GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY: AN OVERVIEW

Harvey J. Miller
Department of Geography
University of Utah
harvey.miller@geog.utah.edu

Jiawei Han
School of Computing Science
Simon Fraser University
han@cs.sfu.ca

Acknowledgments: Thanks to Mark Gahegan and Phoebe McNeally for some helpful comments on this chapter.

1. INTRODUCTION

Similar to many research and application fields, geography has moved from a data-poor and computation-poor to a data-rich and computation-rich environment. The scope, coverage and volume of digital geographic datasets are growing rapidly. Public and private sector agencies are creating, processing and disseminating digital data on land use, socioeconomic and infrastructure at very detailed levels of geographic resolution. New high spatial and spectral resolution remote sensing systems and other monitoring devices are gathering vast amounts of geo-referenced digital imagery, video, and sound. Geographic data collection devices linked to global positioning system receivers allow field researchers to collect unprecedented amounts of data. Position aware devices such as cell phones, in-vehicle navigation systems and wireless Internet clients allow tracking of individual movement behavior in space and time. Information infrastructure initiatives such as the U. S. National Spatial Data Infrastructure are facilitating data sharing and interoperability. Digital geographic data repositories on the World Wide Web are growing rapidly in both number and scope. The amount of data that geographic information processing systems can handle will continue to increase exponentially through the mid-21st century.

Traditional spatial analytical methods were developed in an era when data collection was expensive and computational power was weak. The increasing volume and diverse nature of

digital geographic data easily overwhelm mainstream spatial analysis techniques that are oriented towards teasing scarce information from small and homogenous datasets. Traditional statistical methods, particularly spatial statistics, have high computational burdens. These techniques are confirmatory and require the researcher to have *a priori* hypotheses. Therefore, traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse geographic datasets.

In March 1999, the National Center for Geographic Information and Analysis (NCGIA) – Project Varenius held a workshop on "Discovering geographic knowledge in data-rich environments" in Kirkland, Washington. The workshop brought together a diverse group of stakeholders with interests in developing and applying computational techniques for exploring large, heterogeneous digital geographic datasets. This includes geographers, geographic information scientists, computer scientists and statisticians. This book is a result of that workshop. This volume brings together some of the cutting-edge research from the diverse stakeholders working in the area of geographic data mining and geographic knowledge discovery in a data-rich environment.

This chapter provides an introduction to geographic data mining and geographic knowledge discovery (GKD). In this chapter, we provide an overview of knowledge discovery from databases (KDD) and data mining. We also provide an overview of the highly interesting special case of geographic knowledge discovery and geographic data mining. We identify why geographic data is a non-trivial special case that requires special consideration and techniques. We also review the current state-of-the-art in GKD, including the existing literature and the contributions of the chapters in this volume.

2. KNOWLEDGE DISCOVERY AND DATA MINING

In this section of the chapter, we provide a general overview of knowledge discovery and data mining. We begin with an overview of knowledge discovery from databases (KDD), highlighting

its general objectives and its relationship to the field of statistics and the general scientific process. We then identify the major stages of the KDD processing, including data mining. We classify major data mining tasks and discuss some techniques available for each task. We conclude this section by discussing the relationships between scientific visualization and KDD.

2.1. Knowledge discovery from databases

Knowledge discovery from databases (KDD) is a response to the enormous volumes of data being collected and stored in operational and scientific databases. Continuing improvements in information technology (IT) and its widespread adoption for process monitoring and control in many domains is creating a wealth of new data. There is often much more information in these databases than the "shallow" information being extracted by traditional analytical and query techniques. KDD leverages investments in IT by searching for deeply hidden information that can be turned into knowledge for strategic decision-making and answering fundamental research questions.

KDD is better known through the more popular term "data mining." However, data mining is only one component (albeit a central component) of the larger KDD process. Data mining involves distilling data into *information* or facts about the mini-world described by the database. KDD is the higher-level process of obtaining information through data mining and distilling this information into *knowledge* (ideas and beliefs about the mini-world) through interpretation of information and integration with existing knowledge.

KDD is based on a belief that information is hidden in very large databases in the form of *interesting patterns*. These are non-random properties and relationships that are valid, novel, useful and ultimately understandable. *Valid* means that the pattern is general enough to apply to new data; it is not just an anomaly of the current data. *Novel* means that the pattern is non-trivial and unexpected. *Useful* implies that the pattern should lead to some effective action: rather than searching for any valid and novel pattern, KDD should inform decision making and scientific

investigation. *Ultimately understandable* means that the pattern should be simple and interpretable by humans (Fayyad, Piatetsky-Shapiro and Smyth 1996).

KDD is also based on the belief that traditional database queries and statistical methods cannot reveal interesting patterns in very large databases. One reason is the type of data that increasingly comprise enterprise databases. Another reason is the novelty of the patterns sought in KDD.

KDD goes beyond the traditional domain of statistics to accommodate data not normally amenable to statistical analysis. Statistics usually involves a small and clean (noiseless) numeric database scientifically sampled from a large population with specific questions in mind. Many statistical models require strict assumptions (such as independence, stationarity of underlying processes and normality). In contrast, the data being collected and stored in many enterprise databases are noisy, non-numeric and possibly incomplete. These data are also collected in an open-ended manner without specific questions in mind (Hand 1998). KDD encompasses principles and techniques from statistics, machine learning, pattern recognition, numeric search and scientific visualization to accommodate the new data types and data volumes being generated through information technologies.

KDD is more strongly inductive than traditional statistical analysis. The generalization process of statistics is embedded within the broader deductive process of science. Statistical models are confirmatory, requiring the analyst to specify a model *a priori* based on some theory, test these hypotheses and perhaps revise the theory depending on the results. In contrast, the deeply hidden, interesting patterns being sought in a KDD process are (by definition) difficult or impossible to specify *a priori*, at least with any reasonable degree of completeness. KDD is more concerned about prompting investigators to formulate *new* predictions and hypotheses from data as opposed to testing deductions from theories through a sub-process of induction from a scientific database (Elder and Pregibon 1996; Hand 1998). A rule-of-thumb is that if the

information being sought can only be vaguely described in advance, KDD is more appropriate than statistics (Adriaans and Zantinge 1996).

KDD more naturally fits in the initial stage of the deductive process when the researcher forms or modifies theory based on ordered facts and observations from the "real world." In this sense, KDD is to information space as microscopes, remote sensing and telescopes are to atomic, geographic and astronomical spaces, respectively: KDD is a tool for exploring domains that are too difficult to perceive with unaided human abilities. For searching through a large information wilderness, the powerful but focused laser beams of statistics cannot compete with the broad but diffuse floodlights of KDD. However, floodlights can cast shadows and KDD cannot compete with statistics in confirmatory power once the pattern is discovered.

2.2. Data warehousing

An infrastructure that often underlies the KDD process is the *data warehouse* (DW). A DW is a repository that integrates data from one or more source databases. The data-warehousing phenomenon results from several technological and economic trends, including the decreasing cost of data storage and data processing, and the increasing value of information in business, governmental and scientific environments. A DW usually exists to support strategic and scientific decision-making based on integrated, shared information, although DWs are also used to save legacy data for liability and other purposes (see Jarke at al. 2000).

The data in a DW are usually read-only historical copies of the operational databases in an enterprise, sometimes in summary form. Consequently, a DW is often several orders of magnitude larger than an operational database (Chaudhuri and Dayal 1997). Rather than just a very large database management system, a DW embodies very different database design principles than operational databases.

Operational database management systems are designed to support *transactional data processing*, that is, data entry, retrieval and updating. Design principles for transactional database

systems attempt to create a database that is internally consistent and recoverable (i.e., can be "rolled-back" to the last known internally consistent state in the event of an error or disruption). These objectives must be met in an environment where multiple users are retrieving and updating data. For example, the normalization process in relational database design decomposes large, "flat" relations along functional dependencies to create smaller, parsimonious relations that logically store a particular item a minimal number of times (ideally, only once; see Silberschatz, et al. 1997). Since data are stored a minimal number of times, there is a minimal possibility of two data items about the same real-world entity disagreeing (e.g., if only one item is updated due to user error or an ill-timed system crash).

In contrast to transactional database design, good DW design maximizes the efficiency of *analytical data processing* or data examination for decision making. Since the DW contains read-only copies and summaries of the historical operational databases, consistency and recoverability in a multi-user transactional environment are not issues. The database design principles that maximize analytical efficiency are contrary to those that maximize transactional stability. Acceptable response times when repeatedly retrieving large quantities of data items for analysis require the database to be non-normalized and connected; examples include the "star" and "snowflake" logical DW schemas (see Chaudhuri and Dayal 1997). The DW is in a sense a buffer between transactional and analytical data processing, allowing efficient analytical data processing without corrupting the source databases (Jarke et al. 2000).

