In Press: Chapter 14 in *Research Challenges in Geographic Information Science*, R. McMaster and L Usery eds. John Wiley & Sons

GEOSPATIAL DATA MINING AND KNOWLEDGE DISCOVERY

Yuan, M., Buttenfield, B., Gahegan, M. N., and Miller, H.

Introduction

The advent of remote sensing and survey technologies over the last decade has dramatically enhanced our capabilities to collect terabytes of geographic data on a daily basis. However, the wealth of geographic data cannot be fully valued when information implicit in data is difficult to discern. This confronts GI scientists with an urgent need for new methods and tools that can intelligently and automatically transform geographic data into information and, furthermore, synthesize geographic knowledge. It calls for new approaches in geographic representation, query processing, spatial analysis, and data visualization (Yuan 1998, Miller and Han 2000, Gahegan, 2000a). Information scientists face the same challenge as a result of the digital revolution that expedites the production of mountains of data from credit card transactions, medical examinations, telephone calls, stock values, and other numerous human activities. Collaborative efforts in artificial intelligence, statistics, and databases communities have been developing technologies of knowledge discovery in databases (KDD) to extract useful information from massive amounts of data in support of decision-making (Gardner 1996, Bhandari *et al.* 1997, Hedberg 1996).

KDD technology has emerged as an empowering tool in the development of the next generation database and information systems through its abilities to extract new, insightful information embedded within large heterogeneous databases and to formulate knowledge. A KDD process includes "*data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted knowledge*" (Fayyad 1997, P5). Specifically, data mining aims to discover something new from the facts recorded in a database. It prescribes the steps toward efficient development of knowledge discovery applications. Hitherto, data mining tools mostly adopt techniques from statistics (Glymour *et al.* 1996), neural networks (Lu *et al.* 1996), and visualization (Lee and Ong 1996) to classify data and extract patterns. But ultimately, KDD aims to enable an information system that transforms the information to knowledge through hypothesis testing and theory formation. It sets new challenges for database technology: new concepts and methods are needed for basic operations, query languages, and query processing strategies (Lmielinski and Mannila 1996).

This chapter provides a research frame for GI scientists in study the integration of geospatial data mining and knowledge discovery. Following an examination of the current state of DM and KDD technology, we identify special needs for geospatial DM and KDD, and outline research challenges and the significance to national research needs. We outline research frontiers in geographic knowledge discovery briefly and propose a research agenda to highlight short-term, mid-term, and long-term objectives.

Background: An overview of the state-of-art in data mining and knowledge discovery

There is currently a good deal of interest in geospatial data as a rich source of structure and pattern, making it ideal for data mining research (e.g. Koperski and Han, 1995; Ester et al., 1996, 1998; Knorr and Ng, 1996; Koperski et al, 1999; Roddick and Spiliopoulou, 1999). Many of the very large consumer, medical and financial transaction databases now being constructed contain spatial and temporal attributes and hence offer the possibility of discovering or confirming geographical knowledge (Miller and Han, 2001). For decision makers this knowledge represents improved decision power. While data mining (DM) and knowledge discovery in databases (KDD) quickly become popular among GI scientists, misconception appears not uncommon. Through discussions of the essence of DM/KDD and academic heritage, this section aims to draw a clear picture of the involved science and technology.

What DM/KDD is, and is not

A generally accepted definition of DM and KDD is given by Fayyad et al. (1996) as: "*…the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data*." From this definition we can see the following:

1. DM is not straightforward analysis or machine learning. It is non-trivial, usually in the sense that the dataset under consideration is massive. If this were not so, then an exhaustive statistical analysis should be possible, and is usually preferable since it is more rigorous. Data mining methods contain a degree of non-determinism to enable them to scale to massive datasets. Smythe (2000) provides some clarification in later work where he challenges the somewhat prevalent view that applying an established inductive machine learning technique to data (such as a decision tree) qualifies as mining.

2. Some aspect is unknown at the start of the DM process and must be found. The term data mining does not apply in cases where the outcome is already known, i.e. deterministic or deductive problems. Perhaps ideally, data mining should be an abductive task (originally named by Peirce (1878) as hypothesis), simultaneously uncovering some structure within the data and a hypothesis to explain it. However, this would require sophisticated conceptual structures by which hypotheses might be represented within a machine. Currently, the focus of knowledge discovery seems to be on inductive learning methods, where the aim is to construct a model for the intension of some category from identified training examples. Because the structure is largely known (by way of these training examples) this is not mining per se, but rather a form of knowledge generalization. One exception to this is where the training examples themselves are a hypothesis only, generated from the data rather than given

a-priori, in an effort to establish classes with which to represent the data.Tools such as AutoClass (Cheeseman and Stutz, 1996) function in this manner.

3. The uncovered structure needs to be valid; i.e. shown to be a significant or reliable inference with some level of confidence. Reliability metrics are required to support the hypotheses presented and to differentiate the significant from the marginal.

4. The findings should be novel, that is, unknown at the outset. Obviously, the machine has no concept of what is known by experts, so has no means by which to map novelty to the application domain of discourse. However, it is possible to post-process results so that similar inferences are grouped together in generalized forms. Bradsil and Konolige (1990) refer to this as meta-learning. Each discovery is thus assured of being distinct from its peers.

5. The uncovered structure needs to be useful, i.e. be explainable and applicable in a manner that makes sense within the context of the current application domain. Large datasets may contain a great deal of structure that is not in itself useful. Focusing effort on those parts that are interesting is problematic because they are by definition unknown at the outset.

Academic Heritage

Successful applications of data mining are not common, despite the vast literature now accumulating on the subject. The reason is that, although it is relatively straightforward to find pattern or structure in data, establishing its relevance and explaining its cause are both very difficult problems. Furthermore, much of what can be 'discovered' may well already be common knowledge to the expert. Addressing these problematic issues requires the synthesis of underlying theory from the database, statistics, machine learning and visualization communities. The issues relevant to data mining from each of these disciplines, including database, statistics, and artificial intelligence, are described below (after Smythe, 2000).

