Computers, Environment and Urban Systems 33 (2009) 403–408

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Spatial data mining and geographic knowledge discovery—An introduction

Diansheng Guo [a,1], Jeremy Mennis [b,*]

[a] Department of Geography, University of South Carolina, 709 Bull Street, Room 127, Columbia, SC 29208, United States
[b] Department of Geography and Urban Studies, Temple University, 1115 W. Berks Street, 309 Gladfelter Hall, Philadelphia, PA 19122, United States

ARTICLE INFO

ABSTRACT

Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. There is an urgent need for effective and efficient methods to extract unknown and unexpected information from spatial data sets of unprecedentedly large size, high dimensionality, and complexity. To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases.

This paper highlights recent theoretical and applied research in spatial data mining and knowledge discovery. We first briefly review the literature on several common spatial data-mining tasks, including spatial classification and prediction; spatial association rule mining; spatial cluster analysis; and geovisualization. The articles included in this special issue contribute to spatial data mining research by developing new techniques for point pattern analysis, prediction in space–time data, and analysis of moving object data, as well as by demonstrating applications of genetic algorithms for optimization in the context of image classification and spatial interpolation. The papers concludes with some thoughts on the contribution of spatial data mining and geographic knowledge discovery to geographic information sciences.

1. Introduction

Many fields of geographic research are observational rather than experimental, because the spatial scale is often too large and geographic problems are too complex for experimentation. Researchers acquire new knowledge by searching for patterns, formulating theories, and testing hypotheses with observations. With the continuing efforts by scientific projects, government agencies, and private sectors, voluminous geographic data have been, and continue to be, collected. We now can obtain much more diverse, dynamic, and detailed data than ever possible before with modern data collection techniques, such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information (Goodchild, 2007). Generally speaking, geography and related spatial sciences have moved from a data-poor era to a data-rich era (Miller & Han, 2009). The availability of vast and high-resolution spatial and spatiotemporal data provides opportunities for gaining new knowledge and better understanding of complex geographic phenomena, such as human–environment interaction and social–economic dynamics, and address urgent real-world problems, such as global climate change and pandemic flu spread.

However, traditional spatial analysis methods were developed in an era when data were relatively scarce and computational power was not as powerful as it is today (Miller & Han, 2009). Facing the massive data that are increasingly available and the complex analysis questions that they may potentially answer, traditional analysis methods often have one or more of the following three limitations. First, most existing methods focus on a limited perspective (such as univariate spatial autocorrelation) or a specific type of relation model (e.g., linear regression). If the chosen perspective or assumed model is inappropriate for the phenomenon being analyzed, the analysis can at best indicate that the data do not show interesting relationships, but cannot suggest any better alternatives. Second, many traditional methods cannot process very large data volume. Third, newly emerged data types (such as trajectories of moving objects, geographic information embedded in web pages, and surveillance videos) and new application needs demand new approaches to analyze such data and discover embedded patterns and information.

There is an urgent need for effective and efficient methods to extract unknown and unexpected information from datasets of unprecedentedly large size (e.g., millions of observations), high dimensionality (e.g., hundreds of variables), and complexity (e.g.,

[*] Corresponding author. Tel.: +1 215 204 4748; fax: +1 215 204 7833.
E-mail addresses: guod@sc.edu (D. Guo), jmennis@temple.edu (J. Mennis).
[1] Tel.: +1 803 777 2989; fax: +1 803 777 4972.

0198-9715/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2009.11.001

heterogeneous data sources, space–time dynamics, multivariate connections, explicit and implicit spatial relations and interactions). To address these challenges, *spatial data mining and geographic knowledge discovery* has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases (Andrienko & Andrienko, 1999; Chawla et al., 2000; Gahegan, 2003; Guo, Peuquet, & Gahegan, 2003; Guo et al., 2006; Han, Koperski & Stefanovic, 1997; Keim et al., 2004; Knorr & Ng, 1996; Kulldorff, 1997; Mennis & Liu, 2005; Miller & Han, 2009; Miller and Han, 2001; Openshaw, Charlton, Wymer, & Craft, 1987; Shekhar et al. 2004).

Spatial data mining has deep roots in both traditional spatial analysis fields (such as spatial statistics, analytical cartography, exploratory data analysis) and various data mining fields in statistics and computer science (such as clustering, classification, association rule mining, information visualization, and visual analytics). Its goal is to integrate and further develop methods in various fields for the analysis of large and complex spatial data. Not surprisingly, spatial data mining research efforts are often placed under different umbrellas, such as spatial statistics, geocomputation, geovisualization, and spatial data mining, depending on the type of methods that a research focuses on.

