Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 14-024

Spatial Data Mining

Shashi Shekhar, Michael R. Evans, and James Kang

September 29, 2014

Abstract

Explosive growth in geospatial data and the emergence of new spatial technologies emphasize the need for automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. In this chapter, we explore the emerging field of spatial data mining, focusing on four major topics: prediction and classification, outlier detection, co-location mining, and clustering. We conclude with a look at future research needs.

Spatial Data Mining

Shashi Shekhar, Michael R. Evans, James M. Kang
Department of Computer Science, University of Minnesota

1 Introduction

The explosive growth of spatial data and widespread use of spatial databases [34, 86, 87, 106] have heightened the need for the automated discovery of spatial knowledge. Spatial data mining [96, 86] is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Geospatial-Intelligence Agency [55], the National Cancer Institute [6], and the US Department of Transportation [93]. These organizations are spread across many application domains such as ecology and environmental management [43, 82, 83, 84], public safety, transportation [57], Earth science [40], epidemiology [27], crime analysis [58], and climatology [96, 110].

Spatial database research has been an active area for several decades. The results of this research are being used in a number of areas. To cite a few examples, the filter-and-refine technique used in spatial query processing has been applied to subsequence mining; multidimensional-index structures are used in computer graphics and image processing; and space-filling curves used in spatial query processing and data storage are applied in dimension reduction problems. The value of its contributions no longer in doubt, current research in spatial databases aims to improve its functionality, extensibility, and performance. The impetus for improving functionality comes from the needs of numerous existing application such as Geographic Information Systems, Location Based Services [85], sensor networks [95].

Commercial examples of spatial database management include ESRI's ArcGIS Geodatabase [11], Oracle Spatial [14], IBM's DB2 Spatial Extender and Spatial Datablade, and systems such as Microsoft's SQL Server 2008 [51]. Spatial databases have played a major role in popular applications such as Google Earth [33] and Microsoft's Virtual Earth [67]. Research prototype examples of spatial database management systems include spatial datablades with PostGIS [80], MySQL's Spatial Extensions [70], Sky Server [2] and spatial extensions. The functionalities provided by these systems include use of spatial data types such as points, line-segments and polygons, and spatial operations such as inside, intersection, and distance. Spatial types and operations may be integrated into query languages such as SQL, which allows spatial querying to be combined with object-relational database management systems [20, 97]. The performance enhancement provided by these systems includes a multi-dimensional spatial index and algorithms for spatial database modeling such as OGIS [73] and 3D Topological modeling; spatial query processing including point, regional, range, and nearest neighbor queries; and spatial data methods using a variety of indexes such as quad trees and grid cells.

General purpose data mining tools like Clementine from Statistical Package for the Social Sciences (SPSS), Enterprise Miner from SAS, Data Mining extensions from relational database vendors such as Oracle and IBM, public domain data mining packages such as Weka [29], and See5/C5.0 are designed for the purpose of analyzing transactional data. Although these tools were primarily designed to identify customer-buying patterns in market basket data, they have also been used in analyzing scientific and engineering data, astronomical data, multi-media data, genomic data, and web data. However, extracting interesting and useful patterns from spatial datasets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

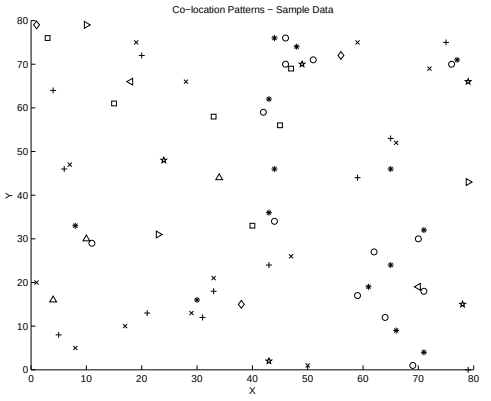


Fig. 1: Illustration of Point Spatial Co-location Patterns. Shapes represent different spatial feature types. Spatial features in sets {'+', '×'} and {'o', '*'} tend to be located together

Specific features of geographical data that preclude the use of general purpose data mining algorithms are: i) the spatial relationships among the variables, ii) the spatial structure of errors, iii) the presence of mixed distributions as opposed to commonly assumed normal distributions, iv) observations that are not independent and identically distributed, v) spatial autocorrelation among the features, and vi) non-linear interactions in feature space. Of course, one can apply conventional data mining algorithms, but these algorithms often perform more poorly on spatial data. A prime example of spatial patterns is co-occurrence patterns, which represent subsets of spatial features whose instances are often located in close geographic proximity, (see Figure 1).

We begin this chapter by describing the characteristics of the data inputs of spatial data mining (Section 2) and by providing an overview of the statistical foundation of spatial data mining (SDM) (Section 3). We then describe in detail four main output patterns of SDM related to anomalies, clustering, co-location and prediction (Section 4). Computational issues regarding these patterns are discussed in Section 5. This chapter concludes with an examination

Tab. 1: Common relationships among non-spatial and spatial data

Non-spatial Relationship	Spatial Relationship
Arithmetic	Set-oriented: union, intersection, membership, ...
Ordering	Topological: meet, within, overlap, ...
Isinstance-of	Directional: North, NE, left, above, behind, ...
Subclass-of	Metric: e.g., distance, area, perimeter, ...
Part-of	Dynamic: update, create, destroy, ...
Membership-of	Shape-based and visibility

of research needs and future directions in Section 6.

