1 Geographic Data Mining and Knowledge Discovery
An Overview

Harvey J. Miller
Jiawei Han

CONTENTS

1.1 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 conditions, 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 location-ware technologies (LATs) such as global positioning system (GPS) receivers allow field researchers to collect unprecedented amounts of data. LATs linked to or embedded in devices such as cell phones, in-vehicle navigation systems, and wireless Internet clients provide location-specific content in exchange for tracking individuals 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 toward 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, USA. The workshop brought together a diverse group of stakeholders with interests in developing and applying computational techniques for exploring large, heterogeneous digital geographic datasets. Drawing on papers submitted to that workshop, in 2001 we published *Geographic Data Mining and Knowledge Discovery,* a volume that brought together some of the cutting-edge research in the area of geographic data mining and geographic knowledge discovery in a data-rich environment. There has been much progress in geographic knowledge discovery (GKD) over the past eight years, including the development of new techniques for geographic data warehousing (GDW), spatial data mining, and geo-visualization. In addition, there has been a remarkable rise in the collection and storage of data on spatiotemporal processes and mobile objects, with a consequential rise in knowledge discovery techniques for these data.

The second edition of *Geographic Data Mining and Knowledge Discovery* is a major revision of the first edition. We selected chapters from the first edition and asked authors for updated manuscripts that reflect changes and recent developments in their particular domains. We also solicited new chapters on topics that were not

covered well in the first edition but have become more prominent recently. This includes several new chapters on spatiotemporal and mobile objects databases, a topic only briefly mentioned in the 2001 edition.

This chapter introduces geographic data mining and GKD. In this chapter, we provide an overview of knowledge discovery from databases (KDD) and data mining. We identify why geographic data is a nontrivial special case that requires distinctive 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.

1.2 KNOWLEDGE DISCOVERY AND DATA MINING

In this section, we provide a general overview of knowledge discovery and data mining. We begin with an overview of 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 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.

1.2.1 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 domain 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 domain) 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 nonrandom 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 nontrivial and unexpected. *Useful* implies that the pattern should lead to some effective action, e.g., successful 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, largely due to the

type of data that increasingly comprise enterprise databases and 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, nonnumeric, 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 guideline 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.

1.2.2 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 DW 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, government, 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 et 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 database design principles very different from 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 multiuser 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 nonnormalized 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 types of "super-queries": more powerful than standard query language such as SQL but shallower than data-mining techniques because 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 zero dimension (e.g., "total sales"), one dimension (e.g., "total sales by item," "total sales by store," "total sales

per week"), two dimensions (e.g., "total sales by item and store") and so on, up to *N* dimensions. 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" because the logical arrangement of aggregations can be viewed as a hypercube in an *N*-dimensional information space (see Gray et al. 1997, Figure 2). 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 on measures computed by different aggregation operators, such as max, min, average, top-10, variance, and so on.

1.2.3 The KDD Process and Data Mining

The KDD process usually consists of several steps, namely, data selection, data pre-processing, data enrichment, data reduction and projection, data mining, and pattern interpretation and reporting. These steps may not necessarily be executed in linear order. Stages may be skipped or revisited. Ideally, KDD should be a human-centered 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; Han and Kamber 2006; 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 functions for revealing hidden information in a database (Klösgen and Żytkow 1996). The type of knowledge to be mined determines the data-mining function to be applied (Han and Kamber 2006). Table 1.1 provides a possible classification of data-mining tasks and techniques. See Matheus, Chan and Piatetsky-Shapiro (1993) and Fayyad, Piatetsky-Shapiro and

TABLE 1.1
Data-Mining Tasks and Techniques

Knowledge Type	Description	Techniques
Segmentation or clustering	Determining a finite set of implicit groups that describe the data.	Cluster analysis
Classification	Predict the class label that a set of data belongs to based on some training datasets	Bayesian classification Decision tree induction Artificial neural networks Support vector machine (SVM)
Association	Finding relationships among itemsets or association/correlation rules, or predict the value of some attribute based on the value of other attributes	Association rules Bayesian networks
Deviations	Finding data items that exhibit unusual deviations from expectations	Clustering and other data-mining methods Outlier detection Evolution analysis
Trends and regression analysis	Lines and curves summarizing the database, often over time	Regression Sequential pattern extraction
Generalizations	Compact descriptions of the data	Summary rules Attribute-oriented induction

Smyth (1996), as well as several of the chapters in this current volume for other overviews and classifications of data-mining techniques.

