Chapter 9
Knowledge Discovery from Geographical Data

Salvatore Rinzivillo, Franco Turini, Vania Bogorny, Christine Körner, Bart Kuijpers, and Michael May

9.1 Introduction

During the last decade, data miners became aware of geographical data. Today, knowledge discovery from geographic data is still an open research field but promises to be a solid starting point for developing solutions for mining spatio-temporal patterns in a knowledge-rich territory. As many concepts of geographic feature extraction and data mining are not commonly known within the data mining community, but need to be understood before advancing to spatio-temporal data mining, this chapter provides an introduction to basic concepts of knowledge discovery from geographical data.

In performing knowledge discovery in a spatial data set, the first important question is how to use the spatial dimension in the discovery process. At least two viewpoints can be considered: either spatial relationships are made explicit prior to data mining or specialised algorithms are directly applied to spatial and non-spatial data. The first viewpoint claims that the spatial dimension is somewhat more basic then the other features, and, then, it can be used to prepare the data set for a successive knowledge extraction step. The exploitation of the spatial dimension for selecting the values of attributes to be used in the mining step can be quite complex, and it may depend both on the structure of the domain and on the kind of model one is looking for. This first approach offers the advantage of allowing the reuse of standard data mining technology on data extracted according to the spatial dimension.

Salvatore Rinzivillo, Franco Turini
KDD Laboratory, Dipartimento di Informatica, Università di Pisa, Italy, e-mail: {rinziv,turini}@di.unipi.it

Vania Bogorny, Bart Kuijpers
Theoretical Computer Science Group, Hasselt University and Transnational University of Limburg, Belgium, e-mail: {vania.bogorny,bart.kuijpers}@uhasselt.be

Christine Körner, Michael May
Fraunhofer Institut Intelligente Analyse- und Informationssysteme, Sankt Augustin, Germany, e-mail: {christine.koerner,michael.may}@iais.fraunhofer.de

The second approach aims at exploiting the spatial features dynamically during the discovery process. This implies a complete reinvention of the data mining technology, but it allows a more flexible use of spatial knowledge.

Mining geographic data poses additional challenges which include the exploitation of background knowledge as well as the handling of spatial autocorrelation and highly erroneous data. Although many data mining algorithms extend over multi-dimensional feature spaces and are thus inherently spatial, they are not necessarily adequate to model geographic space. The first specialised algorithms for geo-referenced data were introduced by Koperski and Han [25] and Ester et al. [14].

This chapter provides an overview of knowledge discovery from geographic data. In Section 9.2 we revise basic spatial concepts and the representation of geographic data. Section 9.3 introduces Geographic Information Systems (GIS) and first approaches to enrich these systems with data mining capabilities. Section 9.4 focuses on the extraction of implicit features and relationships from geographic data. Algorithms for mining geo-referenced data are discussed in Section 9.5, and in the subsequent section we provide an example that connects all presented aspects of the knowledge discovery process. In Section 9.7, we construct a roadmap which views approaches to geographic knowledge discovery in the light of spatio-temporal data, and we conclude the chapter with a short summary.

9.2 Geographic Data Representation and Modelling

9.2.1 Conceptual Models of Space

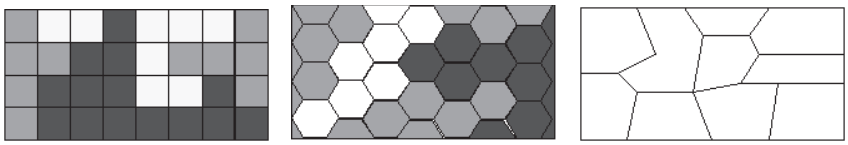
Conceptual models are an abstract representation of reality, reflecting the main characteristics of objects and events from a user's point of view. In the spatial domain they depict measurements or observations (of objects) referenced in space. Conceptual models of space are independent of any restrictions imposed by a subsequent representation in information systems. On the conceptual level, two major approaches can be distinguished [19, 27]. The first model regards the spatial domain as a continuous surface, each point of which can be mapped to one and only one value of some attribute. This paradigm is called *field* model and represents a function of location in two- or three-dimensional space. Typical applications of field models are measurements of mineral and pollutant concentrations or temperature in soil and air. The second conceptual model is based on discrete *objects*. An object is either of type point, line or polygon which may represent a tree, street or city respectively. In contrast to fields, the world of object models is empty except for places that are occupied by objects.

9.2.2 Representation of Spatial Data

The continuous geographic space must be digitalised before it can be stored in a computer. Two main data structures, tessellation and vector, have been developed to represent geographic data in a discrete way. Although apparently related, both forms can be used to represent the concept of fields or objects.

Tessellation

Tessellation models partition the space into a number of cells which each store a value of the associated attribute [37, 7]. The grid can be regularly spaced, in this case it is also called a raster, or irregularly spaced. Figure 9.1 shows a regular grid of square and hexagonal cells as well as an irregular tessellation. The intensity of colour indicates different attribute values. All variation within a cell is lost. Thus, the size of the cell defines the level of resolution. Regular tessellation models possess very efficient indexing structures (run-length encoding, quadtrees) and are well-suited to model continuous change. Their disadvantages include the memory space and computational costs involved to manage high resolutions. Regular tessellation models are commonly applied to satellite and environmental data. Irregular tessellations are used, for example, to represent administrative units.



(a) Square cells (b) Hexagonal cells (c) Irregular cells

Fig. 9.1 Regular and irregular tesselations

Vector

The vector model is most commonly used in current Geographic Information Systems [37, 26]. In the vector model, infinite sets of points in space are represented as finite geometric structures. More precisely, a vector datum consists of a tuple of the form (*geometry*, *attributes*), where a geometry can be a point, polyline or polygon. A point is typically given by its rational coordinates. A polyline is represented by a sequence of points and a polygon takes the form of a closed polyline. Examples of vector data are shown in Figure 9.2. The advantage of the vector model is the concise representation of objects. However, it involves complex data structures, and

the computation of spatial operations, such as intersection and overlay, may take considerable time and resources [37, 7].

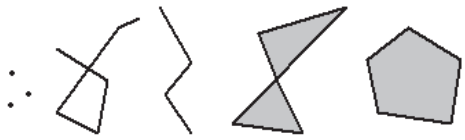


Fig. 9.2 Vectors: points, polylines and polygons

9.3 Geographic Information Systems

9.3.1 Definition

Geographic Information Systems (GIS) have been defined in many ways. Today, it is no longer easy to give a clear definition of GIS. During the past twenty or more years GIS have evolved from Systems to Science, a complex interaction of theory, technology and systems. But the main concern of GIS remains to handle geographic information about places on the earth's surface and to deal with knowledge about *what is where when* (recently, also time has taken its place in GIS [35]). A popular definition [10, 21] says that a GIS is "a system of hardware, software, and procedures designed to support the capture, management, manipulation, analysis, modelling and display of spatially-referenced data for solving complex planning and management problems."