2 Input: Spatial Data

The data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons in vector representation and field data in regular or irregular tessellation such as raster data. The data inputs of spatial data mining have two distinct types of attributes: non-spatial attributes and spatial attributes. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used to define the spatial location and extent of spatial objects [15, 32]. The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation defined in a spatial reference frame, as well as shape.

Spatial datasets are discrete representations of continuous phenomena. Discretization of continuous space is necessitated by the nature of digital representation. There are two basic models to represent spatial data, namely, raster (grid) and vector. Satellite images are good examples of raster data. On the other hand, vector data consists of points, lines, polygons and their aggregate (or multi-) counter parts. Spatial networks are another important data type. This distinction is important as many of the techniques that we describe later favor one or more of these data types. Vector data over a space is a framework to formalize specific relationships among a set of objects. Depending on the relationships of interest, the space can be modeled many different ways, i.e., as set-based space, topological space, Euclidean space, metric space and network space [107].

Set-based space uses the basic notion of elements, element-equality, sets, and membership to formalize set relationships such as set-equality, subset, union, cardinality, relation, function, and convexity. Relational and object-relational databases use this model of space.

Topological space uses the basic notion of a neighborhood and points to formalize extended object relations such as boundary, interior, open, closed, within, connected, and overlaps, which are invariant under elastic deformation. Combinatorial topological space formalizes relationships such as Euler's formula (number of faces + number of vertices - number of edges = 2 for planar configuration). Network space is a form of topological space in which the connectivity property among nodes formalizes graph properties such as connectivity, isomorphism, shortest-path, and planarity.

Euclidean coordinatized space uses the notion of a coordinate system to transform spatial properties and relationships into properties of tuples of real numbers. Metric space formalizes distance relationships using positive symmetric functions that obey the triangle inequality. Many multidimensional applications use Euclidean coordinatized space with metrics such as distance.

Widely used gazetteers employ spatial referencing with identifiers of a location that can be transformed into coordinates, such as a postal code (street addresses) or geo-name which is more natural to human understanding. Time is usually included in the spatial data as a time stamp.

During data input, relationships among non-spatial objects are made explicit through arithmetic relation, ordering, instance-of, subclass-of, and membership-of. In contrast, relationships among spatial objects are often implicit, such as overlap, intersect, and behind. Table 1 gives examples of spatial and non-spatial relationships. One possible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques such as those described in [81, 103, 4, 5, 44]. However, the materialization can result in loss of information. Usually, spatial and temporal vagueness, which naturally exists in data and relationships, creates further modeling and processing difficulty in spatial data mining. Another way to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the spatial data mining process.

3 Statistical Foundations of Spatial Data Mining

(a) Spatial Framework (b) Neighbor relationship (c) Contiguity Matrix

Fig. 2: A spatial framework and its four-neighborhood contiguity matrix.

The specialty of spatial data mining originates from three central concepts in spatial statistics: spatial autocorrelation, and spatial non-stationarity [28, 23]. Spatial statistics is a branch of statistics concerned with the analysis and modeling of spatial data [12]. The field classifies spatial data into three basic types for ease of interpretation: (a) point referenced data , which is modeled as a fixed collection of spatial locations, S, in a two-dimensional framework D (e.g. set of police stations in a metropolitan city); (b) areal data, modeled as a finite set of irregular shaped polygons in a two-dimensional framework D (e.g. set of police districts in a metropolitan city); and (c) point process data which is modeled as a random collection of spatial events, collectively referred to as the spatial point pattern over a two-dimensional framework D (e.g. home locations of patients infected by a disease).

Statistical models [22] are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on probability theory. Spatial data can be thought of as resulting from observations on the stochastic process $Z(s) : s \in D$, where s is a spatial location and D is possibly a random set in a spatial framework. Here we present three types of spatial statistical problems one might encounter: point process, lattice, and geostatistics.

Point process: A point process is a model for the spatial distribution of the points in a point pattern. Several natural processes can be modeled as spatial point patterns, e.g., positions of trees in a forest and locations of bird habitats in a wetland. Spatial point patterns can be broadly grouped into random or non-random processes. Real point patterns are often compared with a random pattern (generated by a Poisson process) using the average distance between a point and its nearest neighbor.

Lattice: A lattice is a model for a gridded space in a spatial framework. Here the lattice refers to a countable collection of regular or irregular spatial sites related to each other via a neighborhood relationship. Several spatial statistical analysis, e.g., the spatial autoregressive model and Markov random fields, can be applied on lattice data.