1. Database

The database community draws much of its motivation from the vast digital datasets now available online and the computational problems involved in analyzing them. Almost without exception, current databases and database management systems are designed without thought to knowledge discovery, so the access methods and query languages they provide are often inefficient or unsuitable for mining tasks (Rainsford and Roddick, 1999).

In geographical analysis, Openshaw's Geographical Analysis Machine (Openshaw et al., 1990) is an example of a more-or-less exhaustive data mining tool. However, such brute force approaches do not scale well. As pointed out above, data mining usually begins from the assumption that the dataset is massive, and accordingly the analysis tools must be designed so that computational performance is given the utmost priority. Approaches to improving performance can take the following forms.

- Optimization of existing methods

Many analysis techniques scale somewhere between O(n^3) and O($n\log(n)$) in terms of computational complexity (Martin, 1991), with the majority falling somewhere around O(n^2). For smaller datasets this causes no problems, but where the number of features (attributes) is large, or the number of records is large, or both, such scaling renders existing techniques unusable. Many breakthroughs have been reported in the last few years to improve computational complexity so that it approaches O(n). Techniques are typically based around optimistic optimization of hierarchical methods, such as decision trees, and include RIPPER (Cohen, 1995) and BOAT (Gehrke et al., 1999).

- Approximation of existing methods

Algorithms attempt to encapsulate all the important structure contained in the original data, so that information loss is minimal and mining algorithms can function more efficiently. The premise here is that the functionality of some existing analysis methods can be approximated either by (a) sampling the data or (b) re-expressing the data in a simpler form. Sampling strategies must try to avoid bias, which is difficult if the target and its explanation are unknown. Data reduction approaches must attempt to 'squash' the data into some lower dimensional form, similar in concept to a principal component transformation or a self-organizing (Kohonen) map.

- New methods for data mining

Smythe (2000) points out that a variety of new approaches to data mining have been created, that can function well using standard query interfaces and languages. He cites association rules (Agrawal et al., 1993) as the most established example, but goes on to caution that they are rather impoverished in the analytic sense as they need further processing before they can represent the statistical significance of findings. However, one of the few documented successes of data mining so far has been in analyzing consumer behavior by applying association rules to databases of purchases (e.g. Berry and Lino, 1997). Such rules can be used to uncover likelihood of one type of purchase, given a set of others. They form the basis of some on-line, consumer analysis applications, too.

2. Statistics

The algorithmic basis of many data mining methods can be traced back to multivariate statistical principles such as maximum likelihood, linear discriminant and k-means functions. Good accounts of multivariate analysis in statistics are given by Dunteman (1984) and Mardia et al., 1979). These parametric approaches to analysis are complemented by clustering methods, as exemplified by the works of Anderberg (1973), Devijver and Kittler (1982) and Kaufman and Rousseeuw (1990). More recently, further progress has been made with the development of specifically spatial clustering techniques (see Murray and Estivill-Castro, 1998, for a useful summary). From a statistical perspective, the challenges posed by data mining are fundamental, forcing the development of new types of inferential analysis techniques focused on discovering and evaluating local patterns within the data (e.g. Anselin, 1995; 1996) rather than validating or refuting established global models.

- Validating the findings

Many of the techniques used to uncover local structure are not statistically rigorous and the challenge is to make them so (Elder and Pregibon, 1996). Data mining techniques, such as association rule construction, are less rigorous that existing statistical methods and do not confirm to significance testing using established statistical theory (Glymour et al., 1997; Smythe, 2000). In a predictive sense this makes reliability assessment problematic. Furthermore, data mining proceeds by constructing many (millions or even billions) of local hypotheses; even using a very high significance value we might reasonably expect a very large, even massive, number of ‘false positives’. This causes two distinct problems. Firstly, how might more reliable measures of significance be constructed and secondly, how can false positives be differentiated from truly significant findings?

3. Artificial Intelligence (AI)

From an AI perspective, difficult problems of a representational nature present themselves. A variety of machine learning methods can be used to perform some of the generalization and inductive learning tasks associated with knowledge construction, including: case-based reasoning, neural networks, decision trees, rule induction, Bayesian belief networks, genetic algorithms, fuzzy and rough sets theory. See Mitchell (1997) for details of the workings of these methods.

- Explaining the findings

A noted above, to be truly abductive, structure must be simultaneously discovered and explained by a hypothesis of some sort. Ideally, this hypothesis would be constructed in the domain of the expert, i.e. a high-level or abstract reason that makes sense within a specific problem context. But more realistically, hypotheses are given in the lower-level language of the data and clustering tools (e.g. an induction rule hierarchy), making them difficult to interpret by the human expert. The need here is for more complex models of geography (or other application domains) to be represented within the computer, which would provide the structure required for a higher (more abstract) form of abduction to take place (e.g. Sowa, 1999: Ch. 7). That is not to say the existing methods are not useful, since any clues to structure in data may well help trigger abductive reasoning by the expert, mapping the low level hypothesis into the application domain.

- Representing the findings

If new objects or categories are being uncovered, then they will also need to be represented in some manner. This topic also involves a significant database component. If findings are to be worked back into the database schema, then this schema must be capable of dynamic update (Drew and Ying, 1998). Furthermore, the semantics of the schema will need to be rich enough to encode the discovered relationships, or again capable of evolving the required relationship-types (e.g. Luger and Stubblefield, 1998). This latter requirement is more difficult because it involves extending the semantic richness of a data model, rather than simply adding in new tables and populating them. Within the geographic sphere, this requirement causes particular difficulty, because implementations of conceptual models vary widely in terms of functionality and level of abstraction. Furthermore, there are only and handful of academic models that might be able to represent discovered spatio-temporal relationships, and none in commercial production at this time.

- Reporting the findings

Related to the above, discovered or uncovered knowledge must be reported to the expert (Gains, 1996), especially since it is unlikely that it can be directly represented in the system (see above). Textual reporting can produce an overwhelming amount of data in an indigestible form. Visual approaches to data mining and knowledge discovery are therefore becoming popular and form part of a growing arsenal of visualization methods by which complex data may be depicted and explored (see below).