In addition to data mining, a DW often supports *online analytical processing* (OLAP) tools. OLAP tools provide multidimensional summary views of the data in a DW. OLAP tools allow the user to manipulate these views and explore the data underlying the summarized views. Standard OLAP tools include *roll-up* (increasing the level of aggregation), *drill-down* (decreasing the level of aggregation), *slice* and *dice* (selection and projection) and *pivot* (re-orientation of the multidimensional data view) (Chaudhuri and Dayal 1997). OLAP tools are in a sense a type of "super-queries": more powerful than standard query language such as SQL but shallower than

data mining techniques since they do not reveal hidden patterns. Nevertheless, OLAP tools can be an important part of the KDD process. For example, OLAP tools can allow the analyst to achieve a synoptic view of the DW that can help specify and direct the application of data mining techniques (Adriaans and Zantinge 1996).

A powerful and commonly applied OLAP tool for multidimensional data summary is the *data cube*. Given a particular measure (e.g., "sales") and some dimensions of interest (e.g., "item", "store," "week") a data cube is an operator that returns the power set of all possible aggregations of the measure with respect to the dimensions of interest. These include aggregations over 0-dimensions (e.g., "total sales"), 1-dimension (e.g., "total sales by item," "total sales by store", "total sales per week"), 2-dimensions (e.g., "total sales by item and store") and so on up to N-dimensions. (In the present example, $N = 3$, with the corresponding aggregations "total sales by item and store and region"). The data cube is an N-dimensional generalization of the more commonly known SQL aggregation functions and "Group-By" operator. However, the analogous SQL query only generates the zero and one-dimensional aggregations; the data cube operator generates these and the higher dimensional aggregations all at once (Gray et al. 1997).

The power set of aggregations over selected dimensions is called a "data cube" since the logical arrangement of aggregations can be viewed as a hypercube in an N-dimensional information space (see Gray et al. 1997, Figure 2; Shekhar et al. this volume). The data cube can be pre-computed and stored in its entirety, computed "on-the-fly" only when requested, or partially pre-computed and stored (see Harinarayan, Rajaman and Ullman 1996). The data cube can support standard OLAP operations including roll-up, drill-down, slice, dice and pivot operations on measures computed by different aggregation operators, such as max, min, average, top-10, variance, and so on.

2.3. The KDD process and data mining

The KDD process usually consists of several generic steps, namely, data selection, data pre-processing, data enrichment, data reduction and projection, and interpretation and reporting. These steps may not be necessarily executed in linear order. Stages may be skipped or revisited. Ideally, KDD should be a human-center process based on the available data, the desired knowledge and the intermediate results obtained during the process (see Adriaans and Zantinge 1996; Brachman and Anand 1996; Fayyad, Piatetsky-Shapiro and Smyth 1996; Matheus, Chan and Piatetsky-Shapiro 1993).

Data selection refers to determining a subset of the records or variables in a database for knowledge discovery. Particular records or attributes are chosen as foci for concentrating the data mining activities. Automated data reduction or "focusing" techniques are also available (see Barbara et al. 1997, Reinartz 1999). *Data pre-processing* involves "cleaning" the selected data to remove noise, eliminating duplicate records, and determining strategies for handling missing data fields and domain violations. The pre-processing step may also include *data enrichment* through combining the selected data with other, external data (e.g., census data, market data). *Data reduction and projection* concerns both dimensionality and numerosity reductions to further reduce the number of attributes or tuples or transformations to determine equivalent but more efficient representations of the information space. Smaller, less redundant and more efficient representations enhance the effectiveness of the *data mining* stage that attempts to uncover the information (interesting patterns) in these representations. The *interpretation and reporting* stage involves evaluating, understanding and communicating the information discovered in the data mining stage.

Data mining refers to the application of low-level algorithms for revealing hidden information in a database (Klösgen and Żytkow 1996). There are many types of data mining techniques and many ways to classify these techniques. Table 1-1 provides a possible classification of data mining tasks and techniques. See Matheus, Chan and Piatetsky-Shapiro (1993), Fayyad, Piatetsky-Shapiro and Smyth (1996) as well as several of the chapters in this

current volume for other overviews and classifications of data mining techniques. Also see Goebel and Gruenwald (1999) for an overview of techniques and a survey of available software tools for KDD and data mining.

Data mining task	Description	Techniques
Segmentation	*Clustering:* Determining a finite set of implicit classes that describes the data. *Classification:* Mapping data items into predefined classes	• Cluster analysis • Bayesian classification • Decision or classification trees • Artificial neural networks
Dependency analysis	Finding rules to predict the value of some attribute based on the value of other attributes	• Bayesian networks • Association rules
Deviation and outlier analysis	Finding data items that exhibit unusual deviations from expectations	• Clustering and other data mining methods • Outlier detection
Trend detection	Lines and curves summarizing the database, often over time	• Regression • Sequential pattern extraction
Generalization and characterization	Compact descriptions of the data	• Summary rules • Attribute-oriented induction

Table 1-1: Data mining tasks and techniques

Segmentation involves partitioning the selected data into meaningful groupings or classes. This can require two major subtasks. *Clustering* determines a finite set of implicit classes that describe the database by examining relationships between data items. *Classification* refers to finding rules to assign data items into pre-existing classes. Some authors consider

clustering and classification to be separate data mining tasks. However, there can be a great deal of overlap and therefore we consider them together as two subtasks of the larger "segmentation" task.

The commonly used data mining technique of *cluster analysis* determines a set of classes and assignments to these classes based on the relative proximity of data items in the information space. Cluster analysis methods for data mining must accommodate the large data volumes and high dimensionalities of interest in data mining; this usually requires statistical approximation or heuristics (see Farnstrom, Lewis and Elkan 2000; Han, Kamber and Tung, this volume). *Bayesian classification* methods, such as AutoClass, determine classes and a set of weights or class membership probabilities for data items (see Cheesman and Stutz 1996). *Decision or classification trees* are hierarchical rule sets that generate an assignment for each data item with respect to a set of known classes. Entropy-based methods such as ID3 and C4.5 (Quinlan 1986, 1992) derive these classification rules from training examples. Statistical methods include the Chi-square Automatic Interaction Detector (CHAID) (Kass 1980) and the Classification and Regression Tree (CART) method (Beiman et al. 1984). Artificial neural networks (ANN) can be used as non-linear clustering and classification techniques. Unsupervised ANNs such as Kohonen Maps are a type of neural clustering where weighted connectivity after training reflects proximity in information space of the input data (see Flexer 1999). Supervised ANNs such as the well-known feedforward/backpropagation architecture require supervised training to determine the appropriate weights (response function) to assign data items into known classes.

Dependency analysis involves finding rules to predict the value of some attribute based on the value of other attributes (Ester, Kriegel and Sander 1997). *Bayesian networks* are graphical models that maintain probabilistic dependency relationships among a set of variables. These networks encode a set of conditional probabilities as directed acyclic networks with nodes representing variables and arcs extending from cause to effect. We can infer these conditional probabilities from a database using several statistical or computational methods depending on the

nature of the data (see Buntine 1996; Heckerman 1997). *Association rules* are a particular type of dependency relationship. An association rule is an expression $X \Rightarrow Y$ (*c*%, *r*%) where *X* and *Y* are disjoint sets of items from a database, *c*% is the *confidence* and *r*% is the *support*. Confidence is the proportion of database transactions containing *X* that also contain *Y*; in other words, the conditional probability $P(Y|X)$. Support is proportion of database transactions that contain *X* and *Y*, i.e., the union of X and Y, $P(X \cup Y)$ (see Hipp, Güntzer and Nakhaeizadeh 2000). Mining association rules is a difficult problem since the number of potential rules is exponential with respect to the number of data items. Algorithms for mining association rules typically use breadth-first or depth-first search with branching rules based on minimum confidence or support thresholds (see Agrawal et. al 1996; Hipp, Güntzer and Nakhaeizadeh 2000).

Deviation and outlier analysis involves searching for data items that exhibit unexpected deviations or differences from some norm. The motivation is that these cases are either errors that should be corrected/ignored or represent unusual cases that are worthy of additional investigation. Outliers are often a byproduct of other data mining methods, particularly cluster analysis. However, rather than treating these cases as "noise," special-purpose outlier detection methods search for these unusual cases as signals conveying valuable information (see Breuing et al. 1999; Ng, this volume).

Trend detection typically involves fitting lines and curves to the data, including linear and logistic regression analysis that are very fast and easy to estimate. These methods are often combined with filtering techniques such as stepwise regression. Although the data often violates the stringent regression assumptions, violations are less critical if the estimated model is used for prediction rather than explanation (i.e., estimated parameters are not used to explain the phenomenon). *Sequential pattern extraction* explores time series data looking for temporal correlations or pre-specified patterns (such as curve shapes) in a single temporal data series (see Agrawal and Srikant 1995; Berndt and Clifford 1996).