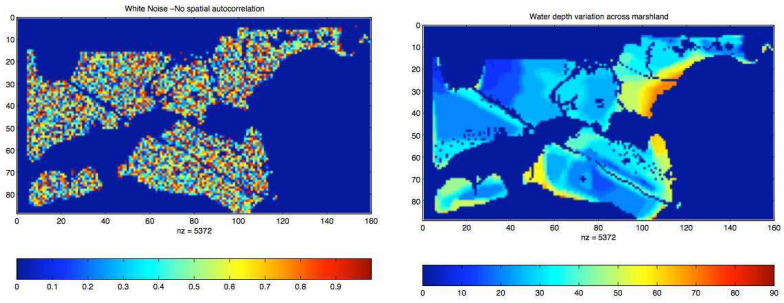
Geostatistics: Geostatistics deals with the analysis of spatial continuity and weak stationarity [22], which is an inherent characteristics of spatial datasets. Geostatistics provides a set of statistics tools, such as kriging, to the interpolation of attributes at unsampled locations.

The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent a neighborhood relationship defined using adjacency or Euclidean distances. Example definitions of a neighborhood using adjacency include a four-neighborhood and an eight-neighborhood.

Figure 2(a) shows a gridded spatial framework with four locations, A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 2(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 2(c). Other contiguity matrices can be designed to model neighborhood relationship based on distance. The essential idea is to specify the pairs of locations that influence each other along with the relative intensity of interaction. More general models of spatial relationships using cliques and hypergraphs are available in the literature [105]. In spatial statistics, correlations among objects (spatial autocorrelation) is quantified using measures such as Ripley's K-function and Moran's I [22].

Spatial Autocorrelation: One of the fundamental assumptions of statistical analysis is that the data samples are independently generated: like successive tosses of coin, or the rolling of a die. However, in the analysis of spatial data, the assumption about the independence of samples is generally false. In fact, spatial data tends to be highly self correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife, and temperature vary gradually over space. The property of like things clustering in space is so fundamental that geographers have elevated it to the status of the first law of geography: "Everything is related to everything else, but nearby things are more related than distant things" [100]. For example, Figure 3 shows the value distributions of an attribute in a spatial framework for an independent identical distribution and a distribution with spatial autocorrelation.

Spatial Non-Stationarity: Spatial Non-Stationarity refers to the inherent variation in measurements of a relation-



(a) Attribute with an Independent Identical Distribution

(b) Attribute with Spatial Autocorrelation

Fig. 3: Attribute values in space with independent identical distribution and spatial autocorrelation

ships over space. In fact, spatial context influences the nature of spatial relationships. For example, human behavior can vary intrinsically over space (e.g., differing cultures). Different jurisdictions tend to produce different laws (e.g., speed limit differences between Minnesota and Wisconsin).

Spatial statistics has explored measures such as Ripley's K Function, Spatial Scan Statistic, Morans I, Local Moran Index, Getis Ord, Gearys C, etc, to quantify spatial correlation. These statistics have found many applications in common spatial data mining tasks including spatial co-location, spatial outlier detection and hotspot discovery.

4 Output: Pattern Families

In this section, we present case studies of four important output patterns for spatial data mining: spatial outliers, co-location patterns, classification and regression models, and spatial clustering.

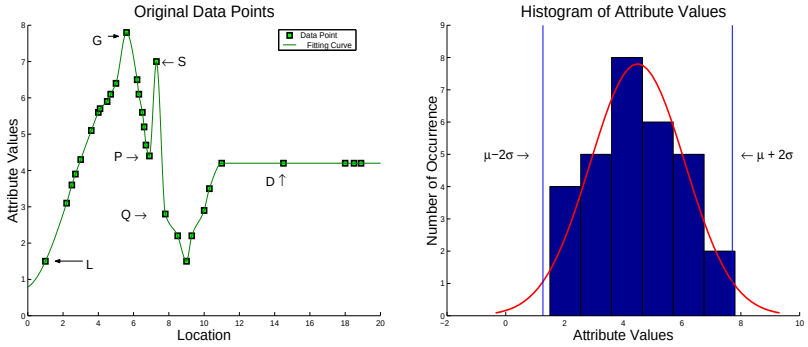
4.1 Spatial Outlier Detection

Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [13], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism [38]. The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as detection of credit card fraud and voting irregularities. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods [109, 99, 79]. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases. These application domains include transportation, ecology, homeland security, public health, climatology, and location-based services.

A spatial outlier [89] is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age.

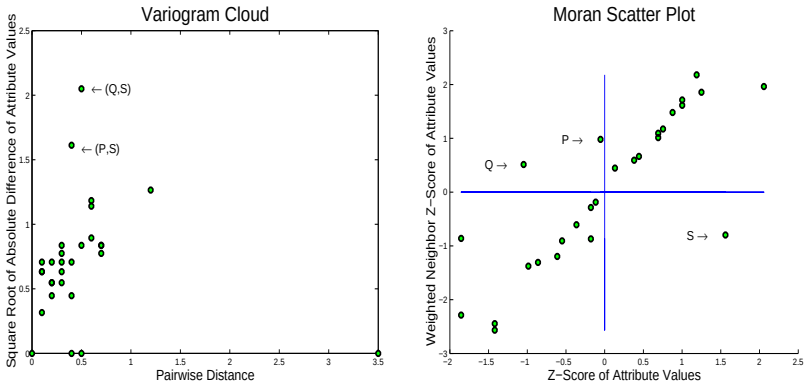
Illustrative Examples and Application Domains: We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 4(a), the X-axis is the location of data points in one-dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point and fit the distribution model to the values of the non-spatial attribute. As shown in Figure 4(b), the outlier detected using this approach is the data point G, which has an extremely high attribute value 7.9, exceeding the threshold of $\mu + 2\sigma = 4.49 + 2 * 1.61 = 7.71$. This test assumes a normal distribution for attribute values. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.