- Visualization

Visual approaches that might support data mining and knowledge discovery have arisen independently in the statistics and database communities as well as within many other branches of science (Gahegan et al., 2001 provide a more detailed overview). However, the terms used to describe these approaches differ by community. Within the database community, the phrase 'visual data mining' is used to describe vast datasets rendered in some summarized form (e.g. Keim and Kriegel, 1996; Card et al., 1998; Ribarsky et al., 1999). Statisticians, on the other hand, use the term ‘exploratory data analysis’, but this also includes statistical techniques as well as graph-based visual methods (e.g. Tukey, 1977; Asimov, 1985; Tufte, 1990; Haslett et al. 1991; Mihalisin et al., 1991). These strands are largely convergent, aiming to capitalize on the pattern recognition of human experts. But perhaps even more important are the rich cognitive structures and mental models that human experts can apply to provide hypotheses and theories to help explain outcomes of computational knowledge discovery (Valdez-Perez, 1999).

Some visualization tools have recently been developed to directly support spatial data mining and knowledge construction activities, such as the selection of useful data dimensions and the search for structure or pattern (e.g. Lee & Ong, 1996; Keim & Herrmann, 1998; MacEachren et al., 1999; Gahegan et al., 2000). Useful overviews of visual data mining are provided by Wong (1999) and Hinneburg et al. (1999).

Summary of techniques and approaches

A summary of the intersection of academic communities and the knowledge discovery tasks of: *finding* structure, *reporting* and *representing* the findings, *validating* their significance and *optimizing* computational performance are given in Table 1 below. though not exhaustive, this table indicates some of the key research initiatives and directions. The high level of interest in knowledge discovery of late is likely to lead to many additional techniques in the near future. But as is often the case with newer academic areas, there is little research evaluating and comparing techniques as yet, so it is difficult to judge their relative merits for a given application.

Table 1: A summary of DM/KD techniques and approaches

	Databases	**Statistics**	**A. I.**	**Visualization**
Finding	Association rules	Local pattern analysis and global inferential tests	Neural networks, decision trees	Exploratory visualization Visual data mining
Reporting	Rule lists	Significance and power	Likelihood estimation, information gain	A stimulus within the visual domain
Representing	Schema update, metadata	Fitted statistical models, local or global	Conceptual graphs, meta models	Shared between the scene and the observer
Validating	Weak significance testing	Significance tests	Learning followed by verification	Human subjects testing.
Optimizing	Reducing computational complexity	Data reduction and stratified sampling strategies	Stochastic search, gradient ascent methods	Hierarchical and adaptive methods, grand tours

Major Research Perspectives

The increasing ability to capture, store and process digital data and information is not unique to geographic information science: similar information revolutions are occurring in diverse fields such as marketing, biology, astronomy and meteorology, just to name a few. Is there anything special about geographic data that requires unique tools and presents unique research challenges? Will progress in geographic knowledge discovery create broader impacts, leading to a better geographic information science?

In this section of the chapter, we identify the unique aspects of geographic knowledge discovery and its potential impacts on geographic information science and geographic research more broadly. These challenges and impacts can be classified into three main areas, namely, geographic information in knowledge discovery, geographic knowledge discovery in geographic information science and geographic knowledge discovery in geographic research. This section summarizes discussion in Miller and Han (2001); see the original source for more detail and references.

Geographic information in knowledge discovery

Geographic data has unique properties that require special consideration and techniques. First, geographic information exists within highly dimensioned geographic measurement frameworks. While other KDD applications involve highly dimensioned information spaces, geographic data is unique since up to four dimensions of the information space are interrelated and provide the measurement framework for the remaining dimensions. The most commonly adopted measurement framework is the topology and geometry associated with Euclidean space. However, some geographic phenomena have properties that are non-Euclidean; examples include travel times within urban areas, mental images of geographic space and disease propagation over space and time (see Cliff and Haggett 1998; Miller 2000). Projecting geographic data into alternative, more appropriate measurement frameworks can aid the search for patterns in geographic data mining. The information inherent in the geographic measurement framework is often ignored in induction and machine learning tools (Gahegan 2000b).

Measured geographic attributes often exhibit the properties of spatial dependency and spatial heterogeneity. The former refers to the tendency of attributes at some locations in space to be related; typically, these are proximal locations. The latter refers to the non-stationarity of most geographic processes, meaning that global parameters do not reflect well the process occurring at a particular location. While these properties have been traditionally treated as nuisances, contemporary research fueled by advances in geographic information technology provides tools that can exploit these properties for new insights into geographic phenomena (e.g., Anselin 1995; Brunsdon, Fotheringham and Charlton 1996; Fotheringham, Charlton and Brunsodn 1997; Getis and Ord 1992, 1996). Some preliminary research in geographic knowledge discovery suggests that ignoring these properties affects the patterns derived from data mining techniques (Chawla et al. 2001). More research is required on scalable techniques for capturing spatial dependency and heterogeneity in geographic knowledge discovery.

A third unique aspect of geographic information in knowledge discovery is the complexity of spatio-temporal objects and patterns. In most non-geographic domains, data objects can be meaningfully represented as points within the information space without losing important properties. This is often not the case with geographic objects: size, shape and boundaries can affect geographic processes, meaning that geographic objects cannot necessarily be reduced to points without information loss. Relationships such as distance, direction and connectivity are also more complex with dimensional objects (see Egenhofer and Herring 1994; Okabe and Miller 1996; Peuquet and Zhang 1987). Transformations among these objects over time are complex but information-bearing (see Hornsby and Egenhofer 2000). The scales and granularities for measuring time can also be complex, preventing a simple "dimensioning up" of space to include time (Roddick and Lees 2001). Developing scalable tools for extracting patterns from collections of diverse spatio-temporal objects is a critical research challenge. Also, since the complexity of derived spatio-temporal patterns and rules can be daunting, a related challenge is making sense of these derived patterns, perhaps through "meta-mining" of the derived rules and patterns (Roddick and Lees 2001).