Generalization and characterization are compact descriptions of the database. As the name implies, *summary rules* are a relatively small set of logical statements that condense the information in the database. The previously discussed classification and association rules are specific types of summary rules. Another type is a *characteristic rule*: this is an assertion that data items belonging to a specified concept have stated properties, where "concept" is some state or idea generalized from particular instances (Klösgen and Żytkow 1996). An example is "all professors in the applied sciences have high salaries." In this example, "professors" and "applied sciences" are high-level concepts (as opposed to low-level measured attributes such as "assistant professor" and "computer science") and "high salaries" is the asserted property (see Han, Cai and Cercone 1993).

A powerful method for finding many types of summary rules is *attribute-oriented induction* (also known as *generalization-based mining*). This strategy performs hierarchical aggregation of data attributes, compressing data into increasingly generalized relations. Data mining techniques can be applied at each level to extract features or patterns at that level of generalization (Han and Fu 1996). Background knowledge in the form of a *concept hierarchy* provides the logical map for aggregating data attributes. A concept hierarchy is a sequence of mappings from low-level to high-level concepts. It is often expressed as a tree whose leaves correspond to measured attributes in the database and the root representing the null descriptor ("any"). Concept hierarchies can be derived from experts or from data cardinality analysis (Han and Fu 1996).

2.5. Visualization and knowledge discovery

KDD is a complex process. The mining metaphor is appropriate: information is buried deeply in a database and extracting it requires skilled application of an intensive and complex suite of extraction and processing tools. Selection, pre-processing, mining and reporting techniques must be applied in an intelligent and thoughtful manner based on intermediate results and background

knowledge. Despite attempts at quantifying concepts such as "interestingness" (e.g., Silberschatz and Tuzhilin 1996), the KDD process is difficult to automate. KDD requires a high-level, most likely human, intelligence at its center (see Brachman and Anand 1996).

Visualization is a powerful strategy for integrating high-level human intelligence and knowledge into the KDD process. The human visual system is extremely effective at recognizing patterns, trends and anomalies. The visual acuity and pattern spotting capabilities that humans acquired for throwing objects at prey and recognize stalking predators can also be exploited in many stages of the KDD process, including OLAP, query formulation, technique selection and interpretation of results. These capabilities have yet to be surpassed by machine-based approaches (Gahegan 2000b, this volume; Wachowicz, this volume).

Keim and Kriegel (1994) and Lee and Ong (1996) describe software systems that incorporate visualization techniques for supporting database querying and data mining. Keim and Kriegel (1994) use visualization to support simple and complex query specification, OLAP, and querying from multiple independent databases. Lee and Ong's (1996) WinViz software uses multidimensional visualization techniques to support OLAP, query formulation and the interpretation of results from unsupervised (clustering) and supervised (decision tree) segmentation techniques.

3. GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY

This section of the chapter describes a very important special case of KDD, namely, *geographic knowledge discovery* (GKD). We will first discuss why GKD is an important special case that requires careful consideration and specialized tools. We will then discuss geographic data warehousing and online geographic data repositories, the latter an increasingly important source of digital geo-referenced data and imagery. We then discuss geographic data mining techniques and the relationships between GKD and *geographic visualization* (GVis), an increasingly active research domain integrating scientific visualization and cartography. We follow this with

discussions of current GKD applications and research frontiers. Throughout this section, we discuss the existing literature in geographic information science and computer science as well as the contributions of this current volume.

3.1. Why geographic knowledge discovery?

3.1.1. Geographic information in knowledge discovery

The digital geographic data explosion is not much different from similar revolutions in marketing, biology and astronomy. Is there anything special about geographic data that requires unique tools and provides unique research challenges? In this section, we identify and discuss some of the unique properties of geographic data and challenges in geographic knowledge discovery (GKD).

Geographic measurement frameworks. While many information domains of interest in KDD are high dimensional, these dimensions are relatively independent. Geographic information are not only high dimensional but also have the property that up to four dimension of the information space are interrelated and provide the measurement framework for all other dimensions. Formal and computational representations of geographic information require the adoption of an implied topological and geometric measurement framework. This framework affects measurement of the geographic attributes and consequently the patterns that can be extracted (see Beguin and Thisse 1979).

The most common framework is the topology and geometry consistent with Euclidean distance. Euclidean space fits in well with our experienced reality and results in maps and cartographic displays that are useful for navigation. However, geographic phenomena often display properties that are consistent with other topologies and geometries. For example, travel time relationships in an urban area usually violate the symmetry and triangular inequality conditions for Euclidean and other distance metrics. Therefore, seeking patterns and trends in transportation systems (such as congestion propagation over space and time) benefits from

projecting the data into an information space whose spatial dimensions are non-metric. Also, disease patterns in space and time often behave according to other topologies and geometries than Euclidean (see Cliff and Haggett 1998; Miller 2000). The useful information implicit in the geographic measurement framework is ignored in many induction and machine learning tools (Gahegan 2000a).

An extensive toolkit of analytical cartographic techniques is available for estimating appropriate distance measures and projecting geographic information into that measurement framework (see, e.g., Cliff and Haggett 1998; Gatrell 1983; Mueller 1982; Tobler 1994). The challenge is to incorporate scalable versions of these tools into GKD. Cartographic transformations can serve a similar role in GKD as data reduction and projection in KDD, i.e., determining effective representations that maximize the likelihood of discovering interesting geographic patterns in a reasonable amount of time.

Spatial dependency and heterogeneity. Measured geographic attributes usually exhibit the properties of *spatial dependency* and *spatial heterogeneity*. Spatial dependency is the tendency of attributes at some locations in space to be related[1]. These locations are usually proximal in Euclidean space. However, direction, connectivity and other geographic attributes (e.g., terrain, land cover) can also affect spatial dependency (see Miller 2000; Rosenberg 2000). Spatial dependency is similar to but more complex than dependency in other domains (e.g., serial autocorrelation in time series data).

Spatial heterogeneity refers to the non-stationarity of most geographic processes. An intrinsic degree of uniqueness at all geographic locations means that most geographic processes vary by location. Consequently, global parameters estimated from a geographic database do not

[1] In spatial analysis, this meaning of spatial dependency is more restrictive than it's meaning in the GKD literature. Spatial dependency in GKD is a rule that has a spatial predicate in either the precedent or antecedent. We will use the term "spatial dependency" for both cases with the exact meaning apparent

describe well the geographic phenomenon at any particular location. This is often manifested as apparent parameter drift across space when the model is re-estimated for different geographic subsets.

Spatial dependency and spatial heterogeneity have historically been regarded as nuisances confounding standard statistical techniques that typically require independence and stationarity assumptions. However, these can also be valuable sources of information about the geographic phenomena under investigation. Increasing availability of digital cartographic structures and geoprocessing capabilities has led to many recent breakthroughs in measuring and capturing these properties (see Fotheringham and Rogerson 1993).

Traditional methods for measuring spatial dependency include tests such as Moran's *I* or Geary's *C*. The recognition that spatial dependency is also subject to spatial heterogeneity effects has led to the development of *local indicators of spatial analysis* (LISA) statistics that disaggregate spatial dependency measures by location. Examples include the Getis and Ord *G* statistic and local versions of the *I* and *C* statistics (see Anselin 1995; Getis and Ord 1992, 1996).

One of the problems in measuring spatial dependency in very large datasets is the computational complexity of spatial dependency measures and tests. In the worse case, spatial autocorrelation statistics are approximately $O(n^2)$, since $n(n-1)$ calculations are required to measure spatial dependency in a database with *n* items (although in practice we can often limit the measurement to local spatial regions). Scalable analytical methods are emerging for estimating and incorporating these dependency structures into spatial models: Pace and Zou (2000) report an $O(n\log(n))$ procedure for calculating a closed form maximum likelihood estimator of nearest neighbor spatial dependency. Another, complementary strategy is to exploit parallel computing architectures. Fortunately, many spatial analytic techniques can be decomposed into parallel computations either due to task parallelism in the calculations or

---

from the context. This should not be too confusing since the GKD concept is a generalization of the

parallelism in the spatial data (see Ding and Densham 1996; Densham and Armstrong 1998; Griffith 1990). Armstrong and Marciano (1995) and Armstrong, Pavlik and Marciano (1994) report promising results with parallel implementations of the Getis-Ord *G* statistic.

Spatial analysts have recognized for quite some time that the regression model is misspecified and parameter estimates are biased if spatial dependency effects are not captured. Methods are available for capturing these effects in the structural components, error terms or both (see Anselin 1993; Bivand 1984). Regression parameter drift across space has also been long recognized. *Geographically weighted regression* uses location-based kernel density estimation to estimate location-specific regression parameters (see Brunsdon, Fotheringham and Charlton 1996; Fotheringham, Charlton and Brunsdon 1997).