Common Methods: Tests to detect spatial outliers separate spatial attributes from non-spatial attributes. Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Spatial statistics literature provides two kinds of bi-partite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests, which are based on the visualization of spatial data, highlight spatial outliers. Example methods include variogram clouds [37] and Moran scatterplots [65, 22]. A variogram cloud displays data points related by neighborhood relationships. Figure 5(a) shows a variogram cloud for the example dataset shown in Figure 4(a). This plot shows that two pairs (P, S) and (Q, S) on



(a) An Example Dataset (b) Histogram

Fig. 4: A Dataset for Outlier Detection.



(a) Variogram cloud (b) Moran scatterplot

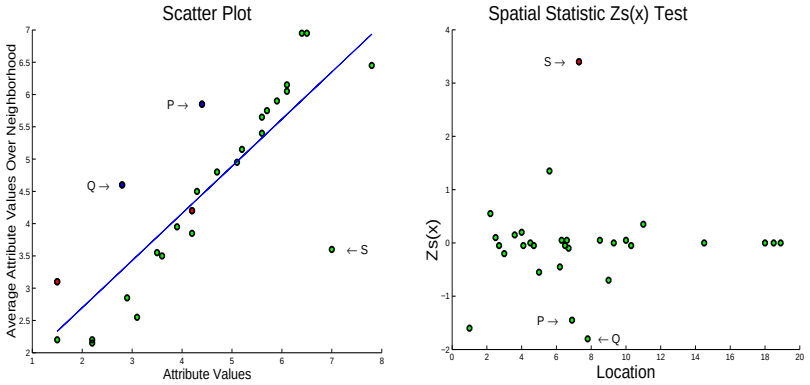
Fig. 5: Variogram cloud and moran scatterplot to detect spatial outliers.

the left hand side lie above the main group of pairs and are possibly related to spatial outliers. A Moran scatterplot shows spatial association or disassociation of spatial close objects. The upper left and lower right quadrants of Figure 5(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values surrounded by low values (e.g,. point S). Figure 5(b) indicates a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values surrounded by low values (e.g,. point S).

A scatterplot [64] shows attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. A least square regression line is used to identify spatial outliers. A scatter sloping upward to the right indicates a positive spatial autocorrelation (adjacent values tend to be similar); a scatter sloping upward to the left indicates a negative spatial autocorrelation. The residual is defined as the vertical distance (Y-axis) between a point P with location (X_p, Y_p) to the regression line $Y = mX + b$, that is, residual $\epsilon = Y_p - (mX_p + b)$. Cases with standardized residuals, $\epsilon_{standard} = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}$, greater than 3.0 or less than -3.0 are flagged as possible spatial outliers, where μ_ϵ and σ_ϵ are the mean and standard deviation of the distribution of the error term ϵ. In Figure 6(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the dataset in Figure 4(a). The point S turns out to be the farthest from the regression line and may be identified as a spatial outlier.

Spatial statistic $S(x)$ is normally distributed if the attribute value $f(x)$ is normally distributed. A popular test for detecting spatial outliers for normally distributed $f(x)$ can be described as follows: Spatial statistic $Z_{s(x)} = |\frac{S(x) - \mu_s}{\sigma_s}| > \theta$. For each location x with an attribute value $f(x)$, the $S(x)$ is the difference between the attribute value at location x and the average attribute value of $x's$ neighbors, μ_s is the mean value of $S(x)$, and σ_s is the value of the standard deviation of $S(x)$ over all stations. The choice of θ depends on a specified confidence level. For example, a confidence level of 95 percent will lead to $\theta \approx 2$.

Figure 6(b) shows the visualization of the spatial statistic method described above. The X-axis is the location of data points in one-dimensional space; the Y-axis is the value of spatial statistic $Z_{s(x)}$ for each data point. We can



(a) Scatterplot (b) Spatial statistic $Z_{s(x)}$

Fig. 6: Scatterplot and Spatial Statistic $Z_{s(x)}$ to detect spatial outliers.

easily observe that point S has a $Z_{s(x)}$ value exceeding 3, and will be detected as a spatial outlier. Note that the two neighboring points P and Q of S have $Z_{s(x)}$ values close to -2 due to the presence of spatial outliers in their neighborhoods.

The techniques presented above are based on single attribute. However, multi-attribute based spatial outlier detection is also possible. For example, average and median attribute value based algorithms are presented in [63].

4.2 Co-location Patterns

Co-location patterns represent subsets of boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species and crime attractors (e.g. Bars, misdemeanors, etc). Boolean spatial features describe the presence or absence of geographic object types at different locations in a two-dimensional or three-dimensional metric space, e.g., the surface of the Earth. Examples of boolean spatial features include plant species, and crime.

