [5] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," 1996.
- [6] D. A. Keim, C. Panse, M. Sips, and S. C. North, "Visual data mining in large geospatial point sets," *Computer Graphics and Applications, IEEE*, vol. 24, pp. 36-44, 2004.
- [7] T. Kohonen, *Self-organizing maps*, 2nd ed: Springer, 1997.
- [8] D. Guo, M. Gahegan, A. M. MacEachren, and B. Zhou, "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach," *Cartography and Geographic Information Science*, vol. 32, pp. 113-132, 2005.