Segmentation or *clustering* involves partitioning a selected set of data into meaningful groupings or classes. It usually applies cluster analysis algorithms to examine the relationships between data items and determining a finite set of implicit classes so that the intraclass similarity is maximized and interclass similarity is minimized. 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). *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).

Classification refers to finding rules or methods to assign data items into preexisting classes. Many classification methods have been developed over many years of research in statistics, pattern recognition, machine learning, and data mining, including decision tree induction, naïve Bayesian classification, neural networks, support vector machines, and so on. *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 (Breiman et al. 1984). Artificial neural networks (ANNs) can be used as nonlinear 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 feed forward/back propagation architecture require supervised training to determine the appropriate weights (response function) to assign data items into known classes.

Associations are rules that predict the object relationships as well as 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).

Deviations are data items that exhibit unexpected deviations or differences from some norm. 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).

Trends are lines and curves fitted 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 violate 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 with the root representing the null descriptor ("any"). Concept hierarchies can be derived from experts or from data cardinality analysis (Han and Fu 1996).

A potential problem that can arise in a data-mining application is the large number of patterns generated. Typically, only a small proportion of these patterns will encapsulate interesting knowledge. The vast majority may be trivial or irrelevant. A data-mining engine should present only those patterns that are interesting to particular users. *Interestingness measures* are quantitative techniques that separate interesting patterns from trivial ones by assessing the simplicity, certainty, utility, and novelty of the generated patterns (Silberschatz and Tuzhilin 1996; Tan, Kumar and Srivastava 2002). There are many interestingness measures in the literature; see Han and Kamber (2006) for an overview.

1.2.4 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, 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 of humans can 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.

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. Fayyad, Grinstein and Wierse (2001) provide a good overview of visualization methods for data mining.

1.3 GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY

This section of the chapter describes a very important special case of KDD, namely, GKD. We will first discuss why GKD is an important special case that requires careful consideration and specialized tools. We will then discuss GDW 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 techniques and applications and research frontiers, highlighting the contributions of this current volume.

1.3.1 Why Geographic Knowledge Discovery?

1.3.1.1 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 GKD.

Geographic measurement frameworks. While many information domains of interest in KDD are high dimensional, these dimensions are relatively independent. Geographic information is not only high dimensional but also has the property that up to four dimensions 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; Miller and Wentz 2003).

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 nonmetric. In addition, disease patterns in space and time often behave according to topologies and geometries other than Euclidean (see Cliff and Haggett 1998; Miller and Wentz 2003). 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 Cliff and Haggett 1998; Gatrell 1983; Müller 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.* 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 and Wentz 2003; 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 nonstationarity 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 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

* In spatial analysis, this meaning of spatial dependency is more restrictive than its 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 from the context. This should not be too confusing since the GKD concept is a generalization of the concept in spatial analysis.

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 worst case, spatial autocorrelation statistics are approximately $O(n^2)$ in complexity, 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 and cyber-infrastructure. Fortunately, many spatial analytic techniques can be decomposed into parallel and distributed computations due to either task parallelism in the calculations or parallelism in the spatial data (see Armstrong and Marciano 1995; Armstrong, Pavlik and Marciano 1994; Densham and Armstrong 1998; Ding and Densham 1996; Griffith 1990; Guan, Zhang and Clarke 2006).