On its technological side, GIS rely on techniques like Global Positioning Systems (GPS) and Remote Sensing. In the past, GIS have cynically been called "maps with a database behind it", but the data models allow a complex representation of the real world that can support querying, analysis and decision support.

The data stored in a GIS is typically divided over thematic layers. Grosso modo, we can say that each of these layers is modelled in the tessellation or vector model as described above.

9.3.2 Thematic Layers

Real world data contains many different aspects. In the description of a city or a region we can, for instance, distinguish between the road network, cadastral information about parcels and houses, hydrographic information, topography (terrain elevation), etc. Following this thematic division, data in a GIS is typically organized

by *layers*, which correspond to themes in the application. For instance, one layer could contain information about the road network, whereas another could contain information about the rivers and lakes and yet another could contain information on elevation. Although data is divided over thematic layers, there is a way of integrating different layers, namely using explicit location on the earth's surface. Using the geographic location as an organizing principal between layers, they can be overlayed or spatially joined.

Each layer represents a common feature and therefore the information in one layer is of a similar type, whereas information in different layers may be of quite different nature. Layers are described by two types of data: spatial data which describes the location of objects and thematic or attribute data which specifies the characteristics of the data in a traditional alpha-numeric way (this data are usually stored in a relational database). The spatial part of the information within one layer can be stored in any physical representation, depending on the need of the data and the application.

9.3.3 Integration of Knowledge Discovery and GIS

Most commercially available GIS provide extended functionality to store, manipulate and visualise geo-referenced data, but rely on the user's ability for exploratory data analysis. This approach is not feasible with regard to the large amount and high dimensionality of geographic data, and demands for integrated data mining technology.

The integration of data mining methods and GIS functionality does not only facilitate data analysis, but also allows for an efficient implementation of algorithms. One prospective area is spatial feature extraction. In general, the application of spatial operations for feature extraction is computationally expensive. When the feature extraction and data mining step are interweaved, a dynamic selection of objects, for which some spatial relationship must be computed, can take place.

To our best knowledge, there are only a few software systems that join the power of data mining techniques and GIS, namely GeoMiner, SPIN! and INGENS. GeoMiner [20] has been among the first approaches to mine geographic data from large spatial databases and focuses on the discovery of spatial association rules. SPIN! [31, 1] is a spatial data mining platform that integrates several algorithms for spatial data mining, which include multi-relational subgroup discovery, rule induction and spatial cluster analysis. It pays special attention to the scalability of algorithms allowing for a tight coupling with the database, and it provides an extensive interface for visual data exploration. INGENS [29] (INductive GEographic iNformation System) is a prototypical GIS which possesses an inductive learning capability. It can generate first-order logic descriptions for geographic objects, and it includes a training facility that allows the interactive selection of examples and counter examples of geographic concepts.

9.4 Spatial Feature Extraction

A spatial feature describes some characteristic of a geographic object. We use the term *feature* in compliance with the definition commonly used in data mining, and not according to the Open GIS Consortium terminology where it corresponds to a real world or abstract entity [34].

A geographic object is characterised by a spatial component, e.g. a geometric object that represents its position in the geographic space, and a set of attributes that describe the non-spatial dimensions of the object, e.g. the type of a road or the construction year of a building. While the non-spatial information can be queried in traditional ways, the information of spatial relationships is implicitly encoded and must be extracted prior to data mining. Spatial feature extraction poses the challenge to reveal meaningful information of geographic objects, with a particular interest in their relationships.

This section describes relation-based and aggregation-based spatial features. It gives an overview of the state-of-the-art of spatial feature extraction. Finally, we conclude with the enhancement of feature extraction using background knowledge.

9.4.1 Relational Features

Information about spatial objects can be derived from single objects or from the relationship between two or more objects. The former are called *unary* features (such as length, area and perimeter), the latter *relational* features. Probably the most prominent relational feature is the *distance* between two objects, which can be measured, for example, using Euclidean distance. In this section we will give an introduction to two further relational features of spatial objects, namely *topological* and *directional* relations.

9.4.1.1 Topological Relations

The topological relations are invariant under homomorphisms, i.e. they are preserved if the considered objects are rotated, scaled, or moved. The formal definition of such relations is based on the *point-set topology* theory. Each geometry G is considered as composed of three parts: the *interior* (denoted by G°), i.e. the set of all the points that are inside the geometry; the *boundary* (denoted by δG), i.e. the limit of the geometry and the *exterior* (denoted by G^{-}), i.e. all the remaining points that do not belong to the object.

The *9-intersection model*, proposed in [12], gives a formal description of the topological relation between two objects. The model is based on the evaluation of all the possible intersection among the interiors, the boundaries and the exteriors of two objects. In particular, given two geometries A and B, the nine possible intersections define the relation between the two geometries, and it is represented by means of

the *9-intersection matrix*:

$$R(A,B) = \begin{pmatrix} \delta A \cap \delta B & \delta A \cap B^{\circ} & \delta A \cap B^{-} \\ A^{\circ} \cap \delta B & A^{\circ} \cap B^{\circ} & A^{\circ} \cap B^{-} \\ A^{-} \cap \delta B & A^{-} \cap B^{\circ} & A^{-} \cap B^{-} \end{pmatrix}$$