The complexity of spatio-temporal objects and rules. Spatio-temporal objects and relationships tend to be more complex than the objects and relationships in non-geographic databases. Data objects in non-geographic databases can be meaningfully represented as points in information space. Size, shape and boundary properties of geographic objects often affect geographic processes, sometimes due to measurement artifacts (e.g., recording flow only when it crosses some geographic boundary). Relationships such as distance, direction and connectivity are more complex with dimensional objects (see Egenhofer and Herring 1994; Okabe and Miller 1996; Peuquet and Ci-Xiang 1987). Transformations among these objects over time are complex but information-bearing (Hornsby and Egenhofer 2000). Developing scalable tools for extracting spatio-temporal rules from collections of diverse geographic objects over time is a major GKD challenge.

In Chapter 2, Roddick and Lees discuss the types and properties of spatio-temporal rules that can describe geographic phenomena. In addition to spatio-temporal analogs of generalization, association and segmentation rules, there are evolutionary rules that describe

---

concept in spatial analysis.

changes in spatial entities over time. They also note that the scales and granularities for measuring time in geography can be complex, reducing the effectiveness of simply "dimensioning up" geographic space to include time. Roddick and Lees suggest that geographic phenomena are so complex that GKD may require *meta-mining*, that is, mining large rulesets that have been mined from data to seek more understandable information.

Diverse data types. The range of digital geographic data also presents unique challenges. One aspect of the digital geographic information revolution is that geographic databases are moving beyond the well-structured vector and raster formats. Digital geographic databases and repositories increasingly contain ill-structured data such as imagery and geo-referenced multimedia (see Câmara and Raper 1999). Discovering geographic knowledge from geo-referenced multimedia data is a more complex sibling to the problem of knowledge discovery from multimedia databases and repositories (see Zaïane et al. 1998).

3.1.2. Geographic knowledge discovery in geographic information science

There are unique needs and challenges for building geographic knowledge discovery into geographic information science. Most GIS databases are "dumb": they are at best a very simple representation of geographic knowledge at the level of geometric, topological and measurement constraints. Knowledge-based GIS is an attempt to capture high-level geographic knowledge by storing basic geographic facts and geographic rules for deducing conclusions from these facts (see, e.g., Srinivasan and Richards 1993; Yuan 1997). GKD is a potentially rich source of geographic facts and rules. A research challenge is building discovered geographic knowledge into geographic databases and models to support intelligent spatial analysis and additional knowledge discovery. This is critical; otherwise, the geographic knowledge obtained from the GKD process may be lost to the broader scientific and problem-solving processes.

3.1.3. Geographic knowledge discovery in geographic research

Geographic information has always been the central commodity of geographic research. Throughout its 3000-year history, the field of geography has operated in a data-poor environment. Geographic information was difficult to capture, store and integrate. Most revolutions in geographic research have been fueled by a technological advancement for geographic data capture, referencing and handling, including sailing ships, satellites, clocks, the global positioning system, the map and GIS. The current explosion of digital geographic and geo-referenced data is the most dramatic shift in the information environment for geographic research since the Age of Discovery in the fifteen and sixteenth centuries, perhaps in history.

Despite the promises of GKD in geographic research, there are some cautions. In Chapter 2, Roddick and Lees note that KDD and data mining tools were mostly developed for applications such as marketing where the standard of knowledge is "what works" rather than "what is authoritative." The question is how to use GKD as part of a defensible and replicable scientific process. As discussed previously in this chapter, knowledge discovery fits most naturally into the initial stages of hypothesis formulation. Roddick and Lees also suggest a strategy where data mining is used as a tool for gathering evidences that strengthen or refute the null hypotheses consistent with a conceptual model. These null hypotheses are a type of focusing technique that constrain the search space in the GKD process. The results will be more acceptable to the scientific community since the likelihood of accepting spurious patterns is reduced.

3.2. Geographic data warehousing

The data warehousing literature contains surprisingly little on the unique challenges associated with geographic data warehousing. Both academic and trade books on data warehousing mention geographic data only in passing, treating location as just another attribute of the data object.

Geographic data warehousing (GDW) shares most of the (considerable) challenges and design issues in standard data warehousing and introduces unique problems to DW design.

In Chapter 3, Bedard, Merrett and Han provide an overview of general DW design issues as well as issues specific to geographic data. As Bedard, Merrett and Han state, "A DW is an enterprise-oriented, integrated, non-volatile read-only collection of data imported from heterogeneous sources at several levels of detail to support decision-making." All of the terms in this definition have non-trivial design implications. The authors discuss the multidimensional DW design philosophy. They also review several system architectures for a DW, including traditional centralized, multi-tiered and *data mart* (mini-warehouses) architectures.

Geographic data introduces complexities that must be accommodated in the DW design and during the data integration process. First is the sheer size: GDW are potentially much larger than comparable non-geographic DWs. Consequently, there are stricter requirements for scalability. Multidimensional GDW design is more difficult since the spatial dimension can be measured using non-geometric, non-geometric generalized from geometric and fully geometric scales. Some of the geographic data can be ill-structured, for example remotely-sensed imagery and other graphics. OLAP tools such as roll-up and drill-down require aggregation of spatial objects and summarizing spatial properties. Spatial data interoperability is critical and particularly challenging since geographic data definitions in legacy databases can vary widely. Metadata management is more complex, particularly with respect to aggregated and fused spatial objects.

A *spatial data cube* is the GDW analog to the data cube tool for computing and storing all possible aggregations of some measure in OLAP. The spatial data cube must include standard attribute summaries as well as pointers to spatial objects at varying levels of aggregation. Aggregating spatial objects is non-trivial and often requires background domain knowledge in the form of a geographic concept hierarchy. Strategies for selectively pre-computing measures in the

spatial data cube include none, pre-computing rough approximations (e.g., based on minimum bounding rectangles), and selective pre-computation (see Han, Stefanovic and Koperski 1998).

In Chapter 4, Shekhar, Lu, Tan, Chawla and Vatsavai introduce the *map cube*. The map cube adds cartographic visualization to the spatial data cube. The map cube operator takes as arguments a base map, associated data files, a geographic aggregation hierarchy and a set of cartographic preferences. The operator generates an album of maps corresponding to the power set of all possible spatial and non-spatial aggregations. The map collection can be browsed using OLAP tools such as roll-up, drill-down and pivot using the geographic aggregation hierarchy.

Distributed geolibraries are becoming increasing prevalent and important as a source of geographically referenced data (National Research Council 1999). Sengupta and Bennett point out in Chapter 5 that while the volume of web-accessible geographically referenced data is growing rapidly, there are some substantial barriers to the effective use of these resources. A diverse set of national, state, and local agencies are developing and posting geographic datasets in a variety of formats, projections, coordinate systems and for different geographic extents. It is often necessary to perform a complicated sequence of data transformations to create a consistent and usable geographical dataset from these and other sources. Unfortunately, many users lack the technical knowledge to perform these transformations.

Sengupta and Bennett discuss the use of intelligent agent and blackboard technologies to resolve these difficulties. The system is designed for a distributed computer network. Individual agents contain the knowledge needed to perform a specific data transformation. Blackboard technologies find and organize a sequenced set of agents capable of performing all needed transformations to integrate diverse geographic data into a usable database. Their chapter discusses the knowledge structures developed to store spatial data processing knowledge, the algorithms used to develop transformation plans and provides an overview of a sample project.

3.3. Geographic data mining

3.3.1. Capturing spatial dependency effects

Geographic data mining involves the application of computational tools to reveal interesting patterns in objects and events distributed in geographic space and across time. These patterns may involve the spatial properties of individual objects and events (e.g., shape, extent) and spatio-temporal relationships among objects and events in addition to the non-spatial attributes of interest in traditional data mining.

In Chapter 6, Chawla, Shekhar, Wu and Ozesmi discuss the effects of spatial dependency in geographic data mining techniques. They note that spatial proximity and dependency patterns have led to major historical breakthroughs in understanding geographic processes and solving problems. These include hints of plate tectonics from the curious fact that the continents could be re-arranged to fit together and the discovery of the transmission mechanism for cholera in the 1800's from the unusual spatial clustering of incidences around a well in London (also see Dobson 1992). Despite these remarkable historical precedents, many data mining techniques ignore or greatly limit their search for spatial dependency patterns among attributes. Traditional data mining techniques search only for explicitly defined relationships among data objects and assume that no dependency effects are present in any relationship not explicitly examined. However, as the spatial dependency literature has shown, this can result in patterns that are biased and do not fit the data well. Chawla et al. demonstrate the effects of including spatial dependency into regression models and clustering techniques.

Difficulties in accounting for spatial dependency in geographic data mining include identifying the spatial dependency structure, the potential combinatorial explosion in the size of these structures and scale-dependency of many dependency measures. Further research is required along all of these frontiers. As noted above, researchers report promising results with parallel implementations of the Getis-Ord *G* statistic. Continued work on parallel implementations of spatial analytical techniques and spatial data mining tools can complement recent work on parallel processing in standard data mining (see Zaki and Ho 2000).