The range and diversity of geographic data formats also presents unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data such as imagery and geo-referenced multi-media (see Câmara and Raper 1999). Discovering geographic knowledge from geo-referenced multimedia data is a more complex sibling of the problem of knowledge discovery from multimedia data (see Zaïne et al. 1998).

Geographic knowledge discovery in geographic information science

There are unique needs and challenges for building discovered geographic knowledge in geographic information science. Most digital geographic databases are at best a very simple representation of geographic knowledge at the level of basic geometric, topological and measurement constraints. Knowledge-based GIS attempts to build higher-level geographic knowledge into digital geographic databases for analyzing complex phenomena (see Srinivasan and Richards 1993; Yuan 1997). Geographic knowledge discovery is a potentially rich source for knowledge-based GIS and intelligent spatial analysis. A critical research challenge is developing representations of discovered geographic knowledge that are effective for knowledge-based GIS and spatial analysis.

Geographic knowledge discovery in geographic research

Geographic information has always been a central commodity of geographic research. For much of history, geographic research has occurred within a data-poor environment. Many of the revolutions in geographic research can be tied to improved technologies for georeferencing, capturing, storing and processing geographic data; examples include sailing ships, satellites, clocks, the map and GIS. The current explosion in digital geographic data may be the most dramatic shift in the environment for geographic research in the history of science. This leads to perhaps one of the most important "meta-questions" for geographic research in the 21st century, namely, what are the questions that we could not answer in the past?

The UCGIS Approach

The UCGIS seeks to facilitate a multidisciplinary research effort on the development of geospatial DM/KDD science and technology. As discussed above, the development of DM and KDD technology has opened new avenues in information science research. It also plays an important role in any research endeavor based upon geospatial information.

The ability to mine data pre-supposes that data delivery mechanisms and access mechanisms are in place. While data delivery services are becoming available in local and distributed computing environments, many impediments remain. In addition to research perspectives discussed above, one emphasis for this UCGIS research theme must address infrastructure support for data mining and knowledge discovery. What mechanisms exist is not designed to handle problems specific to geospatial information. Yet, a robust data foundation is critical to promoting a multidisciplinary involvement in the research arena.

Three characteristics of geospatial data create special challenges to development of a robust data foundation. The characteristics that make geospatial data “special” as a computing problem have been acknowledged in many other writings, of course. Moreover, development of a data infrastructure needed to support GIScience in general forms a focus in another UCGIS initiative (spatial data infrastructure). Let us point out that the focus here is not on developing the spatial data infrastructure *per se* but on developing data mining within the emerging infrastructure. As argued below, the research problems solved by generating a solid data foundation can be shown to create the need for new developments in data mining and knowledge discovery. UCGIS researchers have the expertise with geospatial data coupled with an understanding of the limitations of existing and emerging infrastructures. In these respects, our community is best qualified to pursue a research agenda addressing the DM/KD topics in a geospatial context.

The first characteristic relevant to DM/KD is that geospatial data repositories tend to be very large. Data volume was a primary factor in the transition at many federal agencies from delivering public domain data via physical mechanisms (CD ROM, for example) to electronic mechanisms (NRC Data Foundations report). Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management (Healey, 1991). Computational procedures from knowledge discovery must also be diversified if they are to become fully operational within a geospatial computing environment, and this forms an important component of this research theme. Even with deployment of newer integrated GIS data models (such as ESRI’s *geodatabase* data model and Smallworld’s object oriented data model*)*, the hybrid data model will be preserved. In practice, knowledge integration will begin to span not only disparate data models in a single archive, but disparate archives in disparate database management systems.

A second characteristic of geospatial data is the extremely short phase of data, which are collected cyclically. Data discovery must accommodate collection cycles that may be unknown (as for example identifying the cycles of shifts in major geological faults) or that may shift from cycle to cycle in both time and space (for example the dispersion patterns of a health epidemic, or of toxic waste). Integration of information from multiple data models (point source field data with multispectral raster data) is acknowledged as a focus in an established UCGIS research theme (Spatial Data

Acquisition and Integration), and does not need to become a focus of attention here. Instead, knowledge discovery researchers can turn attention to problems of reasoning and modeling on very short temporal cycles. For example, geospatial knowledge discovery could support real-time tornado tracking, or avalanche prediction, or other localized weather events. Infrastructure issues that need to be researched include (for example) the development of real-time data mining, and to utilize knowledge discovery tools to guide correlation of discovered data patterns across time, determination of temporal drift, validation of data trends across temporal discontinuities, and so forth. Because we understand much less about the nature of time than of space, methodologies for archiving data to facilitate cyclic spatial searching remain crude. The extent to which one can identify data patterns will be determined in whole or in part by the organization of the data in an archive. Research on how best to structure data or to reorder data for specific knowledge discovery tasks is not covered in other UCGIS research themes, and must be addressed if a robust geospatial data foundation can develop.

A third characteristic of relevance to DM/KD applies to a characteristic of the data foundation rather than of the data. Emergence of the Internet has supported development of data clearinghouses, digital libraries, and online repositories wherein one does not access data, but pointers to data. It is paradoxical that as increasing amounts of digital data become available via the Internet, they become increasingly difficult to locate, retrieve, and analyze. This is due in large part to the fact that the Internet lacks a comprehensive catalog or index (Buttenfield 1998). Without a coordinating infrastructure, many data sources and services available today remain essentially inaccessible. Currently, over three million Websites are online; and yet even the best search engines can locate only one third of the accessible pages (NPR, 1998; NRC, 1995). Data mining tools need to be established to locate environmental data sources in the “gray literature” areas of the Internet. Such data sources include but are not limited to field data collected in developing countries, very localized community data sites such as inner city neighborhood and community activist sites, and similar data sources not known to or known by conventional doorways into the geospatial data infrastructure. This type of knowledge discovery treats the entire Internet as a very large, decentralized data repository, and provides a venue for contributions to a global information infrastructure.

