

# Discovering Knowledge from High-Dimensional Geographic Data: Integrating Visual and Computational Approaches

Diansheng Guo, Mark Gahegan, Alan M. MacEachren, Donna J. Peuquet

GeoVISTA Center and Department of Geography  
Pennsylvania State University, University Park, PA 16802  
Email: {dguo, mng1, maceachren, peuquet}@psu.edu

## 1. Introduction

It has been widely recognized that spatial data analysis capabilities have not kept up with the need for analyzing the increasingly large volumes of geographic data of various themes that are currently being collected and archived (Openshaw 1991; Miller and Han 2001; Shekhar, Vatsaval et al. 2002; Guo 2003; Guo, Peuquet et al. 2003; Muntz, Barclay et al. 2003). On one hand, such a wealth of data holds great opportunities for geographers, environmental scientists, public health researchers, and others to address urgent and sophisticated geographic problems, e.g., global change, epidemics such as SARS, etc. On the other hand, existing data analysis methods fall short for the extraction of meaningful patterns from datasets of such unprecedentedly *large size* (in terms of the number of observations) and *high dimensionality* (in terms of the number of variables).

Data mining and knowledge discovery refers to the overall *process* of discovering useful knowledge from data, which generally involves data selection, data pre-processing, data transformation, incorporation of appropriate prior knowledge, data mining, and proper interpretation of the results (Fayyad, Piatetsky-Shapiro et al. 1996). While data mining and KDD research has been widely conducted in areas of business, bioinformatics, text mining, etc., it is still at a very early stage in geographic domains.

Geography is an integrative discipline and geographic data under analysis often span across multiple domains. The complexity of spatial data and geographic problems, together with intrinsic spatial relationships, constitute an enormous challenge to conventional data mining methods and call for both theoretical research and development of new techniques to assist in deriving information from large and heterogeneous spatial datasets (Han and Kamber 2001; Miller and Han 2001; Gahegan and Brodaric 2002).

## 2. Size of Feature Space and Hypothesis Space: Problems and Opportunities

A feature space consists of all input data objects, each of which is typically described by many variables. Simply put, a feature space can be imagined as a table in a spreadsheet, with tens of thousands of rows as observations and hundreds of columns containing values for variables (some of which, in a spatial dataset, may represent geographic characteristics and relationships). Unknown and unexpected patterns, trends or relationships can hide deep in such a huge feature space and make it very hard for analytical methods or visual approaches to find (Miller and Han 2000).

A hypothesis space is formed by all possible configurations of the tools used to detect patterns in a feature space. Existing analysis methods often limit (or reduce) the hypothesis space by assuming specific simple forms for patterns within the feature space. For example, the K-Means clustering method assumes clusters in a dataset are of a

circular shape and of similar size. Then the hypothesis space for a K-Means clustering method consists of all possible combinations of  $K$  values (the number of clusters) and positions for the  $K$  clusters. If  $K$  is an input parameter, then the hypothesis space is further reduced to all possible combination of  $K$  cluster positions.

Characteristically, however, the hypothesis space for a large and high-dimensional geographic dataset has an extreme degree of complexity. This is caused by several factors. First, each pattern may involve a different subset of variables from the original data, and the number of such subsets (hereafter *subspaces*), i.e., possible combinations of attributes, is huge. Second, inside a subspace, potential patterns can be of various forms (e.g., clusters can be various shapes). Third, for a specific pattern form (e.g., cluster of a specific shape), its parameter space is still huge, i.e., there are many ways to configure its parameters. Fourth, patterns can vary over geographic space, i.e., patterns can be different from region to region.

Traditional data analysis approaches rely on the human to (1) formulate a hypothesis in the first place and then test it, or (2) provide a simple pattern form (e.g., a mathematical function) and then use data to learn its parameters (e.g., coefficients in regression analysis), or (3) visually sift through the data to search patterns (e.g., information or scientific visualization). However, approaches (1) and (2) are not exploratory means so that we will not learn answers to any questions we do not know how to pose. While visualization is often designed to support exploration and hypothesis generation, it is inefficient, and indeed insufficient, (as an exploratory method) on its own when applied to large datasets.

The richness of attributes (variables, or dimensions) in a data set can provide both opportunities and challenges for data analysis. On one hand, the availability of many attributes within the data enables the identification of complex (and preferably unexpected) patterns (e.g., multivariate relationships across domains). On the other hand, it is inevitable that irrelevant attributes exist in the data and the result can be misleading or useless if the analysis method is unable to discriminate between relevant and irrelevant attributes.