Spatial dependency can also manifest itself across spatial relationships other than Euclidean distance. Non-Euclidean distances, topological, directional relationships or some combination may be more appropriate for some geographic processes. In Chapter 7, Ester, Kriegel and Sander discuss efficient methods for capturing complex neighborhood relationships in spatial data mining. The authors argue that neighborhood effects are the major difference between mining in relational databases and mining in geographic databases. They present algorithms for major geographic data mining tasks and discuss typical applications for each algorithm. Their discussion highlights the requirements for efficient processing of neighborhood relations. Ester, Kriegel and Sander introduce general concepts for neighborhood relations and efficient computational strategies for implementing these concepts in geographic data mining. Their strategy allows a tight and efficient integration of geographic data mining algorithms with geographic database management systems, speeding up the development and the execution of the mining algorithms.

3.3.2. Geographic data mining techniques

Many of the traditional data mining tasks discussed previously have analogous tasks in the geographic data mining domain. See Ester, Kriegel and Sander (1997) and Han and Kamber (2000) for overviews. Also see Roddick and Spiliopoulou (1999) for a useful bibliography of spatio-temporal data mining research. The volume of geographic data combined with the complexity of spatial data access and spatial analytical operations implies that scalability is particularly critical.

Spatial segmentation tasks include spatial clustering and spatial classification. *Spatial clustering* groups spatial objects such that objects in the same group are similar and objects in different groups are unlike each other. This generates a small set of implicit classes that describe the data. Clustering can be based on combinations of non-spatial attributes, spatial attributes (e.g., shape) and proximity of the objects or events in space, time and space-time. Spatial

clustering has been a very active research area in both the spatial analytic and computer science literatures. Research on the spatial analytic side has focused on theoretical conditions for appropriate clustering in space-time (see O'Kelly 1994; Murray and Estivill-Castro 1998). Research on the computer science side has resulted in several scalable algorithms for clustering very large spatial datasets and methods for finding proximity relationships between clusters and spatial features (Knorr and Ng 1996; Ng and Han 1994).

In Chapter 8, Han, Tung and Kamber present an overview of major spatial clustering methods recently developed in the data mining literature. They classify spatial clustering methods into five categories, namely, partitioning, hierarchical, density-based, grid-based and model-based methods. Although traditional *partitioning metho*ds such as k-means and k-medoids are not scalable, scalable versions of these tools are available (also see Ng and Han 1994). *Hierarchical methods* group objects into a tree-like structure, that progressively reduces the search space. Hierarchical methods can build clusters from the bottom-up (by aggregation) or from the top-down (by splitting). Some methods combine hierarchical clustering and iterative relocation to improve their solutions. *Density-based methods* can find arbitrarily-shaped clusters by growing from a seed as long as the density in its neighborhood exceeds certain threshold. *Grid-based methods* divide the information spaces into a finite number of grid cells and cluster objects based on this structure. Finally, *model-based methods* first develop hypotheses for clusters and then find the best fit of the data to that model.

Spatial classification selects a relevant set of attributes and attribute values that determine an effective mapping of spatial objects into predefined target classes. Ester, Kriegel and Sander (1997) present a learning algorithm based on ID3 for generating spatial classification rules based on the properties of each spatial object as well as spatial dependency with its neighbors. The user provides a maximum spatial search length for examining spatial dependency relations with each object's neighbors. Adding a rule to the tree requires meeting a minimum information gain threshold.

Mining for *spatial dependency* involves finding rules to predict the value of some attribute based on the value of other attributes, where one or more of the attributes are spatial properties. *Spatial association rules* are association rules that include spatial predicates in the precedent or antecedent. Spatial association rules also have confidence and support measures. Spatial association rules can include a variety of spatial predicates, including topological relations such as "inside" and "disjoint," as well as distance and directional relations. Koperski and Han (1995) provide a detailed discussion of the properties of spatial association rules. They also present a top-down search technique that starts at the highest level of a geographic concept hierarchy (discussed below), using spatial approximations (such as minimum bounding rectangles) to discover rules with large support and confidence. These rules form the basis for additional search at lower levels of the geographic concept hierarchy with more detailed (and computationally-intensive) spatial representations.

Spatial objects that do not belong to any cluster are called outliers; distance measures can be used to identify outliers and analyze their properties. Han, Tung and Kamber also discuss distance-based outlier analysis and its use in outlier detection. As mentioned previously in this chapter, clustering and outlier detection are inverse problems. Outlier analysis can also be used in monitoring, embedded in a real-time system looking for unusual cases with respect to spatial and temporal trends. Ng discusses outlier analysis in Chapter 9, highlighting its use in a real-time monitoring system. Ng first reviews different classes of outlier detection, including *noise-based*, *distribution-based*, *depth-based* and *distance-based*. Ng also discusses the use of distance-based outlier detection in a video surveillance system monitoring a small geographic space. Video images of people walking through the space are translated to spatio-temporal trajectories. A distance-based outlier method determines cases that exhibit unusual trajectories, indicating unusual movement patterns through the geographic space. Although this type of system can be potentially abused, it also has great potential benefits for detecting thieves, terrorists or rapists in sensitive locations. The distance-based outlier analysis applied to space-time trajectories could

also be used to spot unusual and interesting movement behavior at other geographic scales and for other movement phenomena. This can include vehicular movement through intelligent transportation systems to spot congestion build-up or other unusual transportation events. Another possibility is tracking wildlife such as bears, wolves or hawks with embedded global positioning system receivers to spot unexpected changes in ecological niches.

Spatial trend detection involves finding patterns of change with respect to the neighborhood of some spatial object. Ester, Kriegel and Sander (1994) provide a neighborhood search algorithm for discovering spatial trends. The procedure performs a breadth-first search along defined neighborhood connectivity paths and evaluates a statistical model at each step. If the estimated trend is strong enough, that neighborhood path is expanded in the next step.

Geographic phenomena often have complex hierarchical dependencies. Examples include city systems, watersheds, location and travel choices, administrative regions and transportation/telecommunications systems. *Geographic characterization and generalization* is therefore an important geographic data mining task. Generalization-based data mining can follow one of two strategies in the geographic case. *Geographic dominant generalization* first spatially aggregates the data based on a user-provided geographic concept hierarchy. A standard attribute-oriented induction method is used at each geographic aggregation level to determine compact descriptions or patterns of each region. The result is a description of the pre-existing regions in the hierarchy using high-level predicates. *Non-geographic dominant generalization* generates aggregated spatial units that share the same high-level description. Attribute-oriented induction is used to aggregate non-spatial attributes into higher-level concepts. At each level in the resulting concept hierarchy, neighboring geographic units are merged if they share the same high-level description. The result is a geographic aggregation hierarchy based on multidimensional information. The extracted aggregation hierarchy for a particular geographic setting could be used to guide the application of confirmatory spatial analytic techniques to the data about that area.

3.4. Geographic knowledge discovery and geographic visualization

Earlier in this chapter, we noted the potential for using visualization techniques to integrate human visual pattern acuity and knowledge into the KDD process. Geographic visualization (GVis) is the integration of cartography, GIS and scientific visualization to explore geographic data and communicate geographic information to private or public audiences (see MacEachren and Kraak 1997). Major GVis tasks include *feature identification*, *feature comparison* and *feature interpretation* (MacEachren et al. 1999).

GVis is related to GKD since it often involves an iterative, customized process driven by human knowledge. However, the two techniques can greatly complement each other. For example, feature identification tools can allow the user to spot the emergence of spatio-temporal patterns at different levels of spatial aggregation and explore boundaries between spatial classes. Feature identification and comparison GVIs tools can also guide spatial query formulation. Feature interpretation can help the user build geographic domain knowledge into the construction of geographic concept hierarchies. MacEachren et al. (1999) discuss these functional objects and a prototype GVis/GKD software system that achieves many of these goals.

MacEachren et al. (1999) suggest that integration between GVis and GKD should be considered at three levels. The conceptual level requires specification of the high-level goals for the GKD process. Operational-level decisions include specification of appropriate geographic data mining tasks for achieving the high-level goals. Implementation level choices include specific tools and algorithms to meet the operational-level tasks.