Spatial Co-location: Co-location rules are models to infer the presence of boolean spatial features in the neighborhood of instances of other boolean spatial features. For example, "Nile Crocodiles → Egyptian Plover" predicts the presence of Egyptian Plover birds in areas with Nile Crocodiles. Figure 1 shows a dataset consisting of instances of several boolean spatial features, each represented by a distinct shape. A careful review reveals two co-location patterns, i.e. ('+','×') and ('o','*').

Co-location rule discovery is a process to identify co-location patterns from large spatial datasets with a large number of boolean features. The spatial co-location rule discovery problem looks similar to, but, in fact, is very different from the association rule mining problem [5] because of the lack of transactions. In market basket datasets, transactions represent sets of item types bought together by customers. The support of an association is defined to be the fraction of transactions containing the association. Association rules are derived from all the associations with support values larger than a user given threshold. In the spatial co-location rule mining problem, transactions are often not explicit. The transactions in market basket analysis are independent of each other. Transactions are disjoint in the sense of not sharing instances of item types. In contrast, the instances of Boolean spatial features are embedded in a continuous space and share a variety of spatial relationships (e.g. neighbor) with each other.

Common Methods: Spatial co-location rule mining approaches can be grouped into two broad categories: approaches that use spatial statistics and algorithms that use association rule mining kind of primitives. Spatial statistics based approaches utilize statistical measures such as cross-K function, mean nearest-neighbor distance, and spatial autocorrelation. However, these approaches are computationally expensive. Association rule-based approaches focus on the creation of transactions over space so that an *apriori* like algorithm [5] can be used. Transactions in space can use a reference-feature centric [53] approach or a data-partition [69] approach. Figure 7(c) shows two possible partitions for the dataset of Figure 7(a), along with the supports for co-location (A, B). The reference feature centric model is based on the choice of a reference spatial feature [53] and is relevant to application domains focusing on a specific boolean spatial feature, e.g. cancer. The event centric model [88] finds subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types (see Figure 7a).

Illustrative Examples and Application Domains: Domain scientists are interested in finding the co-locations of other task relevant features (e.g. asbestos) to the reference feature. For example, consider the spatial dataset in Figure 7(a) with three feature types, A, B and C. Each feature type has two instances. The neighbor relationships

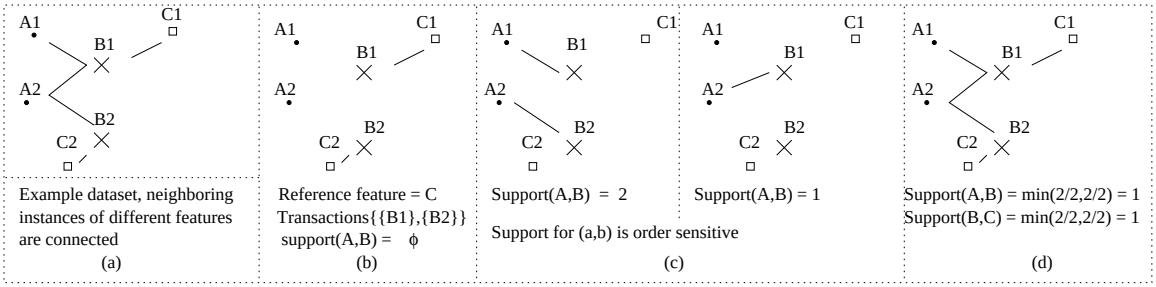


Fig. 7: Example to illustrate different approaches to discovering co-location patterns a) Example dataset. b) Data partition approach. Support measure is ill-defined and order sensitive c) Reference feature centric model d) Event centric model

between instances are shown as edges. Co-locations (A, B) and (B, C) may be considered as frequent in this example. Figure 7(b) shows transactions created by choosing C as the reference feature. Co-location (A, B) will not be found since it does not involve the reference feature.

4.3 Regression and Classification

Classification and regression are similar types of patterns in data mining. Given a sample set of input-output pairs, the objective of supervised learning is to find a function that learns from the given input-output pairs, and predicts an output for any unseen input (but assumed to be generated from the same distribution), such that the predicted output is as close as possible to the desired output. For example, in remote sensing image classification, the input attribute space consists of various spectral bands or channels (e.g., blue, green, red, infra-red, thermal, etc.) The input vectors (x_i's) are reflectance values at the i^{th} location in the image; and the outputs (y_i's) are thematic classes such as forest, urban, water, and agriculture. Depending on the type of output attribute, two supervised learning tasks can be distinguished:

The fact that classical data mining techniques ignore spatial autocorrelation and spatial heterogeneity in the model-building process is one reason why these techniques do a poor job. A second, more subtle but equally important reason is related to the choice of the objective function to measure classification accuracy. For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. However, this measure may not be the most suitable in a spatial context. *Spatial accuracy*−how far the predictions are from the actuals−is as important in this application domain due to the effects of the discretization of a continuous wetland into discrete pixels, as shown in Figure 8. Figure 8(a) shows the actual locations of nests and 8(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled 'A' and are quite close to other blank pixels, which represent 'no-nest'. Now consider two predictions shown in Figure 8(c) and 8(d). Domain scientists prefer prediction 8(d) over 8(c), since the predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 8(c) and 8(d), and a measure of spatial accuracy is needed to capture this preference.