The decentralization of data delivery via ftp and the Internet revokes many assumptions of what can be known in advance about an archive about to be explored. In addition to the format and data model issues described above, one must consider the semantic issues. Infrastructure support for DM/KD must facilitate thesaurus support. Data definitions are acknowledged to vary widely from agency to agency within a single country. Witness for example the difference between the definition of “address” by 911 Dispatchers (the location of a front door) and by the U.S. Post Office (the location of a mailbox). In urban areas, these two items (front door and mailbox) may be co-located. In rural areas, however, the locations may differ by half a kilometer or more (example from Jack Estes, 1993, personal communication). Knowledge discovery tools working across data sets must be embedded with functions to discriminate semantic differences from errors; and in a decentralized data mining environment, linkages between data and thesauri may not be explicit. Herein lies another important research area for this UCGIS theme.

To summarize, the development of data mining and knowledge discovery tools must be supported by a solid geographic foundation. The emergence of a geospatial data infrastructure has been ad-hoc. Contributed data has not been coupled with contributed tools for data analysis and modeling. Data mining, knowledge discovery methods have not been implemented to deal effectively with geospatial data, whose sensitivities are known widely to geographers. As our understanding of the nature of geographic information and its sensitivities to spatial, temporal and spectral resolution improve, it is probable that refinement of DM algorithms will prove insufficient; and design of new procedures and knowledge validation procedures will begin to emerge. We view the acceptance of the need for new DM/KD designs as one of the primary indicators of the success of the research agenda we propose.

Importance to national research needs

In the information age, geospatial data are collected from diverse sources at a rate that exceeds state-of-the-art capabilities for data management. In a historical context, the amount of information generated in 1999 alone is estimated to be more than 12% of the total volume generated by humankind in all of recorded history (Bradley et al. 2001). Recent studies on the increasing amounts of data collected and stored digitally suggest a widening gap between the volumes of stored data and human's ability to process the data. Moore's Law states that computing power doubles every 18 months. Gray and Shenoy (2000) claim that the cost of storage media drops in half every 9 months. Given this, it is not surprising that data archives will continue to grow and we can expect that the vast majority of archived items will not be accessible by human searches, which are based upon existence of a catalog.

Many human activities could benefit from advances in data mining and knowledge discovery research. In some cases, timely access to archived information could minimize loss of life, harm to geographic or societal groups, and costs to society as a whole. This is particularly compelling in the context of mitigating natural hazards.

New developments in technologies, such as Internet 2, will extend Internet protocols and increase bandwidths for data transmission. It will become possible for people to search larger archives which are globally distributed, according to their own particular needs. As individuals are able to extract data from electronic repositories more effectively, they become more empowered to participate in their own governance. Additionally, they become alert to inconsistencies in archived data. Particularly for data of local geographic significance, this opens the door for grass-roots contributed data updates and correction notices. “*Developing the technical and institutional means to support incorporation of local knowledge into networked repositories presents a novel challenge*” (NRC, 1999: 2).

The actual benefits to the nation of improved access to information are difficult to predict. Many are intangible; others do not have a fixed cost. However, the value of an informed and participatory citizenry to environment and society makes a clear and compelling argument for pursuit of research in data mining and knowledge discovery, and points also to the importance of teaching the nation's youth in utilizing such tools to tap into the vast archives of geospatial information that are and will continue to be archived.

Example Projects

We are still at a very early juncture in the history of geospatial knowledge discovery (GKD). Many attempts are being made to develop new GKD tools and applications. Below, we outline a list of geographic knowledge discovery applications in geographic information science and broader geographic research.

Map interpretation and information extraction

Malebra et al. (2001) demonstrate the use of inductive machine learning tools within a GIS environment. Their system can extract and interpret complex human and physical features from topographic maps for input into a GIS and for analysis.

Information extraction from remotely sensed imagery

The increasing detailed spatial, temporal and spectral resolutions provided by advances in remote sensing technologies are creating massive imagery databases. These databases are overwhelming the ability of researchers to analyze and understand the information implicit within these data. Gopal, Liu and Woodcock (2001) use artificial neural networks combined with visualization techniques to interpret and understand the patterns extracted from remotely sensed images

Mapping environmental features

Many geographic phenomena have complex, multidimensional attributes that are difficult to summarize and integrating using traditional analytical methods. Eklund, Kirkby and Salim (1998) using inductive learning techniques and artificial neural networks to classify and map soil types. Lees and Ritman (1991) use decision tree induction methods for mapping vegetation types in areas where terrain and unusual disturbances (e.g., fire) confound traditional remote sensing classification methods

Extracting spatio-temporal patterns

Identifying unusual patterns in massive spatio-temporal databases can be difficult since the number of possible patterns can be very large. Mesrobian et al. (1996) develop the Open Architecture Scientific Information System (OASIS) for querying, exploring and visualizing geophysical phenomena from large, heterogeneous and distributed databases. The Conquest Scientific Query Processing System, a component of OASIS, identifies cyclonic activity from weather and climate data by extracting unusual patterns in air pressure and winds over time. In another domain, Openshaw and colleagues (Openshaw et al. 1987, Openshaw 1994) develop exploratory techniques based on simple querying and artificial life methods for spotting spatial-temporal clusters in crime data.

Interaction, flow and movement

Spatial interaction, flow and movement in geographic space can provide insights into the spatial structure of physical and human geographic systems. Spatial structure and spatial interaction are intimately related: location influences interaction patterns while interaction patterns influence the location of entities and activities. For tractability purposes, traditional spatial and network analytic make strong assumptions about influences among flow, interaction, movement and location, essentially only capturing direct and proximal effects in space and time. More complex *n*-th order influences may be buried in the massive interaction, flow and movement databases are being captured by real-time monitoring systems, intelligent transportation systems and "position-aware" devices such as cellular telephones and wireless internet clients. Marble et al. (1997) describe visualization methods for exploring massive interaction matrices. Smyth (2001) explores the possibilities for geographic knowledge discovery from the space-time trajectories of mobile devices.