In Chapter 10, Wachowicz presents the *GeoInsight* approach that builds on this conceptual model. At the conceptual level, GeoInsight helps define goals of the GKD process by considering what type of geographic knowledge is to be constructed and how knowledge can be constructed given the general nature of the geographic data. At the operational level, a *method-task-operation* strategy dictates that the data mining task to be applied consider the potential

forms of visual representation and user interaction modes. At the implementation level, GeoInsight approach dictates seamless integration of GKD and GVis tools to provide an effective and easy-to-use ESA environment. Wachowicz supports this strategy by first discussing the shared concepts and objectives as well as the integration challenges among GVis, GKD and ESA

In Chapter 11, Gahegan argues that portraying geographic data in a form that a human can understand frees ESA from some of the representational constraints that GIS and geographic data models impose. The chapter describes techniques that are currently available for exploratory visual analysis and visual data mining, highlighting the reasoning and inference implicit in each. Gahegan categorizes techniques according to the roles required from the system and user and the primary mode of reasoning (*deduction*, *induction* and *abduction* or inference to the best explanation). Following this review, Gahegan provides a formal description of mapping data from the database to a visual space with respect to how to control what is observed within the data. Gahegan briefly describes visual geographic knowledge construction and provides some general conclusions and future directions for continued research on the interface between GVis and GKD in exploratory spatial analysis.

3.5. Applications of geographic data mining and knowledge discovery

Although there is a need for more widespread "test cases" or benchmarks for GKD (see below), there have been some successful applications of these techniques for both human and physical geographic phenomena. In this subsection, we discuss some example applications in the existing literature as well as in this volume.

Map interpretation and information extraction. Map interpretation is a basic but daunting task that challenges even human-level intelligence. A particularly complex map product is a *topographic map*. Roughly, a topographic map is a large-scale (1:10,000 to 1:100,000) composite map usually depicting relief (terrain), land cover, hydrography and human-made

features. Although many GIS software can handle topographic map data, the ability to extract geographic information from these products is limited. This requires the ability to recognize natural versus human environmental features, interpret the spatial interrelationships among these features and understand territorial organization of political and administrative units. Although humans can often recognize features such as a road and distinguish this from other linear features such as hydrography, formal definitions suitable for automated feature extraction and interpretation are difficult to state. For example, consider the difficulty of stating a consistent and complete definition of the feature "road" that would be suitable for automated map interpretation.

In Chapter 12, Esposito, Lanza, Lisi and Malebra present a GIS for extracting information from topographic maps, INGENS (Inductive Geographic Information System). INGENS integrates traditional GIS data acquisition, storage, editing and graphical display capabilities with an inductive learning function. The learning subsystem discovers first-order summary rules implicit in the topographic map. The user provides primitive background knowledge (basic geographic constraints and properties, such as "two regions that share a geographic boundary are geographically proximal") and a set of geographic concepts to be learned (e.g., "downtown"). A learning algorithm follows a parallel search strategy, building a (possibly recursive) logical theory of the concepts based on observations in the database and the provided background knowledge. The authors provide a detailed description of the formal learning problem implemented within INGENS and present an example application from topographic maps in a region of Italy.

Information extraction from remotely-sensed imagery. Increasingly detailed information about the Earth's surface and its natural and human features is available through remote sensing (RS) systems. Traditionally, the spatial resolution of passive (reflectance-based) remote sensors has been limited, generally not better than 10 meters (m). New high-resolution passive sensing systems (e.g., IKONOS) can achieve spatial resolution down to 1 m, greatly increasing their

potential for socio-economic applications. Active sensing systems such as synthetic aperture radar (SAR) can presently achieve spatial resolutions down to the sub-meter level (see Lillesand and Kiefer 2000). Laser-based LIDAR systems under development can also achieve sub-meter accuracy. In addition to greater spatial resolution, RS systems are also improving with respect to spectral resolution. New hyperspectral sensor systems such as the Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) capture over two hundred electromagnetic bands, generating very detailed spectral signature for each pixel. These technological breakthroughs are creating new domains for RS, particularly for socio-economic applications since detailed human features can now be obtained. However, a challenge is dealing with the large amount of data and information in these imagery files.

Artificial neural networks (ANNs) have been used effectively for classification and information extraction from remotely-sensed imagery for several decades. However, a problem with ANNs is their "black box" nature: while ANNs are very good at classification and generalization, their internal workings are difficult to interpret. This means that much of the information available in these images is lost. While this may be acceptable for pure classification and prediction problems, the ability to interpret the parameters of a trained ANN can provide better understanding of the underlying data relationships and properties as well as the ability to generalize results across different geographic extents, time periods and RS systems.

In Chapter 13, Gopal, Liu and Woodcock explore the use of visualization in interpreting the inner workings of a trained ANN. Gopal, Liu and Woodcock work with a family of ANNs known as fuzzy Adaptive Resonance Theory (ART) networks. An advantage over standard backpropagation ANNs is that ART networks can incorporate new information without substantially disrupting the previously learned parameters. Fuzzy ART networks can recognize "non-crisp" patterns; this capability is often critical in geographic applications where features are often fuzzy. The authors train a fuzzy ART network to recognize a conifer forest (at least 40% needle leaf evergreen forest). Visualization guides several aspects of this process, including

feature selection, analyzing the source and nature of misclassification errors, training data selection, understanding generalization problems and pruning the ANN to obtain a more parsimonious representation.

Mapping environment features. Many geographic phenomena have complex, multidimensional attributes that are difficult to summarize and combine using traditional analytical methods. For example, phenomena such as soil type are very difficult to classify and map. Soil type is a function of several geographic variables, including the soil composition, soil depth, underlying geology, slope, elevation and hydrology, all of which combine in a complex manner to form particular soil regimes. This makes it very difficult to develop a precise formula or rule set that can unambiguously and accurately classify soil types. Eklund, Kirkby and Salim (1998) apply data mining techniques to classify soil types. They compare several inductive learning algorithms (including C4.5) and feedforward/backpropagation ANNs with ground truth data in classifying soils based on salinity. The data mining techniques are embedded within a knowledge-based spatial decision support system for environmental monitoring and planning. The data mining tools allow refinement of the knowledge base within the decision support system as new data are acquired.

Lees and Ritman (1991) apply data mining techniques to the difficult problem of mapping vegetation types and structure in hilly and disturbed areas. While remotely-sensed imagery can often be useful in vegetation mapping, these images are not very useful in determining detailed speciation, particularly in hilly areas and disturbed areas (e.g., due to fire or human action). Ancillary data can be helpful, but they can be difficult integrate with imagery within a traditional modeling framework. Lees and Ritman (1991) use decision tree induction methods to integrate imagery with other environmental variables in a GIS to classify vegetation types. Their method resolves topographic shadowing problems in the imagery due to terrain and can also identify disturbed, cleared and agricultural areas.

Extracting spatio-temporal patterns. Mesrobian et al. (1996) describe the OASIS (Open Architecture Scientific Information System) information system for querying, exploratory analysis and visualization of geophysical phenomena from large, heterogeneous, distributed database systems. OASIS facilitates collaborative exploratory data analysis, knowledge discovery and visualization among scientists at diverse sites. A component of OASIS is the Conquest Scientific Query Processing System. This is a data mining system that locates cyclonic storms and their trajectories from weather and climate data. The system extracts air pressure minima from the data, refines the location of the minima, and assigns minima to storm tracks (also see Mesrobian et al. 1994). Clustering and normalization tools help spot unusual atmospheric patterns corresponding to cyclonic activity.

Openshaw (1994) describes two computational techniques for exploring space-time-attribute patterns in digital geographic data. The first technique has its origins in the Geographic Analysis Machine (GAM) developed by Openshaw et al. (1987). During each iteration, the GAM searches within a small, pre-specified radius each of geographic object, looking for neighboring observations with similar values. If a spatial dependency is found, the map displays that circle. After all observations have been scanned, a new iteration begins with a slightly larger circle based on a user-specified step-size. This continues up to a user-specified limit. The resulting map shows possible spatial clusters. Later versions of the GAM include spatial cluster identification rules. Openshaw (1994) develops an extension of this brute-force approach to include temporal and attribute search ranges. He also describes the use of artificial life that explore databases feeding from spatio-temporal correlations. Openshaw (1994) illustrates the use of these tools in a geo-referenced crime database.

Interaction, flow and movement in geographic space. Geographers refer to the movement of people, materials, capital and information between geographic locations collectively as *spatial*

interaction. Spatio-temporal patterns in spatial interaction can illuminate the hidden and subtle geographic structures that generate these patterns. Examples include characterizing the internal and external structure of regional economies, illuminating the relationships between migration, demographics and socioeconomic characteristics at detailed levels of spatio-temporal resolution and recognizing the predicates of congestion in transportation network. Traditional spatial analytic techniques also have difficulty in capturing *n*-order cascade effects among interaction and flows (e.g., flow from A to B creating second-order flows from B to C, B to D, etc.).

Spatial interaction data are often recorded as flows between origin and destinations or as flows within structures such as riparian systems and transportation networks. Analyzing interaction and flow data is a long-standing concern in geography. However, existing techniques are easily overwhelmed by the huge sizes of the origin-destination matrices being published by entities such as the United States Bureau of Transportation Statistics, real-time data being collected through intelligent transportation systems and operational data being collected by utility companies. Visualization and exploratory tools could discover the hidden knowledge within these rich flow matrices.