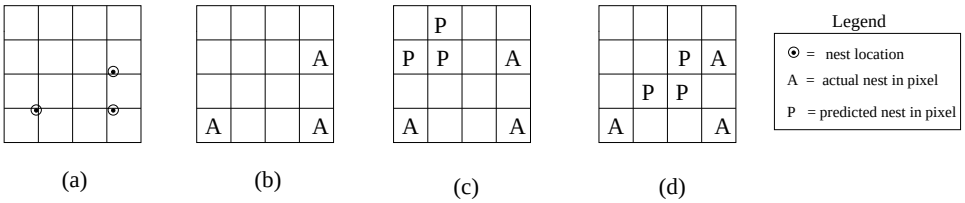


Fig. 8: (a)The actual locations of nests, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another model. Prediction(d) is spatially more accurate than (c).

- **Classification:** Here, the input vectors x_i are assigned to a few discrete numbers of classes, for example, image classification [24] y_i.
- **Regression:** In regression, also known as function approximation or prediction, the input-output pairs are generated from an unknown function of the form $y = f(x)$, where y is continuous. Typically regression is used

in regression and estimation, for example, crop yield prediction [62], daily temperature prediction, and market share estimation for a particular product. Regression can also be used in inverse estimation, that is, given that we have an observed value of y, we want to determine the corresponding x value.

Illustrative Examples and Application Domains: The prediction of events occurring at particular geographic locations is very important in several application domains such as crime analysis, cellular networks, and natural disasters. However, prediction and regression with spatial data requires taking into account the varying relationships among variables, such as autocorrelation.

Common Methods: Several previous studies [47], [94] have shown that the modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. An example spatial framework and its four-neighborhood contiguity matrix is shown in Figure 2. In this section we present two spatial data mining techniques, namely the Logistic Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

Logistic Spatial Autoregressive Model(SAR): Logistic SAR decomposes a classifier $\hat{f}_C$ into two parts, namely spatial autoregression and logistic transformation. Spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation [8]. If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{1}$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable via the logistic function for binary dependent variables.

Markov Random Field-based Bayesian Classifiers: Maximum likelihood classification (MLC) is one of the most widely used parametric and supervised classification technique in the field of remote sensing [39, 98]. However, MLC is a per-pixel based classifier and assumes that samples are independent and identically distributed (i.i.d). Ignoring spatial autocorrelation results in *salt and pepper* kind of noise in the classified images. One solution is to use Markov Random Field (MRF)-based Bayesian classifiers [61] to model spatial context via the *a priori* term in Bayes' rule. This uses a set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix). A more detailed theoretical and experimental comparison of these two methods can be found in [91].

4.4 Spatial Clustering

Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters.

In spatial statistics, the standard against which spatial point patterns are often compared is a completely spatially point process, and departures indicate that the pattern is not completely spatially random. Complete spatial randomness (CSR) [22] is synonymous with a homogeneous Poisson process. The patterns of the process are independently and uniformly distributed over space, i.e., the patterns are equally likely to occur anywhere and do not interact with each other. In contrast, a clustered pattern is distributed dependently and attractively in space.

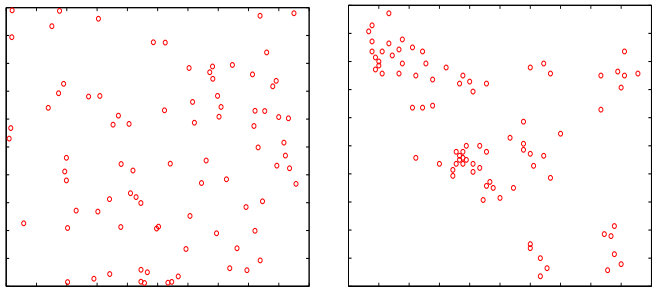
An illustration of complete spatial random patterns and clustered patterns is given in Figure 9, which shows realizations from a completely spatially random process and from a spatial cluster process respectively (each conditioned to have 85 points in a unit square).

Illustrative Examples and Application Domains: Cluster analysis is used in many spatial and spatiotemporal application domains such as in remote sensing data analysis as a first step to determine the number and distribution of spectral classes, in epidemiology for finding unusual groups of health-related events, and in detection of crime hot spots by police officers.

Notice from Figure 9 (a) that the complete spatial randomness pattern seems to exhibit some clustering. This is not an unrepresentative realization, but illustrates a well known property of homogeneous Poisson processes: event-to-nearest-event distances are proportional to χ^2_2 random variables, whose densities have a substantial amount of probability near zero [22]. True clustering, by contrast, is shown in Figure 9 (b).

Common Methods: After verification of the statistical significance of spatial clustering, clustering algorithms are used to discover clusters of interest. Because of the multitude of clustering algorithms that have been developed, it is useful to categorize them into groups.