Data mining and knowledge discovery is an iterative process that involves multiple steps, including data selection, cleaning, preprocessing, and transformation; incorporation of prior knowledge; analysis with computational algorithms and/or visual approaches, interpretation and evaluation of the results; formulation or modification of hypotheses and theories; adjustment to data and analysis method; evaluation of result again; and so on (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Data mining and knowledge discovery is exploratory in nature, more inductive than traditional statistical methods. It naturally fits in the initial stage of a deductive discovery process, where researchers develop and modify theories based on the discovered information from observation data (Miller & Han, 2009, pp. 4).

In the literature, knowledge discovery refers to the above multistep process while data mining is narrowly defined as the application of computational, statistical or visual methods. In practice, however, the application of any data mining method should be carried out following the above process to ensure meaningful and useful findings. In this paper, "spatial data mining" and "geographic knowledge discovery" are used interchangeably, both referring to the overall knowledge discovery process.

2. Common spatial data-mining tasks

Spatial data mining is a growing research field that is still at a very early stage. During the last decade, due to the widespread applications of GPS technology, web-based spatial data sharing and mapping, high-resolution remote sensing, and location-based services, more and more research domains have created or gained access to high-quality geographic data to incorporate spatial information and analysis in various studies, such as social analysis (Spielman & Thill, 2008) and business applications (Brimicombe, 2007). Besides the research domain, private industries and the general public also have enormous interest in both contributing geographic data and using the vast data resources for various application needs. Therefore, it is well anticipated that more and more new uses of spatial data and novel spatial data mining approaches will be developed in the coming years. Although we attempt to present an overview of common spatial data mining methods in this section, readers should be aware that spatial data mining is a new and exciting field that its bounds and potentials are yet to be defined.

Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, whether computational, statistical, visual, or some combination of them. Here we only briefly introduce a selected set of tasks and related methods, including classification (supervised classification), association rule mining, clustering (unsupervised classification), and multivariate geovisualization.

2.1. Spatial classification and prediction

Classification is about grouping data items into classes (categories) according to their properties (attribute values). Classification is also called supervised classification, as opposed to the unsupervised classification (clustering). "Supervised" classification needs a training dataset to train (or configure) the classification model, a validation dataset to validate (or optimize) the configuration, and a test dataset to evaluate the performance of the trained model. Classification methods include, for example, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbor methods and case-based reasoning (CBR).

Spatial classification methods extend the general-purpose classification methods to consider not only attributes of the object to be classified but also the attributes of neighboring objects and their spatial relations (Ester, Kriegel, & Sander, 1997; Koperski, Han, & Stefanovic, 1998). A visual approach for spatial classification was introduced in (Andrienko & Andrienko, 1999), where the decision tree derived with the traditional algorithm C4.5 (Quinlan, 1993) is combined with map visualization to reveal spatial patterns of the classification rules. Decision tree induction has also been used to analyze and predict spatial choice behaviors (Thill & Wheelerm, 2000). Artificial neural networks (ANN) have been used for a broad variety of problems in spatial analysis (Fischer, 1998; Fischer, Reismann and Hlavackova-Schindler, 2003; Gopal, Liu and Woodcock, 2001; Yao & Thill, 2007). Remote sensing is one of the major areas that commonly use classification methods to classify image pixels into labeled categories (for example, Cleve, Kelly, Kearns, & Morltz, 2008).

Spatial regression or prediction models form a special group of regression analysis that considers the independent and/or dependent variable of nearby neighbors in predicting the dependent variable at a specific location, such as the spatial autoregressive models (SAR) (Anselin, Syabri, & Kho, 2006; Cressie, 1983; Pace, Barry, Clapp, & Rodriquez, 1998). However, spatial regression methods such as SAR often involve the manipulation of an n by n spatial weight matrix, which is computationally intensive if n is large. Therefore, more recent research efforts have sought to develop approaches to find approximate solutions for SAR so that it can process very large data sets (Griffith, 2004; Kazar, Shekhar, Lilja, Vatsavai, & Pace, 2004; Smirnov & Anselin, 2001).