The above problem is often bypassed by requiring the user (who should be an expert on the application problem) to specify a subspace (or several subspaces) for analysis. Given the high dimensionality of currently available data (which may have hundreds of attributes), such a manual approach for choosing attributes is inefficient. Moreover, depending on the user to choose attributes for analysis makes it hard to find *unexpected* patterns, while finding such unexpected patterns is one of the main purposes of data mining and knowledge discovery (Fayyad, Piatetsky-Shapiro et al. 1996).

### **3. The Issues**

#### **3.1. Multivariate spatial patterns**

Most existing spatial analysis methods can only deal with a low-dimensional data space, which usually consists of a geo-referenced attribute and a 2-D or 3-D geographic space. In other words, they generally have very limited ability in identifying multivariate spatial patterns within a dataset of many attributes. On the other hand, geographic considerations related to spatial data analysis are hard to integrate within high-dimensional data mining methods.

### **3.2. Efficiency and scalability**

Large data volumes and/or high dimensionality can cause serious efficiency problems. If a method does not scale well with the data size ( $n$ ) and/or the dimensionality ( $d$ ), its performance (in terms of execution time) can degrade rapidly as  $n$  or  $d$  increases. For example, assume the running time of an analysis method is proportional to  $n^2$  and  $d^2$ . If it takes 1 second to finish for a dataset of size 100 (i.e., having 100 data objects) and dimensionality 10 (i.e., having 10 variables), then, for a dataset of size 10,000 and dimensionality 100, it will take nearly two weeks to finish! Many traditional spatial analysis methods, which often examine each pair of objects, do not scale well with large data size. The efficiency problem becomes extremely crucial when human interaction and guidance are needed during the discovery process.

### **3.3. Understandability and interaction**

The understandability of the output patterns of a KDD system is very important. It becomes even more critical when human interaction, interpretation, and guidance is needed for addressing sophisticated geographic problems. Many data mining methods (or KDD systems) work in a black-box manner and output only static (either descriptive or visual) presentation of patterns, which can leave the user in doubt on many questions. What may change in the pattern if I change this parameter or exclude that variable? How can I know what is the best configuration for this parameter? I got a different set of patterns with another method—how do these two sets of patterns correspond to each other? Will the patterns change over different geographic regions or time periods?

## **4. Research Questions**

### **4.1. Integrating and coordinating different methods in a unified environment**

To achieve both efficiency and effectiveness for exploring complex datasets, a powerful data mining strategy lies in tightly coupling visualization techniques and analytical processes into a unified framework to integrate the best of both human and machine capabilities (Jong and Rip 1997; MacEachren, Wachowicz et al. 1999; Wong 1999; Ankerst, Ester et al. 2000; Peuquet 2002; Guo 2003; Guo, Peuquet et al. 2003). A practical geographic knowledge discovery environment should integrate a suite of computational and visual approaches and support a human-led, interactive, and iterative discovery process. Research questions here may include:

- ◆ How to open computational methods for human interaction and intermediate input/output?
- ◆ How to categorize and standardize the cooperation among different methods so that various methods can be integrated in a unified framework and environment?
- ◆ How to link and brush patterns across components (methods), each of which focuses on a different perspective or space?
- ◆ How to leverage the power of both computational and visual approaches? For example, how to utilize computational methods to suggest most interesting views in visual components, and how to utilize interactive visualization to properly configure/guide computational procedures?
- ◆ How to support a developmental discovery process (Gahegan and Brodaric 2002), with patterns being progressively refined through an iterative process?

- ◆ How to efficiently process large data size and high dimensionality so that real-time human interaction is possible?

#### **4.2. Exposing and understanding the behavior of integrated tools**

Research questions here include:

- ◆ How to help the user choose tools that will work for his/her problem?
- ◆ How to present or visualize, and hence help the user understand, where in the hypothesis space the tools have searched? (Where have we looked?)
- ◆ In a collaborative data exploration task, where several analysts are involved, we may also want to know where our collaborators have looked and what they have found. In other words, how to communicate the operations, findings, and interpretations among collaborators? How to integrate or compare these operations/findings/interpretations to steer further exploration?
- ◆ How to validate the findings and make sure that (1) the findings are not spurious, and (2) the tools do not miss out important patterns? How to measure the validity of high-dimensional patterns and make sure that the patterns are independent of the tools that are used to find them?