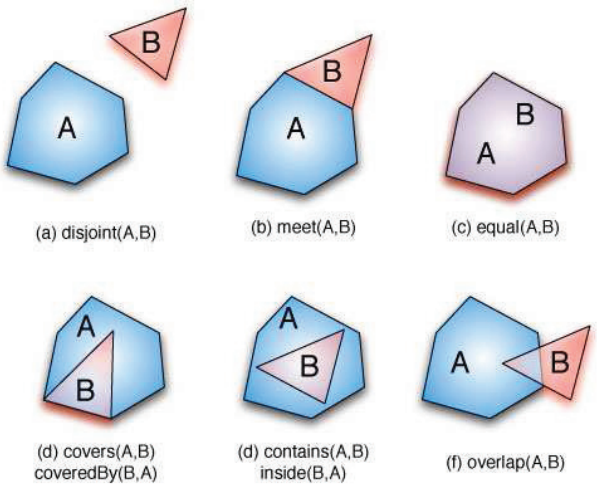


Fig. 9.3 Topological relations

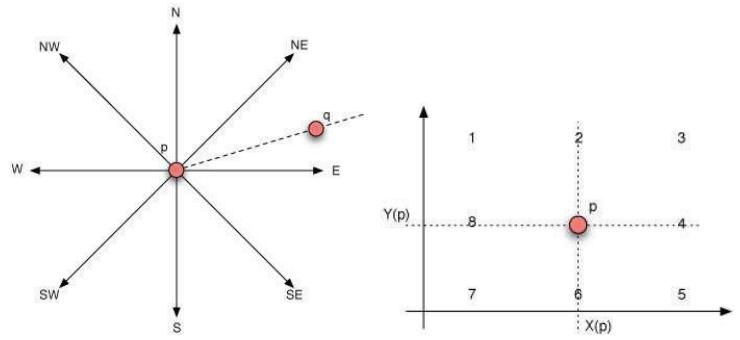
Each of these intersections is tested if it is empty or not, which results in a total of 2^9 combinations. However, many of these cases can be discarded due to geometric properties of the considered objects. For example, if we consider two 2-dimensional objects, say A and B, there are only eight possible relations between A and B (showed in Figure 9.3), i.e. there are eight possible distinct configurations of the matrix.

9.4.1.2 Directional Relations

Directional relations are defined over a reference system determined by two orthogonal axes, x and y. Based on relationships between point objects, the definition of directional relations can be extended to objects of arbitrary shape [36]. The approaches used for the formal definition of directional relations are mainly based on two methods [17]: *cone-shaped areas* and *projections*.

The *cone-shaped areas* method relates the cardinal direction between two points p and q by considering the angular direction with reference to some fixed direction in space. For example, the directional symbols for the system presented in Figure 9.4(a) are: $V_9 = \{N, NE, E, SE, S, SW, W, NW, 0\}$. The direction 0 (zero) represents the situations when two points are not distinct. The direction through the two

points p and q is closer to the E direction, therefore the symbol E is assigned to the direction pq.

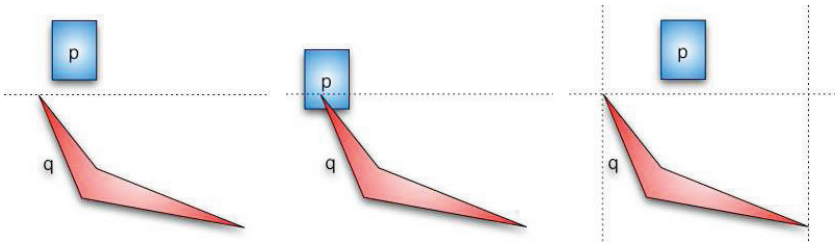


(a) Named directions for cone-shaped reference systems (b) Plane partitions using one reference point

Fig. 9.4 Examples of directional relations

The *projection-based* method uses projection lines to determine the direction between two points in space. Let us consider a reference point p. If we draw two orthogonal projections from point p we obtain nine partitions of the space (four open line segments, four open regions and the intersection point, see Figure 9.4(b)). The position of a second point q inside one of these regions determines the direction.

Directional relations can be generalised for two objects of arbitrary shape using the above definitions. Given two spatial objects p and q, we denote with p_i (q_i) a generic point of object p (q). The relation $strong\_north \equiv \forall p_i \forall q_i north(p_i, q_i)$ denotes that all the point of p are north of all points of q (Figure 9.5(a)). The relation *weak_north* holds when some points of p are north of all point of q, for each point of p there exist some points of q such that p_i is north of q_i, and some points of p are south of some points of q (Figure 9.5(b)).



(a) *strong_north* (b) *weak_north* (c) *strong_bounded_north*

Fig. 9.5 Directional relations for extended spatial objects

Directional relations can be defined by using the minimum bounding box (MBB) approximation. A MBB is the axis-parallel rectangle that is spanned by the two coordinates $c_1 = (min(x_p), min(y_p))$ and $c_2 = (max(x_p), max(y_p))$. We denote with $MBB(p)$ the bounding box of object p. The relation

$$strong\_bounded\_north(p,q) \equiv strong\_north(MBB(p), MBB(q))$$

holds when all the points of p are bounded by the horizontal line that passes through the northernmost point of q and by the two vertical lines that also bound q (Figure 9.5(c)). Similarly, the relations for other cardinal directions can be defined.

9.4.2 Spatial Aggregation

Aggregation of data is commonly applied to summarise information and to derive features that cannot be measured at a single point. Within the spatial domain, aggregation is also used to attach information about the local environment to some entity. For example, in order to compare birth rates of European countries the number of live births and inhabitants must be summarised for each country. The birth rate itself is a variable that cannot be measured at a single location but must be derived for some areal unit. For urban planning a smaller areal unit may be chosen. For example, a city council might evaluate locations for a new kindergarten based on socio-demographic data of the respective municipal districts.

This example shows that spatial information can be aggregated at several levels of resolution. The choice of resolution is not always obvious, which gives rise to the modifiable areal unit problem. The modifiable areal unit problem [16] comprises two parts, the scale effect and the zoning effect. The scale effect may lead to different statistical results if information is grouped at different levels of spatial resolution. The zoning effect refers to the variability of statistical results if the borders of spatial units are differently chosen at a given scale of resolution. Both effects need to be carefully considered when aggregating spatial data.

In the example above, administrative borders were chosen to define spatial units. This paragraph presents two techniques to aggregate spatial data based on distance and time. Distance-based units, also called buffers, contain all objects that lie within a predefined distance to the object in question. Continuing the example from above, the city council could count the number of households with young children within a 2 km distance to each potential location. Yet, distance alone may not always yield the desired result. Imagine, one location is situated close to a river and no bridge is nearby. In this case, it may be more important *how long* it takes for a parent to reach the kindergarten. Units that contain all objects reachable within a given amount of time are called drive time zones. They are calculated using Dijkstra's algorithm to determine the shortest path between two points in a graph. The graph is formed by the underlying street network with edges weighted according to the average or

maximal speed allowed. Figure 9.6 contrasts a 1500 m buffer and a 4 minute drive time zone in the middle of Frankfurt, Germany.

(a) 1500 m buffer (b) 4 minute drive time zone

Fig. 9.6 Aggregations within Frankfurt, Germany

9.4.3 State-of-the-art Feature Extraction

The extraction of spatial features from geographic data such as topological and distance, is the most effort and time consuming step in the whole discovery process [41], but only little attention has been devoted to this problem. On one side the user must choose the appropriate spatial and non-spatial features. On the other side the extraction process itself requires high computational costs. Spatial features can be extracted from geographic data by functionalities provided by GIS and geographic database management systems. Several approaches to extract spatial features for data mining and knowledge discovery have been proposed. Spatial features can be extracted either in the data preprocessing or during the data mining task.