Marble, Gou, Liu and Saunders (1997) review traditional approaches to analyzing interaction and flow data and describe emerging research on the design and proof–of–concept testing of scientific visualization and dynamic graphics–based tools for the exploratory analysis of spatial–temporal, interregional flow systems. This research is attempting to improve the computer-based flow mapping techniques developed by analytical cartographers (e.g., Tobler 1981). Of particular interest are visualization tools allowing the user to examine the total set of flows and identify interesting subsets within the total flow set. Other exploratory tools include forward and backward brushing techniques. These tools can quickly and selectively explore the relationships that exist between various components of the flow data set (e.g., inflows vs. outflows) and between the various interregional flows and selected characteristics of the origin and destination regions (e.g., out–migration as related to unemployment in the region of origin).

A third set of exploratory techniques involve the projection pursuit tools for reducing the dimensionality of multi–dimensional flow matrices to allow visualization within geographic and other spaces.

Geographers and transportation analysts have long recognized that aggregate flows between locations or within networks is an incomplete picture of spatial interaction phenomena. Individuals often exhibit complex *trip chaining* behavior where multiple trip purposes and multiple stops are combined into a single travel episode for efficiency or due to time constraints. Data on complex trip chaining behavior can be acquired through activity diaries, a collection method that is becoming increasingly useful and accurate due to developments in data collection devices such as global positioning systems (GPS) combined with handheld recording devices such as personal digital assistants (PDA) (see Murakami and Wagner 1999). Arentze et al. (2000) apply the rule induction algorithms C4.5, CART and CHAID to derive decision trees for classifying different types of multi-stop/multi-purpose travel episodes based on a large activity database. Results suggest that decision tree induction methods can capture both compensatory and non-compensatory behavioral rules as well as interactions among different travel choices. Forer (1998) and Kwan (2000) develop 3-D geographic visualizations of these paths to support exploratory analysis of activity data and spatio-temporal constraints on travel behavior.

The increasing number of position-aware, wirelessly connected mobile devices, including cell phones, personal information managers (PIMs) and personal digital assistants (PDAs), are a potentially rich source of information on human movement within geographic space. Technology for determining the detailed geographic location is becoming very inexpensive. Also, the social and commercial forces driving deployment of positioning capability are very strong (e.g., E-911 services, intelligent transportation systems, mobile geographic information services). Consequently, there are growing possibilities to access space-time trajectories of these personal devices and their human companions in real-time and to deliver them to a data warehouse for analysis. These mobile trajectories contain detailed information about personal and vehicular

mobile behavior. They present interesting opportunities to find patterns, extract their content, and use this discovered knowledge to enhance people's ability to travel efficiently, effectively, comfortably and with a minimum of stress.

In Chapter 14, Smyth explores the possibilities for geographic knowledge discovery from the spatio-temporal trajectories of mobile devices. Smyth reviews general characteristics of the "mobile world" (the perspective of a person traveling though geographic space), position-aware wireless devices and the type of mobile geographic information services that could be provided through these and other mobile devices. Smyth also identifies how recorded mobile trajectories reflect human spatial behavior and requirements for data mining from these trajectories. Smyth discusses how the discovered knowledge can lead to better, more scalable, and less expensive mobile geographic information services. One possibility is resolving the high cost of mobile geographic information services. It may be possible to build self-maintaining mobile geographic information systems that can provide valuable assistance to an individual based only on past behavior and some knowledge of current conditions.

4. GKD RESEARCH FRONTIERS

There are several critical research frontiers in GKD, presented as below (in no particular order) (also see Miller and Han 2000).

Developing and supporting geographic data warehouses. A glaring omission from current research in GKD techniques is the development and supporting infrastructure for *geographic data warehouses* (GDW). To date, a true GDW does not exist. This is alarming since data warehouses are central to the knowledge discovery process. Creating true GDWs requires solving issues in geographic and temporal data compatibility, including differences in semantics,

referencing systems, geometry, accuracy and precision. Supporting GDWs may also require restructuring of transaction-oriented databases systems, particularly for flow and interaction data.

Better spatio-temporal representations in GKD. Current GKD techniques use very simple representations of geographic objects and geographic relationships, for example, point objects and Euclidean distances. Other geographic objects (including lines, polygons and more complex objects) and geographic relationships (including non-Euclidean distances, direction, connectivity, attributed geographic space such as terrain and constrained interaction structures such as networks) should be recognized by GKD techniques. Time needs to be more completely integrated into geographic representations and relationships. This includes a full range of conceptual, logical and physical models of spatio-temporal objects. Finally, we need to formulate multiple representations (in particular, robust geographic concept hierarchies) and granularities in spatio-temporal representation in order to manage the complexity of GKD.

GKD using richer geographic data types. Geographic datasets are rapidly moving beyond the well-structured vector and raster formats to include semi-structured and unstructured data, in particular, geo-referenced multimedia data. GKD techniques should be developed that can handle these heterogeneous datasets.

User interfaces for GKD. GKD needs to move beyond the technically-oriented researchers to the broader geographic and related research communities. This requires interfaces and tools that can aid these diverse researchers in the GKD process. These interfaces and tools should be based on useful metaphors that can guide the search for geographic knowledge and make sense of discovered geographic knowledge.

What are the new questions for geographic research? A fundamental question for the geographic and related research communities is "What questions do we want to answer that we could not answer previously?" These communities need to form well-structured (but possibly open-ended) questions in order to guide the computer science and related communities in their tool and algorithm development.

Proof of concepts and benchmarking problems for GKD. There is a strong need for some "examples" or "test cases" to illustrate the usefulness of GKD. This includes a demonstration of GKD techniques leading to new, unexpected knowledge in key geographic research domains. Also important is benchmarking to determine the effects of varying data quality on discovering geographic knowledge. A related issue is research and demonstration projects that illustrate the usefulness of GKD techniques in forecasting and decision support for the public and private sectors.

Building discovered geographic knowledge into GIS and spatial analysis. Current GIS software uses simple representations of geographic knowledge. Discovered geographic knowledge should be integrated into GIS, possibly through inductive geographic databases or online analytical processing (OLAP)-based GIS interfaces. There is also a need for new, intelligent spatial analysis methods that can represent the high-level discovered through GKD.

5. CONCLUSIONS

Due to explosive growth and wide availability of geo-referenced data in recently years, traditional spatial analysis tools are far from adequate at handling the huge volumes of data and the growing complexity of spatial analysis tasks. Geographic data mining and knowledge discovery represents an important direction in the development of new generation of spatial analysis tools in

data-rich environment. In this chapter, we provide a brief introduction to knowledge discovery from databases and data mining, with special reference the applications of these theories and techniques to geo-referenced data.

As shown in this chapter, geographic knowledge discovery is an important and interesting special case of knowledge discovery from databases. Much progress has been made recently in geographic knowledge discovery techniques, including heterogeneous spatial data integration, spatial or map data cube construction, spatial dependency and/or association analysis, spatial clustering methods, spatial classification and spatial trend analysis, spatial generalization methods, and geographic visualization tools. Applications of geographic data mining and knowledge discovery have also been actively developed, including map interpretation and information extraction tools, information extraction from remotely-sensed imagery, identification of spatial-temporal patterns, mapping environment features, and analysis of flow and movements in geographic space. However, according to our view, geographic data mining and knowledge discovery is a promising but young discipline, facing many challenging research problems. We hope this book will introduce some recent works in this direction and motivate researchers to make contributions at developing new methods and applications in this promising field.

6. LITERATURE CITED

Adriaans P. and Zantinge, D. (1996) *Data Mining*; Harlow, U.K.: Addison-Wesley.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1996) "Fast discovery of association rules," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 307-328

Agrawal, R. and Srikant, R. (1995) "Mining sequential patterns," *Proceedings, 11th International Conference on Data Engineering*, Los Alamitos, CA: IEEE Computer Society Press, 3-14.

Anselin, L. (1993) "Discrete space autoregressive models," in M.F. Goodchild, B. O. Parks and L. T. Steyaert (eds.) *Environmental Modeling with GIS*, New York: Oxford University Press, 454-469

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

Arentze, T. A., Hofman, F., van Mourik, H., Timmermans, H. J. P. and Wets, G. (2000) "Using decision tree induction systems for modeling space-time behavior," *Geographical Analysis*, in press.

Armstrong, M. P. and Marciano, R. (1995) "Massively parallel processing of spatial statistics," *International Journal of Geographical Information Systems,* 9, 169-189.

Armstrong, M. P., Pavlik, C. E. and Marciano, R. (1994) "Experiments in the measurement of spatial association using a parallel supercomputer," *Geographical Systems*, 1, 267-288

Barbara, D., DuMouchel, W., Faloutos, C., Haas, P. J., Hellerstein, J. H., Ioannidis, Y., Jagadish, H. V., Johnson, T., Ng, R., Poosala, V., Ross, K. A. and Servcik K. C. (1997), "The New Jersey data reduction report," *Bulletin of the Technical Committee on Data Engineering*, 20(4): 3-45.