#### **4.3. Developing effective unsupervised feature selection methods**

Traditional feature selection methods have been studied in the area of supervised classification (Liu and Motoda 1998). Several unsupervised feature selection methods have recently been developed to select an “optimal” subset of attributes (Dy and Brodley 2000; Dy and Brodley 2000), or produce a pool of “good” subsets of attributes (Kim, Street et al. 2000), for unsupervised clustering. Several subspace clustering methods have been developed to detect clusters residing in different subspaces (Agrawal, Gehrke et al. 1998; Procopiuc, Jones et al. 2002). Research questions here include:

- ◆ How to evaluate the interestingness of a subspace?
- ◆ Given a high-dimensional dataset, how to efficiently and effectively locate interesting subspaces without an exhaustive search, which is extremely expensive in computation?
- ◆ How to visualize the feature selection process, help the expert understand the searching process, and eventually incorporate expert knowledge into the searching process?

#### **4.4. Integrating spatial and non-spatial information**

As introduced in section 3, most of existing spatial methods can only deal with a low-dimensional data space while general-purpose data mining methods have very limited power in dealing with spatial dimensions in a meaningful way. Special considerations related to spatial data are hard to integrate within conventional data mining methods.

Research questions here include:

- ◆ How to encode spatial information into non-spatial data mining methods?
- ◆ How to integrate spatial and aspatial methods for data mining at a more fundamental level than transforming one to the other?
- ◆ How to include geospatial information that goes beyond simple position (e.g., accessibility of a place on a network, “roughness” of the terrain for a region, shape of a region perimeter, etc)?

#### **4.5. Detecting multivariate patterns of various forms**

The challenges here include: (1) not to impose any *a priori* model, or assume a specific form of pattern (i.e., let the data speak for themselves), (2) to be efficient, and (3) to make the result and the discovery process easy to understand. Research questions here include:

- ◆ How to find patterns of various forms?
- ◆ How to efficiently process large datasets?
- ◆ How to progressively formulate patterns when (1) it is too computationally expensive; or (2) human interaction is needed to steer the searching process?
- ◆ How to use visualization to discover multivariate spatial patterns (at a more fundamental level than linking conventional visualization to geographic maps)?
- ◆ How to visualize multivariate spatial patterns? How to validate them?

#### **4.6. Developing interactive geovisualization techniques**

Various visualization techniques are needed to enable a fully open and interactive knowledge discovery environment. As an integral part in a knowledge discovery process, visualization techniques and computational methods are complimentary approaches to each other and share similar goals (MacEachren, Wachowicz et al. 1999). Research questions here include:

- ◆ How to achieve a more intuitive interaction between tools with linking & brushing across different tools, displays, scales & times?
- ◆ How to organize and compartmentalize the feature space according to the mental model(s) of the user, and hence facilitate visual inferences for patterns?
- ◆ How to visually depict the process of searching for patterns in the data space and searching for good configurations in the hypothesis space?
- ◆ How to support collaborative visual data mining that takes advantage of the complementary expertise of a group of analysts?
- ◆ How about using integrated display, sound, motion, animation, VR, etc.?

#### **4.7. Managing data and knowledge**

The representation of spatial-temporal data and geographic knowledge is an important part in a geographic knowledge discovery process (Yuan 1998; Peuquet 2002; Mennis and Peuquet 2003). Research questions here include:

- ◆ How to represent data for efficient data access and effective mining of complex spatial-temporal patterns?
- ◆ How to store and integrate knowledge into the discovery process?
- ◆ How to represent and archive discovered knowledge?
- ◆ How can discovered knowledge be remembered, communicated, and re-used?

### **5. Summary**

There are many research questions that need to be answered, at conceptual and implementation levels. Nevertheless, it is imperative that we develop the tools for unlocking the treasure of information hidden away in data we already have but cannot truly use. We emphasize the necessity of integrating both visual and computational approaches to discover patterns from complex geographic data and address sophisticated geographic problems.