Research frontiers and Priority areas for research

There are several critical research challenges in geographic knowledge discovery and data mining. Miller and Han (2000) offer the following list of emerging research topics in the field:

Developing and supporting geographic data warehouses

To date, a true geographic data warehouse (GDW) does not exist. Spatial properties are often reduced to simple aspatial attributes in mainstream data warehouses. Creating an integrated GDW requires solving issues in spatial and temporal data interoperability, including differences in semantics, referencing systems, geometry, accuracy and position.

Better spatio-temporal representations in geographic knowledge discovery

Current GKD techniques generally use very simple representations of geographic objects and spatial relationships. Geographic data mining techniques should recognize more complex geographic objects (lines and polygons) and relationships (non-Euclidean distances, direction, connectivity and interaction through attributed geographic space such as terrain). Time needs to be more fully integrated into these geographic representations and relationships.

Geographic knowledge discovery using diverse data types

GKD techniques should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia.

User interfaces for geographic knowledge discover

GKD needs to move beyond technically-oriented researchers to the broader GIScience and other research communities. This requires interfaces and tools that can aid diverse researchers in applying these techniques to substantive questions.

Proof of concepts and benchmarking

As in other KDD and DM domains, there needs to be some definitive test cases or benchmarks to illustrate the power and usefulness of GKD to discover unexpected and surprising geographic knowledge. A related issue is benchmarking to determine the effects of varying data quality on discovered geographic knowledge.

Building discovered geographic knowledge into GIS and spatial analysis

We require effective representations of discovered geographic knowledge that are suitable for GIS and spatial analysis. This may include inductive geographic databases, online analytical processing (OLAP)-based GIS interfaces and intelligent tools for guiding spatial analysis.

Expanding upon the above list, we propose the following research agenda in geospatial data mining and knowledge discovery:

Short-term objectives

- Apply DM and KD techniques to the new generations of geospatial data models and identify analytical and visualizational needs for geospatial DM and KD;
- Survey the existing spatial analysis methods, evaluate their potential for large data sets, and, when appropriate, extend their computational abilities in large data sets;
- Apply data warehousing techniques and models to the geographic context and examine methodologies for distributed databases and distributed processing that accommodate the spatial nature of both the data and potential retrieval queries.

Medium-term objectives

- Develop a taxonomy of geographic knowledge and categorize models (methods) for geographic information computing;
- Develop a system for geographic knowledge acquisition and synthesis;
- Develop robust spatial and temporal representations and develop algorithms to automate complex geographic queries in large, distributed, heterogeneous, and dynamic databases;

- Develop robust spatial and temporal reasoning and analytical models to support geographic knowledge formulation through interactive and recursive query processes;
- Develop multi-dimensional, interactive visualization techniques with dynamic links to distributed GIS databases to greatly enhance user's capabilities to detect hidden patterns and inspect potential correlations among geographic variables.

Long-term objectives

- Develop an integrated theory for geographic information representation, processing, analysis, and visualization. The theory will suggest the best geographic representation, analytical methods, and visualization techniques to extract the highest level of geographic information and knowledge in a GIS database;
- Enable a full implementation of geographic knowledge discovery across distributed databases that allow the general public to inspect climate patterns and regional demographic dynamics, for example, on the internet.

Acknowledgement

We would like to thank the forty participants of the one-day workshop on geospatial data mining and knowledge discovery at UCGIS Summer Assembly in Buffalo June 20, 2001, in particular, Dr. Robert Edsall, Dr. Kathleen Hornsby, Dr. Lin Lou, and Dr. Shashi Shekhar. Comments and suggestions from the forty participants have strengthened the manuscript. Yuan's effort is in part supported by the National Science Foundation under Grant No. 0074620.

References:

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD*, pp. 207-216.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Boston, USA, Academic Press.

Anselin, L. (1995). Local indicators of spatial association - LISA, *Geographical Analysis*, 27, 93-115.

Anselin L, (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association, in *Spatial Analytical Perspectives on GIS* Eds M Fischer, H Scholten and D Unwin (London: London), pp. 111-125

Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM Journal of Science and Statistical Computing*, Vol. 6, pp. 128-143.

Berry, M. J. A. and Lino, G. (1997) *Data Mining Techniques For Marketing, Sales, and Customer Support*, New York, NY: John Wiley and Sons.

Bhandari, E. Colet, E., Parker, J., Pines, Z, Pratap, R., Pratap, R. and Ramanujam, K. Advanced scout: data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery*, 1, 121-125.

Bradley, P.S., Fayyad, U., Sarawagi, S., and Shim, K. 2001. Editorial. *SIGKDD Explorations*, 2, 2, i-iii.

Bradsil, P. B. and Konolige, K. (Eds.) (1990). *Meta-Learning, Meta-Reasoning and Logics*. Boston, MA, USA, Kluwer Academic Press.

Buttenfield, B.P. 1998 Looking Forward: Geographic Information Services and Libraries in the Future. *Cartography and GIS***,** 25(3): 161-171.

Brunsdon, C., Fotheringham, A. S. and Charlton, M. E. (1996) "Geographically weighted regression: A method for exploring spatial nonstationarity," *Geographical Analysis*, 28 281-298.

Card, S., Mackinlay, J. and Shneiderman, B. (1998). Information visualization. Readings in Information Visualization (Eds. Card, S., Mackinlay, J. and Shneiderman, B.), San Francisco: Morgan-Kaufmann, pp. 1-34.

Câmara, A. S. and Raper, J. (eds.) (1999) *Spatial Multimedia and Virtual Reality*, London: Taylor and Francis.

Cheeseman, P. and Stutz, J. (1996). Bayesian Classification: Theory and results. In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press, pp. 153-189.

Chawla, S., Shekhar, S., Wu, W. L. and Ozesmi, U. (2001) "Modeling spatial dependencies for mining geospatial data: An introduction," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, in press.

Cliff, A. D. and Haggett, P. (1998) "On complex geographical space: Computing frameworks for spatial diffusion processes," in P. A. Longley, S. M. Brooks, R. McDonnell and B. MacMillan (eds.) *Geocomputation: A Primer*, Chichester, U.K.: John Wiley and Sons, 231-256.