1. *Hierarchical* clustering methods start with all patterns as a single cluster and successively perform splitting or merging until a stopping criterion is met. This results in a tree of clusters, called *dendograms*. The dendogram can be cut at different levels to yield desired clusters. Hierarchical algorithms can further be divided into *agglomerative* and *divisive* methods. The hierarchical clustering algorithms include balanced



(a) CSR Pattern (b) Clustered Pattern

Fig. 9: Complete Spatial Random (CSR) and Spatially Clustered Patterns

iterative reducing and clustering using hierarchies (Birch), clustering using inter-connectivity (Chameleon), clustering using representatives (Cure), and robust clustering using links (Rock). More discussion of these methods can be found in [19, 111, 50].

2. *Partitional* clustering algorithms start with each pattern as a single cluster and iteratively reallocate data points to each cluster until a stopping criterion is met. These methods tend to find clusters of spherical shape. *K-Means* and *K-Medoids* are commonly used partitional algorithms. Squared error is the most frequently used criterion function in partitional clustering. The recent algorithms in this category include partitioning around medoids (Pam), clustering large applications (Clara), clustering large applications based on randomized search (Clarans), and expectation-maximization (EM). Related papers include [41, 72].
3. *Density-based* clustering algorithms try to find clusters based on the density of data points in a region. These algorithms treat clusters as dense regions of objects in the data space. The density-based clustering algorithms include density-based spatial clustering of applications with noise (Dbscan), ordering points to identify clustering structure (Optics), and density based clustering (Decode). Related research is discussed in [56, 66, 78, 71, 3]

More details on various clustering methods can also be found in a recent survey paper [36]. Many of the clustering algorithms discussed here do not take into account spatial autocorrelation and spatial constraints. However, algorithms for spatial clustering in the presence of obstacles have been proposed in [101, 112]. These approaches show improved clustering results and stress the importance of modeling neighborhood relationships in clustering.

4.5 Other Pattern Families: Hotspots

Hotspots are a special kind of clustered pattern. Like general clustered patterns, objects in hotspot regions have high similarity in comparison to one another, and are quite dissimilar to all the objects outside the hotspot. One important feature that distinguishes a hotspot from a general cluster is that the objects in the hotspot area are more active compared to all others (density, appearance, etc). Hotspot discovery/detection in spatial data mining is a process of identifying spatial regions where more events are likely to happen, or more objects are likely to appear, in comparison to other areas.

Hotspot detection is mainly used in the analysis of crime and disease data. Crime data analysis [102] aims at finding areas that have greater than average numbers of criminal or disorderly events, or areas where people have a higher than average risk of victimization. In cancer/disease data analysis, hotspots of locations where disease are reported intensively are detected, which may indicate a potential breakout of this disease, or suggest an underlying cause of the disease. Other domains of application include: transportation (to identify unusual rates of accidents along highways), and ecological science (to conduct geoinformatic surveillance for geospatial hot-spot detection [46]).

The pattern and shape of spatial hotspot varies. For example, in the crime hotspot detection, the results may be given in the form of crime hotspot streets, hotspot areas or hotspot cities (see Figure 10).

Common Methods: Hotspot detection methods fall in three main catagories based on the types of hotspot they are looking for.:

Global Hotspot Detection: The K-means algorithm creates k (user-defined number) clusters by partitioning the crime point data into groups. The process finds the best positioning of the K centers and then assigns each point to the center that is nearest.

Hierarchical Hotspot Detection: The Nearest Neighbor Hierarchical (NNH) Clustering [45] algorithm is based on a nearest neighbor analysis technique. For example, crime incident locations are first grouped into nearest neighbor clusters containing a minimum number of point locations specified by the user and these first-order clusters are further

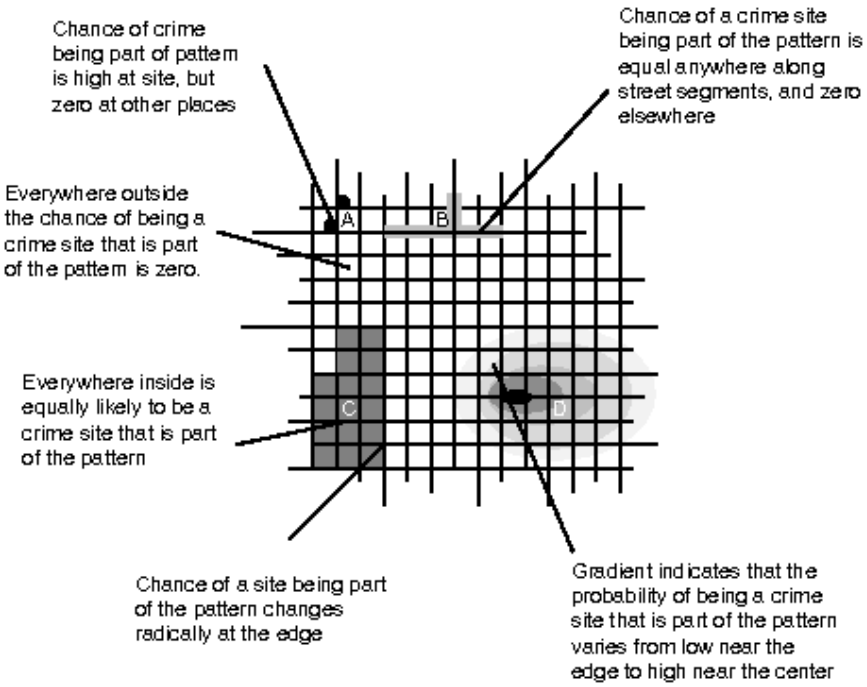


Fig. 10: The hotspots region may be location spots, street segments or block areas.

grouped into larger, second-order clusters, with the process continuing until no more clustering is possible and its variation.