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 spatiotemporal objects and rules. Spatiotemporal objects and relationships tend to be more complex than the objects and relationships in nongeographic databases. Data objects in nongeographic 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 spatiotemporal rules from collections of diverse geographic objects is a major GKD challenge.

In their update of Chapter 2 from the first edition of this book, Roddick and Lees discuss the types and properties of spatiotemporal rules that can describe geographic phenomena. In addition to spatiotemporal analogs of generalization, association, and segmentation rules, there are evolutionary rules that describe 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 rule sets 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 Petrushin and Khan 2006).

1.3.1.2 Geographic Knowledge Discovery in Geographic Information Science

There are unique needs and challenges for building GKD into geographic information systems (GIS). 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 Srinivasan and Richards 1993; Yuan 1997). The semantic web and semantic geospatial web attempt to make information understandable to computers to support interoperability, findability, and usability (Bishr 2007; Egenhofer 2002).

GKD is a potentially rich source of geographic facts. A research challenge is building discovered geographic knowledge into geographic databases and models to support information retrieval, interoperability, 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.

1.3.1.3 Geographic Knowledge Discovery in Geographic Research

Geographic information has always been the central commodity of geographic research. Throughout much of its 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 the map, accurate clocks, satellites, GPS, and GIS. The current explosion of digital geographic and geo-referenced data is the most dramatic shift in the information environment for geographic research 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 kinds of focusing techniques 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.

1.3.2 Geographic Data Warehousing

A change since the publication of the first edition of this book in 2001 is the dramatic rise of the geographic information market, particular with respect to web-mapping services and mobile applications. This has generated a consequent heightened interest in GDW.

A GDW involves complexities that are unique to standard DWs. First is the sheer size. GDWs are potentially much larger than comparable nongeographic DWs. Consequently, there are stricter requirements for scalability. Multidimensional GDW design is more difficult because the spatial dimension can be measured using nongeometric, nongeometric 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 because geographic data definitions in legacy databases can vary widely. Metadata management is more complex, particularly with respect to aggregated and fused spatial objects.

In Chapter 3, also an update from the first edition, Bédard and Han provide an overview of fundamental concepts underlying DW and GDW. After discussing key concepts of nonspatial data warehousing, they review the particularities of GDW, which are typically spatiotemporal. They also identify frontiers in GDW research and development.

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 nontrivial 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 2000).

In Chapter 4, Lu, Boedihardjo, and Shekhar update a discussion of the *map cube* from the first edition. The map cube extends the data cube concept to GDW. 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 nonspatial aggregations. The map collection can be browsed using OLAP tools such as roll-up, drill-down, and pivot using the geographic aggregation hierarchy. They illustrate the map cube through an application to highway traffic data.

GDW incorporates data from multiple sources often collected at different times and using different techniques. An important concern is the quality or the reliability of the data used for GKD. While error and uncertainty in geographic information have been long-standing concerns in the GIScience community, efforts to address

these issues have increased substantially since the publication of the first edition of this book in 2001 (Goodchild 2004).

Chapter 5 by Gervais, Bédard, Levesque, Bernier, and DeVillers is a new contribution to the second edition that discusses data quality issues in GKD. The authors identify major concepts regarding quality and risk management with respect to GDW and spatial OLAP. They discuss possible management mechanisms to improve the prevention of inappropriate usages of data. Using this as a foundation, Chapter 5 presents a pragmatic approach of quality and risk management to be applied during the various stages of a spatial data cube design and development. This approach manages the potential risks one may discover during this development process.

1.3.3 Geographic Data Mining

Many of the traditional data-mining tasks discussed previously in this chapter have analogous tasks in the geographic data-mining domain. See Ester, Kriegel and Sander (1997) and Han and Kamber (2006) for overviews. Also, see Roddick and Spiliopoulou (1999) for a useful bibliography of spatiotemporal 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.

1.3.3.1 Spatial Classification and Capturing Spatial Dependency

Spatial classification builds up classification models based on 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.