Cohen, W. W. (1995). "Fast, effective rule induction." Proc. *12th International Conference on Machine Learning*, San Francisco, California, USA, Morgan-Kaufmann, pp. 115-123.

Devijver, P. A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*, London, Prentice-Hall International.

Drew, P. and Ying, J. (1998). Metadata management for geographic information discovery and exchange. In: Sheth, A. and Klas, W. (Eds.). Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media. McGraw-Hill, pp. 89-121.

Dunteman, G. H. (1984). *Introduction to Multivariate Analysis*, Beverly Hills, CA, USA, Sage.

Egenhofer, M. J. and Herring, J. R. (1994) "Categorizing binary topological relations between regions, lines and points in geographic databases," in M. Egenhofer, D. M. Mark and J. R. Herring (eds.) *The 9-intersection: Formalism and its Use for Natural-language Spatial Predicates*, National Center for Geographic Information and Analysis Technical Report 94-1, 1-28.

Eklund, P. W., Kirkby, S. D. and Salim, A. (1998) "Data mining and soil salinity analysis," *International Journal of Geographical Information Science*, 12, 247-268.

Elder, J. F. and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press, pp. 83-113

Ester M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231.

Ester, M., Kriegel, H.-P. and Sander, J. (1998). Algorithms for characterization and trend detection in spatial databases. Proc. *4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New York, USA, pp. 44-50.

Fayyad, U. 1997. Editorial. *Data Mining and Knowledge Discovery*. 1, 5-10.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, Fall 1996, pp. 37-54.

Fotheringham, A. S., Charlton, M. and Brunsdon, C. (1997) "Two techniques for exploring non-stationarity in geographical data," *Geographical Systems*, 4, 59-82

Gahegan, M. (2000a). The case for inductive and visual techniques in the analysis of spatial data. *Geographical Systems*, 7(2) 77-83.

Gahegan, M. (2000b) "On the application of inductive machine learning tools to geographical analysis," *Geographical Analysis*, 32, 113-139

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2000.) GeoVISTA *Studio*: A Geocomputational Workbench. Proc. 4th Annual Conference on GeoComputation, UK, August 2000. URL: http://www.ashville.demon.co.uk/gc2000/.

Gahegan, M., Wachowicz, M., Harrower, M. and Rhyne, T.-M. (2001). The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Systems* (special issue on the ICA research agenda).

Gains, B. R., (1996). Transforming Rules and Trees into Comprehensible Knowledge Structures, In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press.

Gardner, C. (1996). *IBM Data Mining Technology*. IBM Corporation, Stamford, Connecticut.

Gray, J. and Shenoy, P. Rules of thumb in data engineering. *Proceedings of 16th International Conference on Data Engineering*. Pp. 3-13. IEEE, 2000.

Gehrke, J., Ganti, V., Ramrakrishnan, R. and Loh, W.-Y. (1999). BOAT—Optimistic decision tree construction. Proc. *SIGMOD 1999*. ACM Press: New York, pp. 169-180.

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1996). Statistical inference and data mining. *Communications of the ACM*, 39(11):35-41.

Glymour, C., Madigan, D., Pregibon, D., and Smyth P. (1997) Statistical themes and lessons for data mining, *Journal of Data Mining and Knowledge Discovery*, Vol. 1, pp. 11-28.

Getis, A. and Ord, J. K. (1992) "The analysis of spatial association by use of distance statistics," *Geographical Analysis*, 24, 189-206.

Getis, A. and Ord, J. K. (1996) "Local spatial statistics: An overview," in P. Longley and M. Batty (eds.) *Spatial Analysis: Modelling in a GIS Environment*, Cambridge, UK: GeoInformation International, 261-277

Gopal, S., Liu, W. and Woodcock, C. (2001) "Visualization based on fuzzy ARTMAP neural network for mining remotely sensed data," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery,* London: Taylor and Francis, in press.

Goodchild, M.F., Buttenfield, B.P., Adler, P., Krygiel, A., Onsrud, H., Kahn, R. (1999). *Distributed Geolibraries.* National Research Council Monograph. Washington, D.C.: National Academy Press.

Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, Vol. 45, No. 3, pp. 234-242.

Hedberg, S. R. 1996. Search for the mother lode: tales of the first data miners. *IEEE Expert*, 11(5): 4-7.

Hinneburg, A., Keim, D. and Wawryniuk, M. (1999). HD-Eye: Visual mining of high dimensional data. IEEE Computer Graphics and Applications, September/October 1999, pp. 22-31.

Healey, R. (1991). Database Management Systems. In Maguire, D., Goodchild, M.F., and Rhind, D. (Eds.) *Geographic Information Systems: Principles and Applications.* London: Longman.

Hornsby, K. and Egenhofer, M. J. (2000) "Identity-based change: A foundation for spatio-temporal knowledge representation," *International Journal of Geographical Information Science*, 14, 207-224.

Lees, B. G. and Ritman, K. (1991) "Decision-tree and rule-induction approach approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments," *Environmental Management*, 15, 823-831.

Lee, H. and Ong, H. (1996). Visualization support for data mining. *IEEE Expert*, 11(5):69-75.

Lmielinski, T. and Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, 39(11):58-64.

Lu, H., Setiono, R., and Liu, H. 1996. Effective data mining using neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):957-961.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, USA, Wiley.

Keim, D. and Kriegel, H.-P.(1996) Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering* (Special Issue on Data Mining),

Knorr, E. M. and Ng, R. T., (1996). Finding aggregate proximity relationships and commonalities in spatial data mining, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 884-897.

Koperski, K. and Han, J. (1995), Discovery of Spatial Association Rules in Geographic Information Databases, Proc. *4th International Symposium on Large Spatial Databases* (SSD95), Maine, pp. 47-66.

Koperski, K. Han, J. and Adhikary, J. (1999). Mining knowledge in geographic data. *Comm. ACM* (to appear). URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html

Luger, G. F. and Stubblefield, W. A (1998). *Artificial Intelligence: structures and strategies for complex problem solving*. Reading, MA: Addison-Wesley.