Most approaches extract spatial features in data preprocessing, where any spatial relation may be computed and geographic objects may have any geometric representation (e.g. point, line). In [25] a top-down progressive refinement method is proposed and spatial approximations are calculated in a first step. In a second step, more precise spatial relationships are computed to the outcome of the first step. This method has been implemented in the GeoMiner system [20]. [13] proposed new operations such as graphs and paths to compute spatial neighbourhoods. However, these operations are not implemented by most GIS, and to compute all spatial relationships between all geographic objects in order to obtain the graphs and paths is computationally expensive for real applications. In [30] all spatial relationships are computed and converted to a first order logic database. This process is computationally expensive for real problems and many spatial relationships might be unnecessarily extracted. A feature extraction module named Featex has been im-

plemented in the Ares system [4], where the user can choose the geographic object types and non-spatial attributes. An approach that uses geo-ontologies as prior knowledge to filter spatial features has been proposed by [6]. In this approach the semantics of geographic objects is considered, and geo-ontologies are used to compute only topological features semantically consistent.

The approach of [22] deals with geographic coordinates directly and extracts spatial features during the mining task, yet it considers only distance features. Another drawback is the input restriction to point primitives. For geographic objects represented by n-dimensional primitives, their centroid is extracted. This process may lose significant information and generates imprecise patterns. For example, the Mississippi River intersects many states considering its real geometry, but it will be far from the same states if only the centroid is considered.

9.4.4 Improvement of Feature Extraction Using Background Knowledge

In geographic space, many features represent natural geographic dependences in which some objects are *always* related to other objects. For instance, islands are naturally *within* water bodies, ports are naturally *adjacent* to water bodies, bus stops *intersect* roads.

A large amount of natural geographic dependences is well-known by geographers and geographic database designers. These dependences are normally explicitly represented in geographic database schemas through one-to-one and one-to-many cardinality constraints [41] in order to warrant the spatial integrity of geographic data [40]. Since natural dependences are intrinsic to geographic data they are also represented in geographic ontologies [6]. Geo-ontologies and geographic database schemas are rich knowledge repositories that can be used as background knowledge to accelerate the spatial feature extraction.

Well-known spatial features require computational time to be extracted and generate patterns which are in most cases non-novel and non-interesting for data mining [5]. But which spatial features can be considered either interesting or well-known? Figure 9.7 shows a geographic map in which it is possible to visualize that all bus stops intersect streets. It is well-known that bus stops only exist on streets, but their spatial relationship is normally not explicitly stored in geographic databases, and needs to be extracted with functionalities provided by GIS. Considering this example, the topological feature between both geographic objects could be retrieved from the knowledge base instead of performing a spatial join operation. Since both geographic objects have a mandatory topological relationship, the distance between these objects is zero, and no spatial operation is required to extract either topological or distance features.

Background knowledge can be used to improve the discovery process from many different perspectives, but only a few approaches have used prior knowledge in geographic data mining. In [4] prior knowledge is defined by the data mining user and

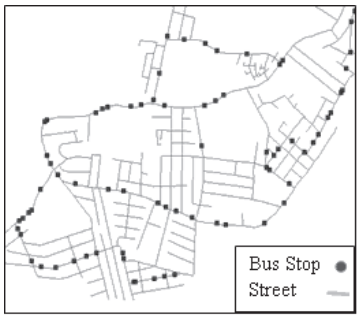


Fig. 9.7 Region of the Porto Alegre city

is used to reduce the number of well-known patterns. In [5] background knowledge is extracted from geographic database schemas to reduce spatial joins in feature extraction and to reduce well-known patterns. In [6] background knowledge is extracted from geo-ontologies and is used to improve topological feature extraction. In this approach only topological relationships that are semantically consistent are computed.

9.5 Spatial Data Mining

This section presents an overview of spatial data mining techniques that are applied to geographic data. It is important to notice the difference between the term *spatial* and *geographic* [33]: "'Spatial' concerns any phenomena where the data objects can be embedded within some formal space that generates implicit relationships among the objects. [...] 'Geographic' refers to the specific case where the data objects are georeferenced and the embedding space relates (at least conceptually) to locations on or near the Earth's surface." Spatial data mining thus includes geographic data mining as a special case.

In the next section we describe challenges of spatial data mining that arise due to the nature of geographic data. The remaining sections present recent approaches to clustering, classification, regression, association rule mining and subgroup discovery using geographic data.

9.5.1 Challenges for Mining Geographic Data

Geographic data often violate assumptions that are essential to traditional data mining techniques. The most predominant characteristic of geographic data is known as *Tobler's Law* [42], which states that "[...] everything is related to everything else,

but near things are more related than distant things." It means that attribute values of spatial objects are stronger correlated the closer two objects are in location. Usually, geographic objects exhibit strong positive autocorrelation and show similar values within their local neighbourhood. This behaviour directly contrasts the often made assumption of independent, identical distributions in classical data mining and causes poor performance of algorithms that ignore autocorrelation [8]. A second characteristic of geographic data is its variation across several scales of resolution. Dependencies on a small scale turn into random variation when analysed using broader units of measurement. Thus, discovered patterns depend on the choice of resolution and are subject to random variation. A third challenge for mining geographic data pose the implicitly defined relationships between spatial objects. They can be extracted as described in Section 9.4 either previously to the application of algorithms or dynamically [3].

In general there are two alternatives how algorithms treat geographic data. The first approach uses traditional algorithms and includes spatial attributes either as ordinary variables or requires feature extraction during pre-processing. The second approach relies on specialised algorithms that incorporate feature extraction or are able to handle geographic dependencies directly. In the remaining section we present several algorithms that are applied to geographic data and emphasise their strategy for feature extraction and their ability to handle autocorrelation.

9.5.2 Clustering

Clustering divides a given set of objects into non-overlapping groups, such that similar objects are within the same group and objects of different groups are most heterogeneous. As clustering relies on the distance between objects, it is inherently spatial. Yet, the assumption of convex clusters (e.g. k-means) is inappropriate for many geographical data sets (see Figure 9.8). Ester et al. [14] developed a density based algorithm for point data that finds clusters of arbitrary shape. The idea of this approach is that a cluster can be recognized by a high density of points within, while only few points are found in the surrounding environment. It requires the definition of a *neighbourhood*, which is used to iteratively join points, and a *density* which is used to delineate the borders of a cluster. In [39] this approach is extended to cluster vector data (e.g. polygons).