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 spatiotemporal relationships among objects and events in addition to the nonspatial attributes of interest in traditional data mining. As noted above, ignoring spatial dependency and spatial heterogeneity effects in geographic data can result in misspecified models and erroneous conclusions. It also ignores a rich source of potential information.

In Chapter 6, also an updated chapter from the first edition, Shekhar, Vatsavai and Chawla discuss the effects of spatial dependency in spatial classification and prediction techniques. They discuss and compare the aspatial techniques of logistic regression and Bayesian classification with the spatial techniques of spatial autoregression and Markov random fields. Theoretical and experimental results suggest that the spatial techniques outperform the traditional methods with respect to accuracy and handling "salt and pepper" noise in the data.

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 implementations of spatial analytical techniques and spatial data-mining tools that exploit parallel and cyber infrastructure environments can complement recent work on parallel processing in standard data mining (see Zaki and Ho 2000).

1.3.3.2 Spatial Segmentation and Clustering

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 nonspatial 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 2002).

In Chapter 7, Han, Lee, and Kamber update a review of major spatial clustering methods recently developed in the data-mining literature. The first part of their chapter discusses spatial clustering methods. They classify spatial clustering methods into four categories, namely, partitioning, hierarchical, density-based, and grid-based. Although traditional *partitioning methods* such as k-means and k-medoids are not scalable, scalable versions of these tools are available (also see Ng and Han 2002). *Hierarchical methods* group objects into a tree-like structure that progressively reduces the search space. *Density-based methods* can find arbitrarily shaped clusters by growing from a seed as long as the density in its neighborhood exceeds certain thresholds. *Grid-based methods* divide the information spaces into a finite number of grid cells and cluster objects based on this structure.

The second part of Chapter 7 discusses clustering techniques for trajectory data, that is, data collected on phenomena that changes geographic location frequently with respect to time. As noted above, these data have become more prevalent since the publication of the first edition of this book; this section of the chapter is new material relative to the first edition. Although clustering techniques for trajectory data are not as well developed as purely spatial clustering techniques, there are two major types based on whether they cluster whole trajectories or can discover sub-trajectory clusters. *Probabilistic methods* use a regression mixture model to cluster entire trajectories, while *partition-and-group* methods can discover clusters involving sub-trajectories.

Closely related to clustering techniques are *medoid queries*. A medoid query selects points in a dataset (known as medoids) such that the average Euclidean distance between the remaining points and their closest medoid is minimized. The resulting assignments of points to medoids are clusters of the original spatial data, with the medoids being a compact description of each cluster. Medoids also can be interpreted

as facility locations in some problem contexts (see Murray and Estivill-Castro 1998). Mouratidis, Papadias, and Papadimitriou discuss medoids in Chapter 8.

1.3.3.3 Spatial Trends

Spatial trend detection involves finding patterns of change with respect to the neighborhood of some spatial object. Ester, Kriegel and Sander (1997) 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, then the neighborhood path is expanded in the next step.

In Chapter 9, a new chapter solicited for the second edition of this book, Fotheringham, Charlton, and Demšar describe the use of geographically weighted regression (GWR) as an exploratory technique. Traditional regression assumes that the relationships between dependent and independent variables are spatially constant across the study area. GWR allows the analyst to model the spatial heterogeneity and seek evidence whether the nonstationarity found is systematic or noise. This allows the analyst to ask additional questions about the structures in the data. GWR is also a technique that benefits greatly from GVis, and Fotheringham, Charlton, and Demšar use GVis analytics to examine some of the interactions in the GWR parameter surfaces and highlight local areas of interest.

1.3.3.4 Spatial Generalization

Geographic phenomena often have complex hierarchical dependencies. Examples include city systems, watersheds, location and travel choices, administrative regions, and transportation/telecommunications systems. *Spatial 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. *Spatial 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. *Nonspatial dominant generalization* generates aggregated spatial units that share the same high-level description. Attribute-oriented induction is used to aggregate nonspatial 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.