MacDougall, E. B. (1992). Exploratory analysis, dynamic statistical visualization and geographic information systems. *Cartography and Geographical Information Systems*, Vol. 19, No. 4, pp. 237-246.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D. and Masters, R. (1999). Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 311-334.

Mardia, K. V., Kent, T. and Bibby, J. M. (1979). *Multivariate Analysis*, London, UK, Academic Press.

Martin, J. C. (1991). *Introduction to Languages and the Theory of Computation*. McGraw Hill, New York, USA.

Mihalisin, T. Timlin, J. and Schwegler, J. (1991). Visualizing multivariate functions, data and distributions. *IEEE Computer Graphics and Applications*, Vol. 19, No. 13, pp. 28-35.

Miller, H. and Han, J. (Eds.) (2001). *Geographic Data Mining and Knowledge Discovery*, London: Taylor & Francis.

Mitchell, T. M. (1997). *Machine Learning*, New York, USA, McGraw Hill.

Malerba, D., Esposito, F. Lanza, A., Lisi., F. A. (2001) "Machine learning for information extraction from topographic maps," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery,* London: Taylor and Francis, in press.

Marble, D. F., Gou, Z., Liu, L. and Saunders, J. (1997) "Recent advances in the exploratory analysis of interregional flows in space and time," in Z. Kemp (ed.) *Innovations in GIS 4*, London: Taylor and Francis.

Mesrobian, E, Muntz, R., Shek, E., Nittel, S., La Rouche, M., Kriguer, M., Mechoso, C., Farrara, J., Stolorz, P. and Nakamura, H. (1996) "Mining geophysical data for knowledge," *IEEE Expert*, 11(5), 34-44.

Miller, H. J. (2000) "Geographic representation in spatial analysis," *Journal of Geographical Systems*, 2, 55-60.

Miller, H. J. and Han, J. (2000) “Discovering geographic knowledge in data rich environments: A report on a specialist meeting,” *SIGKDD Explorations: Newsletter of the (Association for Computing Machinery) Special Interest Group on Knowledge Discovery and Data Mining*, 1(2), 105-108; available at http://www.acm.org/sigs/sigkdd/explorations/

Miller, H. J. and Jiawei, H. (2001) "Geographic data mining and knowledge discovery: An overview," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery,* London: Taylor and Francis, in press.

Murray, A. T. and Estivill-Castro, V. (1997). Cluster discovery techniques for exploratory spatial data analysis. *International journal of Geographical Information Science*, 12(5), 431-443.

NPR 1998 National Public Radio Morning Edition, report on emergence of commercial Internet search engines such as Yahoo and other “dot-coms”, 3 April, 1998.

NRC National Research Council 1999 *Distributed Geolibraries: Spatial Information Resources. Panel on Distributed Geolibraries, Goodchild, M.F. Chair.* Washington, D.C.: National Academy Press .

NRC National Research Council 1995 *Data Foundation for the National Spatial Data Infrastructure***.** Mapping Science Committee, Sugarbaker, L.A., Chair. Washington, D.C.: National Academy Press .

Okabe, A. and Miller, H. J. (1996) "Exact computational methods for calculating distances between objects in a cartographic database," *Cartography and Geographic Information Systems*, 23, 180-195.

Openshaw, S., Cross, A. and Charlton, M. (1990), Building a Prototype Geographical Correlates Machine. *International Journal of Geographical Information Systems*, Vol. 4, No. 4, pp. 297-312.

Openshaw, S. (1994) "Two exploratory space-time-attribute pattern analysers relevant to GIS," in A. S. Fotheringham and P. A. Rogerson (eds.) *Spatial Analysis and GIS*, London: Taylor and Francis, 83-104.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987) "A mark 1 geographical Analysis Machine for automated analysis of point data sets," *International Journal of Geographical Information Systems*, 1, 335-358.

Peuquet, D. J. and Zhang, C.-X.. (1987) "An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane," *Pattern Recognition*, 20, 65-74.

Pierce, C. S. (1878). "Deduction, induction and hypothesis." *Popular Science Monthly*, 13, 470-482.

Rainsford, C. P., and Roddick, J. F. (1999). Database issues in knowledge discovery and data mining, *Australian Journal of Information Systems*, Vol. 6, No. 2, pp. 101-128.

Ribarsky, W., Katz, J. and Holland, A. (1999). Discovery visualization using fast clustering. *IEEE Computer Graphics and Applications*, September/October 1999, pp. 32-39.

Roddick, J. F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research. SIGKDD Explorations. Vol. 1, No. 1, URL: http://www.cis.unisa.edu.au/~cisjfr/STDMPapers/.

Roddick , J. F. and Lees, B. (2001) "Paradigms for spatial and spatio-temporal data mining," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, in press.

Smythe, P. (2000). Data mining: Data analysis on a grand scale? *Statistical Methods in Medical Research*, September, 2000.

Sowa, J. F. (1999). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Srinivasan, A. and Richards, J. A. (1993) "Analysis of GIS spatial data using knowledge-based methods," *International Journal of Geographical Information Systems*, 7, 479-500.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley.

Valdez-Perez, R. E. (1999). Principles of human computer collaboration for knowledge discovery in science. Artificial Intelligence, Vol. 107, No. 2, pp. 335-346.

Wong, P. C. (1999) Visual data mining. *IEEE Computer Graphics and Applications*, Vol. 19, No. 5, pp. 20-21.

Yuan, M. (1997) "Use of knowledge acquisition to build wildfire representation in geographic information systems," *International Journal of Geographical Information Systems*, 11, 723-745.

Yuan, M. (1998). Representing Spatiotemporal Processes to Support Knowledge Discovery in GIS databases. Proceedings: 8th International Symposium on Spatial

Data Handling Spatial Data Handling, T. K. Poiker and N. Chrisman, eds. Pp. 431-440.

Zaïane, O. R., Han, J., Li, Z.-N. and Hou, J. (1998) "Mining Multimedia Data", in *Proceedings, CASCON'98*: Meeting of Minds, Toronto, Canada, November 1998; available at: http://db.cs.sfu.ca/sections/publication/smmdb/smmdb.html.