(a) Convex shape (b) Arbitrary shape

Fig. 9.8 Spatial clusters

9.5.3 Classification and Regression

In classification and regression, the unknown target value of some object is predicted given a set of training instances. If the target variable is discrete, the learning task is called classification. If it is continuous it is referred to as regression. We start with the well-known k-nearest neighbour method, which can be applied to both, classification and regression tasks. The second part presents spatial model trees, geographically weighted regression, and we conclude this section with Kriging. Kriging is a popular regression method in geostatistics and takes explicitly advantage of autocorrelation.

k-Nearest Neighbour

The *k*-nearest neighbour algorithm (kNN) is an instance based learning method that classifies unknown instances according to the target value of the *k* most similar training examples. It assumes that objects with similar characteristics also possess similar class values. In case of classification, the most frequent target value among the neighbours will be assigned to the instance. In case of regression, a (weighted) mean is calculated. In order to determine the similarity between two objects, kNN requires a distance measure for each attribute. As geographic coordinates can be used to determine the distance between two locations, they can be directly included in the algorithm. Thus, kNN relies on objects that are within the geographic neighbourhood and exploits positive autocorrelation of the target variable.

Model Trees

Model trees [45] operate similar to decision trees, but possess leaves that are associated with (linear) functions instead of fixed values. While internal nodes of the tree partition the sample space, leave nodes construct local models for each part of the sample space. Malerba, Ceci and Appice [28] developed a spatial model tree which is able to model local as well as global effects. Their induction method, Mrs-SMOTI (Multi-relational Spatial Stepwise Model Tree Induction), places regression nodes also within inner nodes of the tree and passes these regression parameters to all child nodes. Mrs-SMOTI exploits spatial relationships over several layers and possesses a tight database integration to extract spatial relations during the induction phase.

Geographically Weighted Regression

Geographically weighted regression (GWR) [15] extends the traditional regression framework such that all parameters are estimated within a local context. The model for some variable *z* at location *i* then takes the following form:

$$z_i = \beta_0(x_{ix}, x_{iy}) + \sum_k \beta_k(x_{ix}, x_{iy}) x_{ik} + \varepsilon_i.$$

In the equation above, (x_{ix}, x_{iy}) denotes the pair of coordinates at location i, $\beta_k(x_{ix}, x_{iy})$ is the localised parameter for attribute k, x_{ik} is the value of attribute k at location i and ε_i denotes random noise. The GWR model assumes that all parameters are spatially consistent. Therefore, parameters at location i are estimated from measurements close to i. This is realised by the introduction of a diagonal weight matrix W_i which states the influence of each measurement for the estimation of regression parameters at i:

$$\widehat{\beta}(x_{ix}, x_{iy}) = (X^T W_i X)^{-1} X^T W_i \, z.$$

The weight matrix can be built according to several weighting schemes, such as a Gaussian or bi-square function. GWR is a local regression method which takes advantage of positive autocorrelation between neighbouring points in space.

Kriging

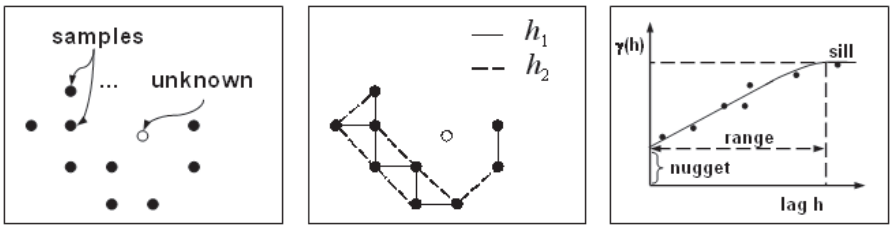
Kriging [44, 9, 11] is an optimal linear interpolation method to estimate unknown values in geographic field data. Let x denote a location in an index set $D \subset \mathbb{R}^n$ in n-dimensional space and $Z(x)$ a random variable of interest at location x. Generally, each variable $Z(x)$ can be decomposed into three terms: a structural component representing a mean or constant trend, a random but spatially correlated component that denotes autocorrelation, and random noise expressing measurement errors or variation inherent to the variable of interest [7].

A technique most widely used is Ordinary Kriging, which assumes intrinsic stationarity with an unknown, but constant mean of the random target variable $Z(x)$. Given a set of measurements, Kriging estimates unknown values as weighted sum of neighbouring sample data (Figure 9.9(a)) and uses the *variogram* to determine optimal weights (Figure 9.9(c)). Variograms model spatial dependency between locations and are a function of distance for any pair of sites:

$$\gamma(h) = \frac{1}{2} Var[Z(x+h) - Z(x)].$$

A variogram of the data can be obtained in two steps. First, the experimental variogram is calculated from the sample by calculating the variance between samples for all increments h. Figure 9.9(b) shows all pairs of sample points with a lag h_1 (solid lines) and a second lag of h_2 (dashed lines). In a second step, the experimental variogram serves to fit a theoretical variogram which is used in Ordinary Kriging. Depending on the data, different model types may be appropriate for the theoretical variogram. Often, a spherical model is used and its parameters are adapted to reflect the experimental variogram. Each variogram is characterised by three parameters: nugget, range and sill as depicted in Figure 9.9(c). The nugget effect represents random noise, as by definition $\gamma(0) = 0$. Within the range, the variance of increments

increases gradually with distance in this example. It directly shows the spatial dependency. The closer two points are the more likely are their values similar. Finally, the curve levels off at the sill. The variance has reached its maximum value and is independent of distance.



(a) Data sample (b) Lag between samples (c) Variogram

Fig. 9.9 Variance of sample increments

Ordinary Kriging estimates the unknown value at a location x_0 as weighted sum of neighbouring sample points x_i $(i = 1 \ldots n)$:

$$Z^*(x_0) = \sum_{i=1}^{n} w_i \, Z(x_i).$$

The weights w_i are determined in conformance with two restrictions. First, $Z^*(x_0)$ must be an unbiased estimate of the true value $Z(x_0)$, which means that on average the prediction error for location x_0 is zero. Because the model assumes a constant mean $m = E\left[Z(x_i)\right]$ $(i = 0..n)$, this claim bounds the sum of weights to one.

$$\begin{aligned} 0 = E\left[Z^*(x_0) - Z(x_0)\right] &= E\left[\sum_{i=1}^{n} w_i \, Z(x_i) - Z(x_0)\right] \\ &= m\left(\sum_{i=1}^{n} w_i - 1\right) \qquad \Rightarrow \sum_{i=1}^{n} w_i = 1 \end{aligned}$$