2.2. Spatial association rule mining

Association rule mining was originally intended to discover regularities between items in large transaction databases (Agrawal, Imielinski, & Swami, 1993). Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items (i.e., items purchased in transactions such as computer, milk, bike, etc.). Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Let X be a set of items and a transaction T is said to contain X if and only if $X \subseteq T$. An association rule is in the form: $X \Rightarrow Y$, where $X \subset I, Y \subset I$ and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of all transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has *support s* in the transaction set D if s% of transactions in D contain $X \cup Y$. Confidence denotes the strength and support indicates the frequencies of the rule. It is often desirable to pay attention to those rules that have reasonably large support (Agrawal et al., 1993).

Similar to the mining of association rules in transactional or relational databases, spatial association rules can be mined in spatial databases by considering spatial properties and predicates (Appice, Ceci, Lanza, Lisi, & Malerba, 2003; Han & Kamber, 2001; Koperski & Han, 1995; Mennis & Liu, 2005). A spatial association rule is expressed in the form $A \Rightarrow B$ [$s\%$, $c\%$], where A and B are sets of spatial or non-spatial predicates, $s\%$ is the support of the rule, and $c\%$ is the confidence of the rule.

Obviously, many possible spatial predicates (e.g., close_to, far_-away, intersect, overlap, etc.) can be used in spatial association rules. It is computationally expensive to consider various spatial predicates in deriving association rules from a large spatial datasets. Another potential problem with spatial association rule mining is that a large number of rules may be generated and many of them are obvious or common knowledge. Domain knowledge is needed to filter out trivial rules and focus only on new and interesting findings.

Spatial co-location pattern mining is spiritually similar to, but technically very different from, association rule mining (Shekhar & Huang, 2001). Given a dataset of spatial features and their locations, a co-location pattern represents subsets of features frequently located together, such as a certain species of bird tend to habitat with a certain type of trees. Of course a location is not a transaction and two features rarely exist at exactly the same location. Therefore, a user-specified neighborhood is needed as a container to check which features co-locate in the same neighborhood. Measures and algorithms for mining spatial co-location patterns have been proposed (Huang, Pei, & Xiong, 2006; Lu & Thill, 2008; Shekhar & Huang, 2001).

2.3. Spatial clustering, regionalization and point pattern analysis

Cluster analysis is widely used for data analysis, which organizes a set of data items into groups (or clusters) so that items in the same group are similar to each other and different from those in other groups (Gordon, 1996; Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999). Many different clustering methods have been developed in various research fields such as statistics, pattern recognition, data mining, machine learning, and spatial analysis.

Clustering methods can be broadly classified into two groups: partitioning clustering and hierarchical clustering. Partitioning clustering methods, such as K-means and self-organizing map (SOM) (Kohonen, 2001), divide a set of data items into a number of non-overlapping clusters. A data item is assigned to the "closest" cluster based on a proximity or dissimilarity measure. Hierarchical clustering, on the other hand, organizes data items into a hierarchy with a sequence of nested partitions or groupings (Jain & Dubes, 1988). Commonly-used hierarchical clustering methods include the Ward's method (Ward, 1963), single-linkage clustering, average-linkage clustering, and complete-linkage clustering (Gordon, 1996; Jain & Dubes, 1988).

To consider spatial information in clustering, three types of clustering analysis have been studied, including spatial clustering (i.e., clustering of spatial points), regionalization (i.e., clustering with geographic contiguity constraints), and point pattern analysis (i.e., hot spot detection with spatial scan statistics). For the first type, spatial clustering, the similarity between data points or clusters is defined with spatial properties (such as locations and distances). Spatial clustering methods can be partitioning or hierarchical, density-based, or grid-based. Readers are referred to (Han, Kamber, & Tung, 2001) for a comprehensive review of various spatial clustering methods.

Regionalization is a special form of clustering that seeks to group spatial objects into spatially contiguous clusters (i.e., regions) while optimizing an objective function. Many geographic applications, such as climate zoning, landscape analysis, remote sensing image segmentation, often require that clusters are geographically contiguous. Existing regionalization methods that are based on a clustering concept can be classified into three groups: (1) multivariate (non-spatial) clustering followed by spatial processing to rearrange clusters into regions (Fovell & Fovell, 1993); (2) clustering with a spatially weighted dissimilarity measure, which considers spatial properties as a factor in forming clusters (Wise, Haining, & Ma, 1997) and (3) contiguity constrained clustering that enforces spatial contiguity during the clustering process (Guo, 2008).