1.3.3.5 Spatial Association

Mining for *spatial association* 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 later), 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.

Chapter 10 by Malerba, Lanza, and Appice discusses INGENS 2.0, a prototype GIS that incorporates spatial data-mining techniques. Malerba and his co-authors reported on INGENS in the first edition of this book; their updated chapter indicates the progress that has been made on this software since 2001. INGENS is a Web-based, open, extensible architecture that integrates spatial data-mining techniques within a GIS environment. The current system incorporates an inductive learning algorithm that generates models of geographic objects from training examples and counter-examples as well as a system that discovers spatial association rules at multiple hierarchical levels. The authors illustrate the system through application to a topographic map repository.

1.3.4 Geovisualization

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 spatiotemporal 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 11, Gahegan updates his chapter from the first edition and argues that portraying geographic data in a form that a human can understand frees exploratory spatial analysis (ESA) from some of the representational constraints that GIS and

geographic data models impose. When GVis techniques fulfill their potential, they are not simply display technologies by which users gain a familiarity with new datasets or look for trends and outliers. Instead, they are environments that facilitate the discovery of new geographical concepts and processes and the formulation of new geographical questions. The visual technologies and supporting science are based on a wide range of scholarly fields, including information visualization, data mining, geography, human perception and cognition, machine learning, and data modeling.

Chapter 12 by Guo is a new chapter solicited for the second edition. In this chapter, Guo introduces an integrated approach to multivariate analysis and GVis. An integrated suite of techniques consists of methods that are visual and computational as well as complementary and competitive. The complementary methods examine data from different perspectives and provide a synoptic view of the complex patterns. The competitive methods validate and crosscheck each other. The integrated approach synthesizes information from different perspectives, but also leverages the power of computational tools to accommodate larger data sets than typical with visual methods alone.

1.3.5 Spatiotemporal and Mobile Objects Databases

Perhaps the most striking change in GKD and data mining since the publication of the first edition of this book in 2001 is the rise of spatiotemporal and mobile objects databases. The development and deployment of LATs and geosensor networks are creating an explosion of data on dynamic and mobile geographic phenomena, with a consequent increase in the potential to discover new knowledge about dynamic and mobile phenomena.

LATs are devices that can report their geographic location in near-real time. LATs typically exploit one or more georeferencing strategies, including radiolocation methods, GPS, and interpolation from known locations (Grejner-Brzezinska 2004). An emerging LAT is radiofrequency identification (RFID) tags. RFID tags are cheap and light devices attached to objects and transmit data to fixed readers using passive or active methods (Morville 2005).

LATs enable location-based services (LBS) that provide targeted information to individuals based on their geographic location though wireless communication networks and devices such as portable computers, PDAs, mobile phones, and in-vehicle navigation systems (Benson 2001). Services include emergency response, navigation, friend finding, traffic information, fleet management, local news, and concierge services (Spiekermann 2004). LBS are widely expected to be the "killer application" for wireless Internet devices; some predict worldwide deployment levels reaching one billion devices by 2010 (Bennahum 2001; Smyth 2001).

Another technology that can capture data on spatiotemporal and mobile phenomena is *geosensor networks*. These are interconnected, communicating, and georeferenced computing devices that monitor a geographic environment. The geographic scales monitored can range from a single room to an entire city or ecosystem. The devices are typically heterogeneous, ranging from temperature and humidity sensors to video cameras and other imagery capture devices. Geosensor networks can also capture the evolution of the phenomenon or environment over

time. Geosensor networks can provide fixed stations for tracking individual vehicles, identify traffic patterns, and determine possible stops for a vehicle as it travels across a given domain in the absence of mobile technologies such as GPS or RFID (Stefanidis 2006; Stefanidis and Nittel 2004).

In the first edition of this book, we included only one chapter dedicated to mining trajectory data (Smyth 2001). Recognizing the growth in mobile technologies and trajectory data, the second edition includes five new chapters on knowledge discovery from spatiotemporal and mobile objects databases.