Second, we require an optimal estimate which minimizes the error variance σ_E^2 of the estimate. The second equation expresses the variance in terms of the variogram.

$$\begin{aligned} \sigma_E^2 &= Var\left(Z^*(x_0) - Z(x_0)\right) = E\left[\left(Z^*(x_0) - Z(x_0)\right)^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(x_i - x_j) - 2\sum_{i=1}^{n} w_i \gamma(x_i - x_0) + \gamma(x_0 - x_0) \end{aligned}$$

The derivatives of the error variance with respect to $w_i (i = 1..n)$ yield a linear system of n equations. In combination with the restriction on the weights, a La-

grange parameter ϕ is introduced and a total of $n+1$ equations is obtained. For each location x_0, the optimal weights w_i are estimated using the following system of equations, given in matrix form:

$$\begin{pmatrix} \gamma(x_1-x_1) & \dots & \gamma(x_1-x_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \gamma(x_n-x_1) & \dots & \gamma(x_n-x_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \phi \end{pmatrix} = \begin{pmatrix} \gamma(x_1-x_0) \\ \vdots \\ \gamma(x_n-x_0) \\ 1 \end{pmatrix}$$

Note that Ordinary Kriging is an exact interpolator. If the value of a location in the data sample is estimated, it will be identical with the measured value. Several variants of Kriging have been developed which extend interpolation to data that contains a trend (Universal Kriging [9, 11]), involves uncertainty (Bayesian Kriging [9]) or contains temporal relations (Spatio-temporal Kriging).

9.5.4 Association Rules

Association rules consist of an implication of the form $X \rightarrow Y$, where X and Y are sets of items co-occurring in a given tuple of the data set ψ [2]. The support s of an itemset X is the percentage of rows in which the itemset X occurs as a subset. The support of the rule $X \rightarrow Y$ is given as $s(X \cup Y)$. The rule $X \rightarrow Y$ is satisfied in ψ with confidence factor $0 \leq c \leq 1$, if at least c% of the instances in ψ that satisfy X also satisfy Y. The confidence factor is given as $s(X \cup Y)/s(X)$.

Spatial association rules (SAR) consist of an implication of the form $X \rightarrow Y$, where X and Y are sets of predicates, and at least one element in X or Y is a spatial predicate [25]. The problem of mining spatial association rules is decomposed in at least three main steps, where the first is usually performed as a data preprocessing method because of the high computational cost:

- Spatial predicate computation: the spatial predicate is a spatial relationship (e.g. distance) between two geographic objects (e.g. closeToRiver);
- Find all frequent predicate sets: a set of predicates is frequent if its support is at least equal to minsup;
- Generate strong association rules: a rule is strong if it reaches minimum support and the confidence is at least equal to the threshold minconf.

Existing spatial association rule mining algorithms are Apriori-like approaches, since the computational cost relies on spatial feature extraction, and not on the candidate generation as in transactional rule mining [41]. Spatial association rule mining algorithms can be classified in two main approaches. The first is based on quantitative reasoning, which mainly computes distance relationships during the frequent set generation. These approaches [22] deal with geographic data (coordinates x,y) directly. Although they have the advantage of not requiring the definition of

a reference object, they have some general drawbacks: usually they deal only with points, consider only quantitative relationships, and they normally do not consider non-spatial attributes of geographic data.

The second approach [25, 30, 4, 38, 43] is based on qualitative reasoning, which usually considers different spatial relationships and features between a reference geographic object type and a set of relevant object types represented by any geometric primitive (e.g. points, lines). Spatial features are normally extracted in a first step, in data *preprocessing* tasks, as explained in section 9.4, while frequent sets are generated in another step.

The main problem in both spatial and non-spatial association rule mining is the generation of huge amounts of rules. Both qualitative and quantitative reasoning approaches have proposed different methods for mining and filtering SAR. [25] presented an approach which exploits taxonomies of both geographic object types and spatial relationships for mining spatial association rules at different granularity levels. Only minimum support is used to prune frequent sets and association rules. In [30] both frequent sets and association rules are pruned a posteriori.

9.5.5 Subgroup Discovery

Subgroup discovery analyses dependencies between a target variable and several explanatory variables. It detects groups of objects that show a significant deviation in their target value with respect to the whole data set. For example, given a discrete target attribute, a subgroup displays an over-proportionally high or low share of a specific target value. More precisely, the quality q of a subgroup h accounts for the difference of target share between the subgroup p and the whole data set p_0, as well as the size n of the subgroup [23]:

$$q(h) = \frac{|p - p_0|}{\sqrt{p_0(1 - p_0)}}\sqrt{n}$$

Subgroups are usually defined by simple conjugation of attribute values, which are then applied to the data set in question. Spatial subgroups are formed if the subgroup definition involves operations on spatial components of the objects. For example, a spatial subgroup could consist of all city districts that are intersected by a river [24]. However, spatial operations are expensive and, due to early pruning it may not be necessary to compute all relations in advance. Klösgen and May [24] developed a spatial subgroup mining system, which integrates spatial feature extraction into the mining process. Spatial joins are performed separately on each search level. Thus, the number of spatial operations can be reduced and redundant storage of features is avoided.

9.6 Example - Frequency Prediction of Inner-City Traffic

Research within the transportation domain as outlined in section 2.6.1 does not only contribute to improved traffic management, but leads also to fruitful applications in other domains. The *Fachverband Außenwerbung* (FAW) is the governing organisation of German outdoor advertisement. Among the development of advertising media and other responsibilities, FAW regulates prices of poster sites. The value of each site is characterised by a quantitative measure, the number of passing pedestrians, vehicles and public transports; and a qualitative measure which specifies the average notice of passers-by. Therefore, in order to calculate poster prices it is vital for FAW to know inner-city traffic frequencies. However, the large number of streets within Germany prohibits empirical measurements for all locations. Within the FAW project, Fraunhofer IAIS developed a method to predict traffic frequencies using spatial data mining [18].

The input data comprises several sources of different quality and resolution. The primary objects of interest are street segments, which generally denote a part of street between two intersections. Each segment possesses a geometry object and has attached information about the type of street, direction, speed class etc. For a small sample of segments one or more frequency measurements are available. In addition, demographic and socio-economic data about the vicinity as well as nearby points of interest (POI) are known. Demographic and socio-economic data usually exist for official districts like post code areas and are directly assigned to all contained streets. In contrast, POI simply mark attractive places like railway stations or restaurants. Clearly, areas with a high density of restaurants will be more frequented than quiet residential areas. In order to utilize POI, the data must first be aggregated. As described in Section 9.4.2 buffers were created around each street segment to calculate the number of relevant POI within the neighbourhood.