Point pattern analysis, which is also known as "hot spot" analysis (Brimicombe, 2007), focuses on the detection of unusual concentrations of events in space, such as geographic clusters of disease, crime, or traffic accidents. The general research problem is to determine whether there is an excess of observed event points (e.g., disease incidents) for an area (e.g., within a certain distance to a location). Several scan statistics have been developed to find such spatial clusters such as the geographic analysis machine (GAM) by Openshaw et al. (1987), Openshaw, Cross, and Charlton (1990) and the family of space–time scan statistics by Kulldorff (1997), Kulldorff, Heffernan, Hartman, Assunção, and Mostashari (2005). Increasingly statistics for the detection of spatial clusters are available for non-Euclidean spaces, particularly network spaces (Shiode & Shiode, 2009; Xie & Yan, 2008; Yamada & Thill, 2007).

The test statistic used in GAM is the count of points (e.g., disease incidents) within an area (i.e., a circular region around a lattice point). To determine whether the count of points in an area is significant, a Monte Carlo procedure is used to generate a large number (e.g., 500) of random data sets, each representing a realization of the null hypothesis in the same area. A test statistic value is calculated for each random data set and thus a distribution of the test statistic values under the null hypothesis is derived. By comparing the actual test statistic value (i.e., the count of points) and the derived distribution, the significance level for the test statistic in the area is obtained. A potential problem with GAM, as noted in (Rogerson & Yamada, 2009), is that it is difficult to adjust for the multiple-testing problem. Its computational workload is also a disadvantage, but more or less all scan statistics need considerable computational power to search and test local clusters.

The spatial scan statistics developed by Kulldorff (1997) and Kulldorff et al. (2005) calculates a likelihood ratio for each local area. To overcome the multiple-testing problem, the scan statistic uses the maximum likelihood ratio (which is the maximum likelihood ratio among all local areas) as the test statistic. Therefore, the scan statistic method reports the most likely cluster, although a set of secondary clusters is also provided. It first calculates the likelihood ratio for each of a collection of zones and finds the maximum. To derive the significance level, replications of the dataset are generated under the null hypothesis, conditioning on the total number of points. For each replication, the test statistic value is calculated again (i.e., the maximum likelihood ratio is found over all enumerated local areas). Then the actual test statistic value is compared to the test values of all replications to derive the significance level for the most likely cluster (and the secondary clusters).

2.4. Geovisualization

Geovisualization concerns the development of theory and method to facilitate knowledge construction through visual exploration and analysis of geospatial data and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication and use (MacEachren & Kraak, 2001). As an emerging domain, geovisualization has drawn interests from various cognate fields and evolved along a diverse set of research directions, as seen in a recently edited volume on geovisualization by Dykes, MacEachren, and Kraak (2005). The main difference between traditional cartography and geovisualization is that, the former focuses on the design and use of maps for information communication and public consumption while the latter emphasizes the development of highly interactive maps and associated tools for data exploration, hypothesis generation and knowledge construction (MacEachren, 1994; MacEachren & Kraak, 1997).

Geovisualization also has close relations with exploratory data analysis (EDA) and exploratory spatial data analysis (ESDA) (Anselin, 1999; Bailey & Gatrell, 1995; Tukey, 1977), which links statistical graphics and maps and relies on the human expert to interact with data, visually identify patterns, and formulate hypotheses/models. However, to cope with today's large and diverse spatial data sets and facilitate the discovery and understanding of complex information, geovisualization needs to address several major challenges, including (1) processing very large datasets efficiently and effectively; (2) handling multiple perspectives and many variables simultaneously to discover complex patterns and (3) the design of effective user interface and interactive strategy to facilitate the discovery process.

To process large data sets and visualize general patterns, visual approaches are often combined with computational methods (such as clustering, classification, and association rule mining) to summarize data, accentuate structures and help users explore and understand patterns (Andrienko & Andrienko, 1999; Guo, Gahegan, MacEachren, & Zhou, 2005; Ward, 2004). To visualize multiple perspectives and many variables, we often need to couple visualization with dimension reduction techniques, such as multidimensional scaling, principle components analysis (PCA), self-organizing maps (SOM) (Agarwal & Skupin, 2008; Kohonen, 2001), or other projection pursuit methods (Cook, Buja, Cabrera, & Hurley, 1995). Multivariate mapping has long been an interesting research problem, for which numerous approaches have been developed, such as specially designed symbols (Chernoff & Rizvi, 1975; Zhang & Pazner, 2004), multiple linked views (Dykes, 1998; MacEachren, Wachowicz, Edsall, Haug, & Masters, 1999; Monmonier, 1989; Yan & Thill, 2009), and clustering-based approaches (Guo et al., 2003,2005). Research efforts for the third challenge have emerged as an active subfield called visual analytics (Thomas & Cook, 2005).