## References

- Agrawal, R., J. Gehrke, D. Gunopulos and P. Raghavan (1998). Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD International Conference on Management of Data*, Seattle, WA, USA, 94-105, ACM Press.
- Ankerst, M., M. Ester and H.-P. Kriegel (2000). Towards an effective cooperation of the user and the computer for classification. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, Boston, Massachusetts, United States, ACM Press New York, NY, USA.
- Dy, J. G. and C. E. Brodley (2000). Feature subset selection and order identification for unsupervised learning. *the Seventeenth International Conference on Machine Learning*, Stanford University, CA, USA, 247-254.
- Dy, J. G. and C. E. Brodley (2000). Visualization and interactive feature selection for unsupervised data. *the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, Massachusetts, United States, 360 - 364, ACM Press.
- Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). From data mining to knowledge discovery-an review. *Advances in Knowledge Discovery*. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusay. Cambridge, MA, AAAI Press/The MIT Press: 1-33.
- Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). The KDD process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM* **39**(11): 27-34.
- Gahegan, M. and B. Brodaric (2002). 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*. D. E. R. a. P. v. Oosterom, Springer: 11 - 25.
- Guo, D. (2003). Coordinating Computational and Visualization Approaches for Interactive Feature Selection and Multivariate Clustering. *Information Visualization* **2**(4): (in press).
- Guo, D., D. Peuquet and M. Gahegan (2003). ICEAGE: Interactive Clustering and Exploration of Large and High-dimensional Geodata. *GeoInformatica* **7**(3): 229-253.
- Han, J. and M. Kamber (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
- Jong, H. d. and A. Rip (1997). The computer revolution in science: steps toward the realization of computer-supported discovery environments. *Artificial Intelligence* **91**(2): 225-256.
- Kim, Y., W. N. Street and F. Menczer (2000). Feature selection in unsupervised learning via evolutionary search. *the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, Massachusetts, United States, 365-369, ACM Press.
- Liu, H. and H. Motoda (1998). Feature selection for knowledge discovery and data mining. Boston, Kluwer Academic Publishers.

- MacEachren, A. M., M. Wachowicz, R. Edsall, D. Haug and R. Masters (1999). Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science* **13**(4).
- Mennis, J. and D. J. Peuquet (2003). The Role of Knowledge Representation in Geographic Knowledge Discovery: A Case Study. *Transactions in GIS* **7**(3): 371 - 392.
- Miller, H. J. and J. Han (2000). Discovering Geographic Knowledge in Data Rich Environments: A Report on a Specialist Meeting. *SIGKDD Explorations* **1**(2): 105-107.
- Miller, H. J. and J. Han (2001). Geographic Data Mining and Knowledge Discovery: an overview. *Geographic Data Mining and Knowledge Discovery*. H. J. Miller and J. Han. London and New York, Taylor & Francis: 3-32.
- Muntz, R. R., T. Barclay, J. Dozier, C. Faloutsos, A. M. MacEachren, J. L. Martin, C. M. Pancake and M. Satyanarayanan (2003). IT Roadmap to a Geospatial Future, report of the Committee on Intersections Between Geospatial Information and Information Technology. Washington, DC, National Academies Press.
- Openshaw, S. (1991). Developing appropriate spatial analysis methods for GIS. *Geographical information systems. Vol. 1: principles*. D. J. Maguire, Longman/Wiley: 389-402.
- Peuquet, D. J. (2002). Representations of space and time, New York : Guilford Press.
- Procopiuc, C. M., M. Jones, P. K. Agarwal and T. M. Murali (2002). A Monte Carlo Algorithm for Fast Projective Clustering. *ACM SIGMOD International Conference on Management of Data*, Madison, Wisconsin, USA, 418-427, ACM Press.
- Shekhar, S., R. R. Vatsaval, M. Gahegan, H. J. Miller, B. Buttenfield and M. Yuan (2002). Geographic Data Mining and Knowledge Discovery. Washington, D.C., University Consortium for Geographic Information Science Research White Paper (available at [http://www.ucgis4.org/priorities/research/2002researchPDF/shortterm/r\\_data\\_mining.pdf](http://www.ucgis4.org/priorities/research/2002researchPDF/shortterm/r_data_mining.pdf)).
- Wong, P. C. (1999). Visual Data mining. *IEEE Computer Graphics & Applications* **19**(5): 20-31.
- Yuan, M. (1998). Representing Spatiotemporal Processes to support Knowledge Discovery in GIS Database. *The 8th International Symposium on Spatial Data Handling*, Burnaby, B.C. Canada, 431-440.