In order to infer reliable frequencies for all remaining street segments, a *k*-nearest neighbour algorithm (kNN, see Section 9.5.3) has been applied [32]. It possesses the advantage to incorporate spatial and non-spatial information based on the definition of appropriate distance functions. The frequency of a street segment is calculated as weighted sum of frequencies from the most similar *k* segments in the data sample. The kNN algorithm is known to use extensive resources as the distance between each street segment and available measurement must be calculated. For a city like Frankfurt this amounts to 43 million calculations (about 21,500 segments and 2,000 measurements). While differences in numerical attributes can be determined very fast, the distance between line segments is computationally expensive. Fraunhofer IAIS implemented the algorithm to perform a dynamic and selective calculation of distance from each street segment to the various measurement locations. First, at any time only distances to the top *k* neighbours are stored, replacing them dynamically during the iteration over measurement sites. Second, a step-wise calculation of distance is applied. If the summarised distance of all non-spatial attributes already exceeds the maximum total distance of the current *k* neighbours, the candidate neighbour can be safely discarded and no spatial calculation is necessary. For the city of Frankfurt this integrated approach sped up calculations from nearly one

day to about two hours. In addition, the dynamic calculations reduced the required disc space substantially.

9.7 Roadmap to Knowledge Discovery from Spatio-temporal Data

Spatial data has proved to be a rich source of information about our environment, taken at a fixed moment in time or aggregated over some period of time. However, spatial patterns do not only develop in space, they also extend in (and possibly change over) time. A great challenge therefore lies in the knowledge discovery from spatio-temporal data. In this section we will look at feature extraction, usage of background knowledge and data mining from a spatio-temporal point of view.

9.7.1 Feature Extraction

The main actors for knowledge discovery from spatio-temporal data are the environment and the objects under consideration. The temporal dimension influences both of them by having an environment that changes along time, and in parallel a group of individuals that change their position. Depending on the type of pattern that the analyst investigates, different approaches for feature extraction can be taken according to which entity is evolving during time. The methods and the techniques discussed in the chapter focus their attention on the relations among objects in the space. Note that also various methods for feature extraction from imagery data sets (e.g., satellite images, field bitmaps, etc.) exist. However, due to lack of space we decided not to include these methods in the chapter.

The feature extraction process is mainly based on the exploration of relations among the objects in the data set. But, how are these relations influenced by the temporal dimension? Some relations, lets call them *time invariant*, do not change during time. For example, the *Leaning Tower of Pisa* has a *contained* relation with the *Piazza dei Miracoli*. And this relation will continue to hold for a long time, at least as long as the tower is still leaning. In contrast, the environment can change over time. Consider, for example, a holiday at the sea with the water coming and going during tide and ebb.

When an object moves in time it modifies its relations with the environment. Actually, it changes only its position: the new location determines new relations with the new neighbourhood. For the feature extraction approaches presented so far, an object located at the same position at different time instants has the same relations with the environment. So, changes in the relations are determined by the modified position alone. However, this is a simplification. Consider, for example, the location of an employee in the morning and evening of a working day. Probably,

the employee will travel along the same road from home to work, but the *status* of the same object is different.

It is already challenging to find valid methods to extract meaningful and useful features from geographic data. The temporal dimension adds a new level of difficulty to this task. The example above, where time is used as a pre-condition to determine if a relation holds or not, represents just a starting point of investigation. The role of time is limited to enabling or disabling a certain feature. The real challenge is one step forward: the definition of feature extraction approaches that explore also the temporal dimension. For example, consider the analysis of pollution traffic density. Here, time is *embedded* in the description of the various scenarios: "clouds" of moving areas moving along together, objects moving far away from each other. The evolution of the whole scene depends both on the positions of each object, but also on the evolution of the status of each observed area: its composition, its density, and all the other properties that characterise it during time.

9.7.2 Background Knowledge

Background knowledge comprises valuable information about an object of interest and originates from explicit domain knowledge of some expert or additional data sources. It fulfils several tasks during knowledge discovery, which include feature extraction and data mining. During feature extraction, background knowledge can be used to distinguish interesting and non-interesting relations, and thus to speed up the feature extraction process. In addition, it advances data mining techniques by restricting the hypothesis space. However, the integration of geographic background knowledge is still a field for exploration.

When we add time to the geographic setting, the integration of background knowledge becomes even trickier. It then does not suffice to treat static information, but necessitates the inclusion of dynamic knowledge. For example, attractive points of interest at daylight differ from those at night time. Shopping centres or museums become desolate places after closing hours, while night clubs just start their business at that time. Also, weekly, monthly or long-term fluctuations need to be considered.

9.7.3 Data Mining Algorithms

Geographic references form an inherent part of spatio-temporal data. Therefore, insights gained in geographic data mining should be applied for spatio-temporal data mining. Yet, how can we incorporate time? Given a trajectory of a moving object, a simple approach might flatten time by reducing the trajectory to its pure spatial dimension. Obviously, this results in a great loss of information. Temporal anomalies as traffic jams, locations of interest to a person (home, work, shops) or the means

of transportation (by car, on foot) cannot be inferred without a temporal reference. A second approach might consider a sequence of time slices, where spatial patterns are discovered independently within each time slice and are later on combined. Basically, this approach performs spatial and temporal mining in a sequential order. It is clearly limited as it relies on synchronous observations and cannot exploit space-time dimensions concurrently. Obviously, both approaches are not optimal to make extensive use of spatio-temporal structures. Again, we will need specialized algorithms which will be discussed in detail in Chapter 10.

9.8 Summary

Knowledge discovery from geographic data is not a trivial process and cannot be solved by classical data mining approaches. On the contrary, it requires an understanding of fundamental geographic concepts, sophisticated feature extraction, and specialised algorithms. In this chapter we presented geographic data models and the role of GIS to manage geographic data. We described several methods to detect hidden relationships between geographic objects and reviewed the state-of-the-art of feature extraction. The section on geographic data mining motivates the use of specialised algorithms. It emphasises the need for dynamic feature extraction and the tight integration of spatial databases in the mining process. The various aspects of knowledge discovery are illustrated by an example from the traffic domain. Finally, we pose a number of open research questions when extending geographic knowledge discovery to the dimension of time.