3. Overview of the articles

Here we provide an overview of the articles in the special issue. These articles make contributions to the spatial data mining literature in a variety of ways. Some of the articles extend established techniques, such as artificial neural networks (ANN) and spatial clustering, to account for issues of spatial dependency and spatial scale. Others develop new techniques for types of spatial data that are only recently becoming widely available, such as path and trajectory data describing moving objects. The contribution of other articles concern new applications of data mining techniques.

As noted earlier, classification and prediction is a fundamental data-mining task and ANNs are among the commonly used classification methods. However, conventional ANN does not consider the spatial dependence and associations between neighboring objects. Cheng and Wang (2009) seek to address this issue in developing an ANN for space–time prediction. Their approach incorporates spatial associations among observations into dynamic recursive neural networks (DRNN), an ANN approach that incorporates feedback from previous iterations of the model inputs and outputs. Such feedbacks make DRNN a good candidate for modeling time-series data. In the present article, the authors propose that a target prediction can be improved by not only incorporating the value of the target at the previous time interval but the values of nearby observations. Three case studies serve to demonstrate this approach using a variety of types of data with varying spatial and temporal records – the prediction of forest fires, economic gross domestic product, and temperature. Results indicate that including spatial association information can improve the computational performance and accuracy of DRNN for space–time prediction.

One of the major challenges to spatial data mining arises from handling new kinds of data. Recent advances in embedding GPS to create location-aware devices have generated a massive volume of data about moving objects. Detecting patterns in these data are challenging due to both the massive volume and temporal nature of the data. Dodge, Weibel, and Forootan (2009) address this challenge in their article, which focuses on the classification of moving trajectories. The authors present a way to characterize moving object trajectories from both a global perspective, i.e. those properties that characterize the entire trajectory of the object, and a local perspective, i.e. those properties that characterize portions of the object's trajectory. Properties include characteristics such as path length and straightness as well as velocity and acceleration. With these extracted characteristics, a SVM is applied to classify trajectories into categories. Two types of data, transportation data of moving vehicles as well as eye-tracking data, are used to demonstrate the proposed approach.

The third paper by Pei, Zhu, Zhou, Li, and Qin (2009) focuses on the development of a new method for point pattern analysis. The authors note that established spatial clustering methods are often sensitive to the parameterization of the clustering algorithm, particularly to the scale at which one theorizes clustering occurs, as such an assumption often must be made a priori to the application of the clustering technique. Consequently, the results of clustering may be highly subjective. To address this issue the authors present a new method of clustering they call the collective nearest neighbor (CLNN) method. The basis for CLNN is the distinction between points whose distribution may be explained by a causal mechanism versus those whose distribution may be explained by random 'noise,' where the distinguishing characteristics between the two processes is intensity of clustering. CLNN extends previous research by developing a procedure for iterating over various scales of measurement to assess intensity. The authors demonstrate CLNN using both synthetic data as well as a case study focusing on identifying clusters of earthquakes in China from seismic data.

Lu, Chen, and Hancock (2009) also focus on data mining of moving object data, but they are interested in clustering and path anomaly detection, i.e., the identification of outliers in a set of moving object paths. These authors propose several new metrics that may be used to identify the degree of similarity among paths, including metrics that capture the global character of a path, such as the perimeter of the region that the path occupies. Other metrics characterize smaller sections of a path, which may be segmented based on the edges shared with another path. The path anomaly detection approach was tested under two scenarios on a set of synthetic paths mapped onto the street network of Oldenburg, Germany. In the first scenario, a set of shortest paths between randomly defined beginning and ending nodes in the city were generated, as well as a smaller set of paths that were forced to visit a vertex not on the shortest path. The algorithm aimed to distinguish those paths that were not shortest-paths from the larger set. In the second scenario, a set of 'normal' paths is defined such that all paths in the set begin in a certain region of the city and end in a certain region of the city, where each normal path is the shortest path. The regions represent locales where people would typically travel to and from, for instance, a shopping mall or residential development. Two outlier sets of paths are defined. In the first set the paths between clusters are not the shortest path. In the second set, the paths have normal length but do not end in a cluster. Results suggest the efficacy of these metrics for identifying path anomalies, with segmentation-based approaches outperforming the globally-derived metrics.