In Chapter 13, Yuan proposes spatiotemporal constructs and a conceptual framework to lead knowledge discovery about geographic dynamics beyond what is directly recorded in spatiotemporal databases. Recognizing the central role of data representation in GKD, the framework develops geographic constructs at a higher level of conceptualization than location and geometry. For example, higher-level background knowledge about the phenomena in question can enhance the interpretation of an observed spatiotemporal pattern. Yuan's premise is that activities, events, and processes are general spatiotemporal constructs of geographic dynamics. Therefore, knowledge discovery about geographic dynamics ultimately aims to synthesize information about activities, events, or processes, and through this synthesis to obtain patterns and rules about their behaviors, interactions, and effects.

Chapter 14 by Wachowicz, Macedo, Renso, and Ligtenberg also addresses the issue of higher-level concepts to support spatiotemporal knowledge discovery. The authors note that although discovering spatiotemporal patterns in large databases is relatively easy, establishing their relevance and explaining their causes are very difficult. Solving these problems requires viewing knowledge discovery as a multitier process, with more sophisticated reasoning modes used to help us understand what makes patterns structurally and meaningfully different from another. Chapter 14 proposes a multitier ontological framework consisting of domain, application, and data ontology tiers. Their approach integrates knowledge representation and data representation in the knowledge discovery process.

In Chapter 15, Cao, Mamoulis, and Cheung focus on discovering knowledge about periodic movements from trajectory data. Discovering periodic patterns (that is, objects following approximately the same routes over regular time intervals) is a difficult problem since these patterns are often not explicitly defined but rather must be discovered from the data. In addition, the objects are not expected to follow the exact patterns but similar ones, making the knowledge discovery process more challenging. Therefore, an effective method needs to discover not only the patterns themselves, but also a description of how they can vary. The authors discuss three algorithms for discovering period motion: an effective but computationally burdensome bottom-up approach and two faster top-down approaches.

Chapter 16 by Laube and Duckham discusses the idea of decentralized spatiotemporal data mining using geosensor networks. In this approach, each sensor-based computing node only possesses local knowledge of its immediate neighborhood. Global knowledge emerges through cooperation and information exchange among network nodes. Laube and Duckham discuss four strategies for decentralized spatial data mining and illustrate their approach using spatial clustering algorithms.

In the final chapter of the book, Kraak and Huisman discuss the *space–time cube* (STC) an interactive environment for the analysis and visualization of spatiotemporal data. Using Hägerstrand's time geographic framework as a conceptual foundation, they illustrate the STC using two examples from the domain of human movement and activities. The first example examines individual movement and the degree to which knowledge can be discovered by linking attribute data to space–time movement data, and demonstrates how the STC can be deployed to query and investigate (individual-level) dynamic processes. The second example draws on the geometry of the STC as an environment for data mining through space–time query and analysis. These two examples provide the basis of a broader discussion regarding the common elements of various disciplines and research areas concerned with moving object databases, dynamics, geocomputation, and GVis.

1.4 CONCLUSION

Due to explosive growth and wide availability of geo-referenced data in recent 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 represent important directions in the development of a new generation of spatial analysis tools in data-rich environment. In this chapter, we introduce knowledge discovery from databases and data mining, with special reference to 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 GKD 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 GVis tools. Application of data mining and knowledge discovery techniques to spatiotemporal and mobile objects databases is also a rapidly emerging subfield of GKD. 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 contribute to developing new methods and applications in this promising field.

REFERENCES

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.

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.

Bennahum, D.S. (2001) "Be here now," *Wired*, 9.11, 159–163.

Benson, J. (2001) "LBS technology delivers information where and when it's needed," *Business Geographics*, 9(2), 20–22.

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.

Bishr, Y. (2007) "Overcoming the semantic and other barriers to GIS interoperability: Seven years on," in P. Fisher (Ed.) *Classics from IJGIS: Twenty Years of the International Journal of Geographical Information Science and Systems*, London: Taylor & Francis, 447–452.