Beguin, H. and Thisse, J.-F. (1979) "An axiomatic approach to geographical space," *Geographical Analysis*, 11, 325-341.

Berndt, D. J. and Clifford, J. (1996) "Finding patterns in time series: A dynamic programming approach," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 229-248.

Bivand, R. S. (1984) "Regression modeling with spatial dependence: An application of some class selection and estimation techniques," *Geographical Analysis*, 16, 25-37

Brachman, R. J. and Anand, T. (1996) "The process of knowledge-discovery in databases: A human-centered approach," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press 37-57

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. and Sander, J. (1999) "OPTICS-OF: Identifying local outliers," in J. M. Żytkow and J. Rauch (eds.) *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence 1704, 262-270.

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

Buntine, W. (1996) "Graphical models for discovering knowledge," U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 59-82

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

Chaudhuri, S. and Dayal, U. (1997) “An overview of data warehousing and OLAP technology,” *SIGMOD Record*, 26, 65-74.

Cheesman, P. and Stutz, J. (1996) "Bayesian classification (AutoClass): Theory and results," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 153-180.

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.

Densham, P. J. and Armstrong, M. P. (1998) "Spatial analysis," in R. Healy, S. Dowers, B. Gittings and M. Mineter (eds.) *Parallel Processing Algorithms for GIS*, London: Taylor and Francis, 387-413.

Ding, Y. and Densham, P. J. (1996) "Spatial strategies for parallel spatial modeling," *International Journal of Geographical Information Systems,* 10, 669-698

Dobson, J. (1992) "Spatial logic in paleogeography and the explanation of continental drift," *Annals of the Association of American Geographers,* 82, 187-206.

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. and Pregibon, D. (1996) “A statistical perspective on knowledge discovery,” pp. 83-113 in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 83-113

Ester, M., Kriegel, H.-P. and Sander, J. (1997) "Spatial data mining: A database approach," M. Scholl and A, Voisard (eds.) *Advances in Spatial Databases*, Lecture Notes in Computer Science 1262, Berlin: Springer, 47-66.

Farnstrom, F., Lewis, J. and Elkan, C. (2000) "Scalability for clustering algorithms revisited," *SIGKDD Explorations*, 2, 51-57.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) "From data mining to knowledge discovery: An overview" in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 1-34.

Flexer, A. (1999) "On the use of self-organizing maps for clustering and visualization," in J. M. Żytkow and J. Rauch (eds.) *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence 1704, 80-88.

Forer, P. (1998) "Geometric approaches to the nexus of time, space and microprocess: Implementing a practical model for mundane soci-spatial systems," in M. J. Egenhofer and R. G. Golledge (eds.) *Spatial and Temporal Reasoning in Geographic Information Systems*, Oxford: Oxford University Press, 171-190.

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) "On the application of inductive machine learning tools to geographical analysis," *Geographical Analysis*, 32, 113-139

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

Gatrell, A. C. (1983) *Distance and Space: A Geographical Perspective*, Oxford: Clarendon Press.

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

Goebel, M. and Gruenwald, L. (1999) "A survey of data mining and knowledge discovery software tools," *SIGKDD Explorations*, 1, 20-33.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H. (1997) "Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals," *Data Mining and Knowledge Discovery*, 1, 29-53.

Griffith, D. A. (1990) "Supercomputing and spatial statistics: A reconnaissance," *Professional Geographer*, 42, 481-492.

Han, J., Cai, Y. and Cercone, N. (1993) "Data-driven discovery of quantitative rules in relational databases," *IEEE Transactions on Knowledge and Data Engineering*, 5, 29-40

Han, J. and Kamber, M. (2000), Data Mining: Concepts and Techniques, San Matel, CA: Morgan Kaufmann.

Han, J., Stefanovic, N. and Koperski, K. (1998) "Selective materialization: An efficient method for spatial data cube construction," in *Proceedings of the 1998 Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science 1394, Berlin: Springer-Verlag, 144-158.

Hand, D. J. (1998) "Data mining: Statistics and more?" *American Statistician,* 52, 112-118

Harinarayan, V., Rajaramna, A. and Ullman, J. D. (1996) "Implementing data cubes efficiently," *SIGMOD Record*, 25, 205-216.

Heckerman, D. (1997) "Bayesian networks for data mining," *Data Mining and Knowledge Discovery*, 1, 79-119.

Hipp, J., Güntzer, U. and Nakhaeizadeh, G. (2000) "Algorithms for association rule mining: A general survey and comparison," *SIGKDD Explorations*, 2, 58-64.

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.

Jarke, M., Lenzerini, M., Vassiliou, Y. and Vassiliadis, P. (2000) *Fundamentals of Data Warehouses*, Berlin: Springer.

Kass, G. V. (1980) "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, 29, 119-127.

Keim, D. A. and Kriegel, H.-P. (1994) "Using visualization to support data mining of large existing databases, " in J. P. Lee and G. G. Grinstein (eds.) *Database Issues for Data Visualization*, Lecture Notes in Computer Science 871, 210-229.

Klösgen, W. and Żytkow, J. M. (1996) "Knowledge discovery in databases terminology," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press 573-592.

Knorr, E. M. and Ng, R. T. (1996) "Finding aggregate proximity relationships and commonalties in spatial data mining," *IEEE Transactions on Knowledge and Data Engineering*, 8, 884-897

Koperski, K. and Han, J. (1995) “Discovery of spatial association rules in geographic information databases,” in M. Egenhofer and J. Herring (eds.) *Advances in Spatial Databases*, Lecture Notes in Computer Science Number 951, Springer-Verlag, 47-66.

Kwan, M.-P. (2000) "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set," *Transportation Research C*, in press.

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

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.

Lillesand, T. M. and Kiefer, R. W. (2000) *Remote Sensing and Image Interpretation*, 4ed. New York: John Wiley

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

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

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.

Matheus, C. J., Chan, P. K. and Piatetsky-Shapiro, G. (1993) "Systems for knowledge discovery in databases," *IEEE Transactions on Knowledge and Data Engineering*, 5, 903-913.

Mesrobian, E, Muntz, R., Santos, J. R., Shek, E., Mechoso, C. R., Farrara, J. D. and Stolorz, P. (1994) "Extracting spatio-temporal patterns from geoscience datasets, *IEEE Workshop on Visualization and Machine Visualization*, Los Alamitos, CA: IEEE Computer Society Press, 92-103.

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*, 1, 105-108

Murakami, E. and Wagner, D. P. (1999) "Can using global positioning system (GPS) improve trip reporting?" *Transportation Research C*, 7, 149-165

Murray, A. T. and Estivill-Castro, V. (1998) "Cluster discovery techniques for exploratory data analysis," *International Journal of Geographical Information Science*, 12, 431-443.

National Research Council (1999) *Distributed Geolibraries: Spatial Information Resources*, Wahshington, D.C.: National Academy Press.

Ng, R. T. and Han, J. (1994) "Efficient and effective clustering methods for spatial data mining," *Proceedings of the 20th Very Large Database Conference*, Santiago, Chile.

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.

O'Kelly, M. E. (1994) "Spatial analysis and GIS," in A. S. Fotheringham and P. A. Rogerson (eds.) *Spatial Analysis and GIS*, London: Taylor and Francis, 65-79

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.

Pace, R. K. and Zou, D. (2000) "Closed-form maximum likelihood estimates of nearest neighbor spatial dependence," *Geographical Analysis*, 32, 154-172

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

Quinlan, J. R. (1986) Induction of decision trees, *Machine Learning,* 1, 81-106.

Quinlan, J. R. (1993) *C4.5 Programs for Machine Learning*, San Matel, CA: Morgan Kaufmann.

Reinartz, T. (1999) *Focusing Solutions for Data Mining*, Lecture Notes in Artificial Intelligence 1623, Berlin: Springer.

Roddick, J. F. and Spiliopoulou, M. (1999) "A bibliography of temporal, spatial and spatio-temporal data mining research," *SIGKDD Explorations*, 1, 34-38.

Rosenberg, M. S. (2000) "The bearing correlogram: A new method of analyzing directional spatial autocorrelation," Geographical Analysis, 32, ???-???

Silberschatz, A., Korth, H.F. and Sudarshan, S. (1997) *Database Systems Concepts*, 3ed., McGraw Hill.

Silberschatz, A. and Tuzhilin, A. (1996), "What makes patterns interesting in knowledge discovery systems", *IEEE Transactions on Knowledge and Data Engineering*, 8, 970-974.

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.

Tobler, W. R. (1981) "A model of geographical movement," *Geographical Analysis*, 13, 1-20

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

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.

Zaki, M. J. and Ho, C.-T. (eds.) (2000) *Large-scale Parallel Data Mining*, Lecture Notes in Artificial Intelligence 1759, Berlin: Springer.