Genetic algorithm is an approach inspired by biological systems to identify optimal, or near-optimal, solutions to optimization problems. The classification of spatial image data is one example of this type of problem to which evolutionary algorithms can be applied, and serves as the focus of the research presented by Momm, Easson, and Kuszmaul (2009). The authors address the classification of multispectral remotely sensed imagery using a non-linear combination of both spectral information and texture metrics. The challenge here is the choice of image texture metric used, as the most informative texture metric differs among problem domains. The choice and combination of texture metric with spectral information, in order to produce the most accurate image classification, can be viewed as an optimization problem solved by genetic programming. The genetic programming approach was compared with a conventional image classifier, K-Means, using a multi-spectral Quickbird image for Oxford, Mississippi, USA, with an emphasis on differentiating classes with similar spectral, but different textural, characteristics.

Shad, Mesgari, and Abkar (2009) also employ genetic algorithms as an approach to optimization. However, in this case the objective is the interpolation of air pollution data, specifically particulate matter, using the fuzzy membership indicator Kriging method. Here, fuzzy logic is used to model uncertainty in the prediction process, where degree of predicted membership can be represented using fuzzy logic. The optimization challenge lies in defining optimal parameterization of the fuzzy logic function for defining degree of membership. For this purpose, the authors use genetic algorithms, where various fuzzy membership functions are allowed to compete to maximize the accuracy of pollution estimation. This approach was applied to a data set of air particulate concentrations gathered from air pollution monitoring stations in Tehran, Iran.

4. Conclusion

Due to the widespread application of geographic information systems (GIS) and GPS technology and the increasingly mature infrastructure for data collection, sharing, and integration, more and more research domains have gained access to high-quality geographic data and created new ways to incorporate spatial information and analysis in various studies. Private industries and the general public also have more and more interest in both contributing and using geographic data. These data have become more diverse, complex, dynamic, and much larger than ever before and therefore are more difficult to analyze and understand. Spatial data mining and knowledge discovery has emerged as an active research field that focuses on the development of theory, methodology, and practice for the extraction of useful information and knowledge from massive and complex spatial databases. The articles in this special issue highlight a selected set of approaches and application in spatial data mining. As noted earlier, spatial data mining is still at a very early stage and its bounds and potentials are yet to be defined. There are both opportunities and challenges facing spatial data mining research.

Spatial data mining is not a push button task. We often claim to "let the data speak for themselves". However, the data cannot tell stories unless we formulate appropriate questions to ask and use appropriate methods to solicit the answers from the data. Data mining is data-driven but also, more importantly, human-centered, with the user controlling the selection and integration of data, cleaning and transformation of the data, choice of analysis methods, and the interpretation of results. It is an iterative and inductive learning process that is embedded in an overall deductive framework.

The abundance of spatial data provides exciting opportunities for new research directions but also demands caution in using these data. The data are often from different sources and collected for different purposes under various conditions, such as measurement uncertainty, biased sampling, varying area unit, and confidentiality constraint. It is important to understand the quality and characteristics of the chosen data.

Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results. What variables should be selected? What measurement framework, such as Euclidean space or non-metric network space, should be used? What spatial relations or contextual information should be considered? Can the chosen data adequately represent the complexity and nature of the problem?

New types of data and new application areas (such as the analysis of moving objects and trajectories, spatially embedded social networks, spatial information in web-based documents, geocoded multimedia, etc.) have significantly expanded the frontier of spatial data mining research. New data types and applications often require the development of new data mining methods and the discovery of new types of patterns.

Handling the very large volume and understanding complex structure in spatial data are another two major challenges for spatial data mining, which demand both efficient computational algorithms to process large data sets and effective visualization approaches to present and explore complex patterns.

Acknowledgment

We would like to express our thanks to all the authors who submitted papers to the special issue and especially Jean-Claude Thill for his assistance in preparing this special issue.

References

Agarwal, P., & Skupin, A. (2008). *Self-organising maps: Applications in geographic information science*. Chichester: Wiley.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD international conference on management of data* (pp. 207–216).