References

1. SPIN! Spatial mining for public data of interest, 2007. http://www.ais.fraunhofer.de/KD/SPIN/.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 487–499. Morgan Kaufmann, 1994.
3. G. Andrienko, D. Malerba, M. May, and M. Teisseire. Mining spatio-temporal data. *Journal of Intelligent Information Systems*, 27(3):187–190, 2006.
4. A. Appice, M. Berardi, M. Ceci, and D. Malerba. Mining and filtering multi-level spatial association rules with ares. In *Proceedings of the 15th International Symposium on the Foundations of Intelligent Systems (ISMIS'05)*, pages 342–353. Springer, 2005.
5. V. Bogorny, S. Camargo, P. Engel, and L. O. Alvares. Mining frequent geographic patterns with knowledge constraints. In *Proceedings of the 14th Annual International Workshop on Geographic Information Systems (GIS'06)*, pages 139–146. ACM, 2006.
6. V. Bogorny, P. Engel, and L. O. Alvares. Enhancing the process of knowledge discovery in geographic databases using geo-ontologies. In H. O. Nigro, S. G. Cizaro, and D. Xodo, editors, *Data Mining with Ontologies: Implementations, Findings and Frameworks*. Idea Group, 2007.
7. P. A. Burrough and R. A. McDonnell. *Principles of Geographical Information Systems*. Oxford University Press, 2000.

8. S. Chawla, S. Shekhar, W. Wu, and U. Ozesmi. Modelling spatial dependencies for mining geospatial data. In H. J. Miller and J. Han, editors, *Geographic Data Mining and Knowledge Discovery*, chapter 6. Taylor & Francis, 2001.
9. J.-P. Chilés and P. Delfiner. *Geostatistics - Modeling Spatial Uncertainty*. Wiley & Sons, 1999.
10. D. J. Cowen. GIS versus CAD versus DBMS: what are the differences? *Journal of Photogrammetric Engineering and Remote Sensing*, 54:1551–1555, 1988.
11. N. A. C. Cressie. *Statistics for Spatial Data*. Wiley & Sons, 1993.
12. M. Egenhofer. Reasoning about binary topological relations. In *Proceedings of the 2nd International Symposium on Advances in Spatial Databases (SSD'91)*, pages 143–160. Springer, 1991.
13. M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. Spatial data mining: Database primitives, algorithms and efficient DBMS support. *Journal of Data Mining and Knowledge Discovery*, 4(2-3):193–216, 2000.
14. M. Ester, J. Sander, H.-P. Kriegel, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231. AAAI Press, 1996.
15. A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression*. Wiley & Sons, 2002.
16. A. S. Fotheringham and P. A. Rogerson. GIS and spatial analytical problems. *International Journal of Geographical Information Systems*, 7(1):3–19, 1993.
17. A. U. Frank. Qualitative spatial reasoning: cardinal directions as an example. *International Journal of Geographical Information Systems*, 10(3):269–290, 1996.
18. Fraunhofer Institut Intelligente Analyse- und Informationssysteme (IAIS). http://www.iais.fraunhofer.de, 2007.
19. R. Haining. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, 2003.
20. J. Han, K. Koperski, and N. Stefanovic. Geominer: a system prototype for spatial data mining. In *Proceedings of the International Conference on Management of Data (SIGMOD'97)*, pages 553–556. ACM, 1997.
21. D. A. Hastings. *Geographic Information Systems: A Tool for Geoscience Analysis and Interpretation*. 1992.
22. Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: A general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12):1472–1485, 2004.
23. W. Klösgen. Subgroup discovery. In W. Klösgen and J. Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3. Oxford University Press, 2002.
24. W. Klösgen and M. May. Spatial subgroup mining integrated in an object-relational spatial database. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, pages 275–286. Springer, 2002.
25. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium on Advances in Spatial Databases (SSD'95)*, pages 47–66. Springer, 1995.
26. R. Laurini and D. Thompson. *Fundamentals of Spatial Information Systems*. Number 37 in APIC Series. Academic Press, 1992.
27. P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind. *Geographic Information Systems and Science*, chapter 3. Wiley & Sons, 2001.
28. D. Malerba, M. Ceci, and A. Appice. Mining model trees from spatial data. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, pages 169–180. Springer, 2005.
29. D. Malerba, F. Esposito, A. Lanza, F. A. Lisi, and A. Appice. Empowering a GIS with inductive learning capabilities: The case of INGENS. *Journal of Computers, Environment and Urban Systems*, 27(3):265–281, 2003.
30. D. Malerba and F. A. Lisi. An ILP method for spatial association rule mining. In *Proceedings of Workshop on Multi-Relational Data Mining (MRDM'01)*, pages 18–29, 2001.

31. M. May and S. Savinov. SPIN! - an enterprise architecture for spatial data mining. In *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'03)*, pages 510–517. Springer, 2003.
32. Michael May. Data mining cup, presentation, 2006. http://www.data-mining-cup.de/2006/Fachkonferenz/Programm/.
33. H. J. Miller. Geographic data mining and knowledge discovery. In J. P. Wilson and A. S. Fotheringham, editors, *Handbook of Geographic Information Science*. Blackwell, 2006.
34. Open GIS Consortium. OpenGIS abstract specification, 1999. http://www.opengeospatial.org/standards/as.
35. T. Ott and F. Swiaczny. *Time-integrative Geographic Information Systems - Management and Analysis of Spatio-Temporal Data*. Springer, 2001.
36. D. Papadias and Y. Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographical Information Science*, 11(2):111–138, 1997.
37. P. Rigaux, M. Scholl, and A. Voisard. *Spatial Databases. With Application to GIS*. Morgan Kaufmann, 2001.
38. S. Rinzivillo and F. Turini. Extracting spatial association rules from spatial transactions. In *Proceedings of the 13th Annual International Workshop on Geographic Information Systems (GIS'05)*, pages 79–86. ACM, 2005.
39. J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Journal of Data Mining and Knowledge Discovery*, 2(2):169–196, 1998.
40. S. Servigne, T. Ubeda, A. Puricelli, and R. Laurini. A methodology for spatial consistency improvement of geographic databases. *Geoinformatica*, 4(1):7–34, 2000.
41. S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2002.
42. W. Tobler. A computer movie simulating urban growth in the detroit region. *Journal of Economic Geography*, 46(2):234–240, 1970.
43. V.Bogorny, J. Valiati, S. Camargo, P. Engel, B. Kuijpers, and L. O. Alvares. Mining maximal generalized frequent geographic patterns with knowledge constraints. In *Proceedings of the 6th International Conference on Data Mining (ICDM'06)*, pages 813–817. IEEE Computer Society, 2006.
44. H. Wackernagel. *Multivariate Geostatistics*. Springer, 1998.
45. Y. Wang and I. Witten. Inducing model trees for continuous classes. In *Proceedings of the 9th European Conference on Machine Learning (ECML'97), Poster Papers*, pages 128–137, 1997.