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 & 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 & 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 & 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.

Egenhofer, M. (2002) "Toward the semantic geospatial web," Geographic Information Systems: Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, New York: ACM Press, 1–4.

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.

Elder, J. and Pregibon, D. (1996) "A statistical perspective on knowledge discovery," 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., Grinstein, G. and Wierse, A. (2001) *Information Visualization in Data Mining and Knowledge Discovery*, San Mateo, CA: Morgan Kaufmann.

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.

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

Fotheringham, A.S. and Rogerson, P.A. (1993) "GIS and spatial analytical problems," International *Journal of Geographical Information Science*, 7, 3–19.

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

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.

Goodchild, M.F. (2004) "A general framework for error analysis in measurement-based GIS," *Journal of Geographical Systems*, 6, 323–324.

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.

Grejner-Brzezinska, D. (2004) "Positioning and tracking approaches and technologies," in H.A. Karimi and A. Hammad (Eds.) *Telegeoinformatics: Location-Based Computing and Services*, Boca Raton, FL: CRC Press, 69–110.

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

Guan, Q., Zhang, T. and Clarke, K.C. (2006) "Geocomputation in the grid computing age," in J.D. Carswell and T. Tezuka (Eds.) *Web and Wireless Geographical Information Systems*: *6th International Symposium, W2GIS 2006, Hong Kong, China, December 4-5, 2006, Proceedings,* Berlin: Springer Lecture Notes in Computer Science 4295, 237–246.

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 Fu, Y. (1996). "Attribute-oriented induction in data mining." In U.M. Fayyad, G., Piatetsky-Shpiro, P., Sayth, and R. Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press/the MIT Press pp. 399–424.

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd ed., San Mateo, CA: Morgan Kaufmann.

Han, J., Stefanovic, N. and Koperski, K. (2000) "Object-based selective materialization for efficient implementation of spatial data cubes," *IEEE Trans. Knowledge and Data Engineering*, 12(6), 938–958.

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.

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

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.

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.

Miller, H.J. and Wentz, E.A. (2003) "Representation and spatial analysis in geographic information systems," *Annals of the Association of American Geographers*, 93, 574–594.

Morville, P. (2005) *Ambient Findability: What We Find Changes Who We Become*, Sebastopol, CA: O'Reilly Media.

Müller, J.-C. (1982) "Non-Euclidean geographic spaces: Mapping functional distances," *Geographical Analysis*, 14, 189–203.

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*, Washington, D.C.: National Academy Press.

Ng, R.T. and Han, J. (2002) "CLARANS: A method for clustering objects for spatial data mining," *IEEE Transactions on Knowledge and Data Engineering*, 14(5), 1003–1016.

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 & Francis, 65–79.

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.

Petrushin V.A. and Khan, L. (2006) *Multimedia Data Mining and Knowledge Discovery*, New York: Springer-Verlag.

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, 267–278.

Silberschatz, A., Korth, H.F. and Sudarshan, S. (1997) *Database Systems Concepts*, 3rd ed., New York, NY: 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.

Smyth, C.S. (2001) "Mining mobile trajectories," H.J. Miller and J. Han (Eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor & Francis, 337–361.

Spiekerman, S. (2004) "General aspects of location-based services," in J. Schiller and A. Voisard (Eds.) *Location-Based Services*, San Francisco, CA: Morgan Kaufmann, 9–26.

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.

Stefanidis, A. (2006) "The emergence of geosensor networks," *Location Intelligence*, 27 February 2006; http://locationintelligence.net.

Stefanidis, A. and Nittel, S. (Eds.) (2004) *GeoSensor Networks*, Boca Raton, FL: CRC Press.

Tan, P.-N., Kumar, V. and Srivastava, J. (2002) "Selecting the right interestingness measure for association patterns," Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02), Edmonton, Canada, 32–41.

Tobler, W. R. (1994) Bidimensional regression," 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

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