Andrienko, G., & Andrienko, N. (1999). Data mining with C4.5 and interactive cartographic visualization. In N. W. G. T. Paton (Ed.), *User interfaces to data intensive systems* (pp. 162–165). Los Alamitos, CA: IEEE Computer Society.

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical information systems–principles and technical issues* (pp. 253–266). New York, NY: John Wiley & Sons, Inc..

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis, 38*(1), 5–22.

Appice, A., Ceci, M., Lanza, A., Lisi, F. A., & Malerba, D. (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis, 7*(6), 541–566.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. New York, NY: John Wiley and Sons, Inc..

Brimicombe, A. J. (2007). A dual approach to cluster discovery in point event data sets. *Computers Environment and Urban Systems, 31*(1), 4–18.

Cheng, T., & Wang, J. (2009). Accommodating spatial associations in DRNN for space–time analysis. *Computers, Environment and Urban Systems, 33*(6), 409–418.

Chernoff, H., & Rizvi, M. H. (1975). Effect on classification error of random permutations of features in representing multivariate data by faces. *Journal of American Statistical Association, 70*, 548–554.

Cleve, C., Kelly, M., Kearns, F. R., & Morltz, M. (2008). Classification of the wildland–urban interface: A comparison of pixel- and object-based classifications using high-resolution aerial photography. *Computers Environment and Urban Systems, 32*(4), 317–326.

Cook, D., Buja, A., Cabrera, J., & Hurley, C. (1995). Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics, 4*(3), 155–172.

Cressie, N. A. C. (1983). *Statistics for spatial data*. New York: Wiley–Interscience.

Dodge, S., Weibel, R., & Forootan, E. (2009). Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems, 33*(6), 419–434.

Dykes, J. (1998). Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using tcl/tk and cdv'. *The Statistician, 47*(3), 485–497.

Dykes, J., MacEachren, A. M., & Kraak, M.-J. (2005). *Exploring geovisualization*. Amsterdam: Elsevier.

Ester, M., Kriegel, H. P., & Sander, J. (1997). Spatial data mining: A database approach. In *Advances in spatial databases* (pp. 47–66). Berlin: Springer-Verlag Berlin.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery—an review. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusay (Eds.), *Advances in knowledge discovery* (pp. 1–33). Cambridge, MA: AAAI Press/The MIT Press.

Fovell, R. G., & Fovell, M.-Y. C. (1993). Climate zones of the conterminous united states defined using cluster analysis. *Journal of Climate, 6*(11), 2103–2135.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *Journal of Geography, 69*(4), 211–221.

Gordon, A. D. (1996). Hierarchical classification. In P. Arabie, L. J. Hubert, & G. D. Soete (Eds.), *Clustering and classification* (pp. 65–122). River Edge, NJ, USA: World Scientific Publisher.

Griffith, D. (2004). Faster maximum likelihood estimation of very large spatial autoregressive models: An extension of the Smirnov–Anselin result. *Journal of Statistical Computation and Simulation, 74*(12), 855–866.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science, 22*(7), 801–823.

Guo, D., Gahegan, M., MacEachren, A. M., & Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science, 32*(2), 113–132.

Guo, D., Peuquet, D., & Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *Geoinformatica, 7*(3), 229–253.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

Han, J., Kamber, M., & Tung, A. K. H. (2001). Spatial clustering methods in data mining: A survey. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 33–50). London and New York: Taylor and Francis.

Huang, Y., Pei, J., & Xiong, H. (2006). Mining co-location patterns with rare events from spatial data sets. *Geoinformatica, 10*(3), 239–260.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*(3), 264–323.

Kazar, B. M., Shekhar, S., Lilja, D. J., Vatsavai, R. R., & Pace, R. K. (2004). In M. J. Egenhofer, C. Freksa, & H. J. Miller (Eds.), *Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis* (pp. 140–161). Berlin: Springer-Verlag.

Kohonen, T. (2001). *Self-organizing maps*. Berlin; New York: Springer.

Koperski, K., Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *The 4th int'l symp. on large spatial databases* (SSD95) (pp. 47–66), Maine, USA.

Koperski, K., Han, J., and Stefanovic, N. (1998). An efficient two-step method for classification of spatial data. In 1998 *international symposium on spatial data handling* SDH'98 (pp. 45–54), Vancouver, BC, Canada.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics—Theory and Methods, 26*, 1481–1496.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R. M., & Mostashari, F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine, 2*, 216–224.

Lu, Q., Chen, F., & Hancock, K. (2009). On path anomaly detection in a large transportation network. *Computers, Environment and Urban Systems, 33*(6), 448–462.

Lu, Y., & Thill, J.-C. (2008). Cross-scale analysis of cluster correspondence using different operational neighborhoods. *Journal of Geographical Systems, 10*(3), 241–261.

MacEachren, A. (1994). Visualization in modern cartography: Setting the agenda. In D. R. F. Taylor & A. M. MacEachren (Eds.), *Visualization in modern cartography* (pp. 1–12). Oxford, UK: Pergamon.

MacEachren, A. M., & Kraak, M. J. (1997). Exploratory cartographic visualization: Advancing the agenda. *Computers and Geosciences, 23*, 335–343.

MacEachren, A., & Kraak, M.-J. (2001). Research challenges in geovisualization. *Cartography and Geographic Information Science*, 283–312.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D., & Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science, 13*(4), 311–334.

Mennis, J., & Liu, J. W. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS, 9*(1), 5–17.

Miller, H., & Han, J. (2009). Geographic data mining and knowledge discovery: An overview. In H. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 1–26). CRC Press, Taylor and Francis Group.

Momm, H., Easson, G., & Kuszmaul, J. (2009). Evaluation of the use of spectral and textural information by an evolutionary algorithm for multi-spectral imagery classification. *Computers, Environment and Urban Systems, 33*(6), 463–471.

Monmonier, M. (1989). Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis, 21*(1), 81–84.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Science, 1*(4), 335–358.

Openshaw, S., Cross, A., & Charlton, M. (1990). Building a prototype geographical correlates exploration machine. *International Journal of Geographical Information Systems, 4*(3), 297–311.

Pace, R. K., Barry, R., Clapp, J. M., & Rodriquez, M. (1998). Spatiotemporal autoregressive models of neighborhood effects. *Journal of Real Estate Finance and Economics, 17*(1), 15–33.

Pei, T., Zhu, A. X., Zhou, C., Li, B., & Qin, C. (2009). Detecting feature from spatial point processes using collective nearest-neighbor. *Computers, Environment and Urban Systems, 33*(6), 435–447.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Rogerson, P., & Yamada, I. (2009). *Statistical detection and surveillance of geographic clusters*. Taylor and Francis Group.

Shad, R., Mesgari, M. S., & Abkar, A. Shad (2009). Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *Computers, Environment and Urban Systems, 33*(6), 472–481.

Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In C. Jensen, M. Schneider, B. Seeger, & V. Tsotras (Eds.), *Advances in spatial and temporal databases, proceedings, lecture notes in computer science* (pp. 236–256). Berlin: Springer-Verlag.

Shiode, S., & Shiode, N. (2009). Detection of multi-scale clusters in network space. *International Journal of Geographical Information Science, 23*, 75–92.

Smirnov, O., & Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristics polynomial approach. *Computational Statistics and Data Analysis, 35*, 301–319.

Spielman, S. E., & Thill, J. C. (2008). Social area analysis, data mining and GIS. *Computers Environment and Urban Systems, 32*(2), 110–122.

Thill, J.-C., & Wheelerm, A. (2000). Tree induction of spatial choice behavior. *Transportation Research Record, 1719*, 250–258.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The research and development agenda for visual analytics*. Los Alametos, CA: IEEE Computer Society.

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley.

Ward, J. H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistic Association, 58*, 236–244.

Ward, M. O. (2004). Finding needles in large-scale multivariate data haystacks. *Computer Graphics and Applications, 24*(5), 16–19.

Wise, S. M., Haining, R. P., & Ma, J. (1997). Regionalization tools for the exploratory spatial analysis of health data. In M. Fischer & A. Getis (Eds.), *Recent developments in spatial analysis: Spatial statistics, behavioural modelling and neuro-computing*. Berlin: Springer-Verlag.

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems, 32*, 396–406.

Yamada, I., & Thill, J.-C. (2007). Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis, 39*(3), 268–292.

Yan, J., & Thill, J.-C. (2009). Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B, 36*, 466–486.

Yao, X., & Thill, J.-C. (2007). Neurofuzzy modeling of context–contingent proximity relations. *Geographical Analysis, 39*(2), 169–194.

Zhang, X., & Pazner, M. (2004). The icon imagemap technique for multivariate geospatial data visualization: Approach and software system. *Cartography and Geographic Information Science, 31*(1), 29–41.