Risk-adjusted Nearest Neighbor Hierarchical Clustering (RNNH) [59] is a variation of NNH. RNNH function is to find the concentration of incidents relative to a baseline. It achieves this goal by adjusting the threshold distance dynamically. The background data is represented as a fine grid using kernel density estimation, and this is used to adjust the threshold distance for clustering the original point set, on a cell-by-cell basis. It should be noted that with both NNh and RNNh, not all events are assigned to clusters, and each point is assigned to either one cluster at a given hierarchical level or none at all.

STAC: The STAC Hot Spot Area function [59] in CrimeStat searches for and identifies the densest clusters of incidents based on the scatter of points on the map. The STAC Hot Spot Area routine creates a set of real units from point data . First it identifies the major concentrations of points for a given distribution and it then represents each dense area by the STAC is a scan-type clustering algorithm in which a circle is repeatedly laid over a grid and the number of points within the circle is counted.

Local Hotspot Detection: LISA: Local Indicators of Spatial Association [10] statistics assess the local association between data by comparing local averages to global averages. For this reason they are useful in adding definition to crime hot spots and placing a spatial limit on those areas of highest crime event concentration

5 Computational Issues

The volume of data, the complexity of spatial data types and relationships, and the need to identify spatial autocorrelation poses numerous computational challenges to the spatial data mining field. When designing spatial data mining algorithms one has to take into account several considerations, such as, space partitioning, predicate approximation, multidimensional data structures, etc. Table 2 summarizes how these requirements compare with classical data mining. Computational issues may arise due to dimensionality, spatial join process required in co-location mining and spatial outlier detection, estimation of SAR model parameters in the presence of large neighborhood matrix W, etc.

To illustrate these computational challenge, we use the example of a case study with parameter estimation for the SAR model. The massive sizes of geospatial datasets in many application domains make it important to develop scalable parameter estimation algorithms of the SAR model solutions for location prediction and classification. As noted previously, many classical data mining algorithms, such as linear regression, assume that the learning samples are *independently and identically distributed (i.i.d.)*. This assumption is violated in the case of spatial data due

Classical Algorithms	Algorithmic Strategies for SDM
Divide-and-Conquer	Space partitioning
Filter-and-Refine	Minimum-Bounding Rectangle (MBR),
	Predicate Approximation
Ordering	Plane Sweeping, Space Filling Curve
Hierarchical Structures	Spatial Index, Tree Matching
Parameter Estimation	Parameter estimation with spatial autocorrelation

Tab. 2: Algorithmic Strategies for SDM

to spatial autocorrelation [9] and in such cases classical linear regression yields a weak model with not only low prediction accuracy [21, 90] but also residual error exhibiting spatial dependence. Modeling spatial dependencies improves overall classification and prediction accuracies significantly.

However, estimation of SAR model parameters is computationally very expensive because of the need to compute the determinant of a large matrix in the likelihood function [60, 74, 75, 76, 52]. The Maximum Likelihood function for SAR parameter estimation contains two terms, a determinant term and an *SSE* term (Equation 2). The former involves computation of the determinant of a very large matrix, which is a well-known hard problem in numerical analysis. To estimate the parameters of a ML-based SAR model solution, the log-likelihood function can be constructed, as shown in (2). The estimation procedure involves computation of the logarithm of the determinant of (log-det) a large matrix, i.e. $(\mathbf{I} - \rho\mathbf{W})$.

$$\ell(\rho|\mathbf{y}) = \frac{-2}{n}\underbrace{\ln|\mathbf{I}-\rho\mathbf{W}|}_{log-det} + \underbrace{\ln((\mathbf{I}-\rho\mathbf{W})\mathbf{y})^T(\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\ \mathbf{x}^T)^T(\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\ \mathbf{x}^T)((\mathbf{I}-\rho\mathbf{W})\mathbf{y})}_{SSE} \quad (2)$$

For example, the exact SAR model parameter estimation for a very small 10,000-point spatial problem can take tens of minutes on common desktop computers. Computation costs make it difficult to use SAR for important spatial problems which involve millions of points, despite its promise to improve prediction and classification accuracy.

In the equation, $\mathbf{y}$ is the n-by-1 vector of observations on the dependent variable, where n is the number of observation points; ρ is the spatial autoregression parameter; $\mathbf{W}$ is the n-by-n neighborhood matrix that accounts for the spatial relationships (dependencies) among the spatial data; $\mathbf{x}$ is the n-by-k matrix of observations on the explanatory variable, where k is the number of features; and β is a k-by-1 vector of regression coefficients. Spatial autocorrelation term $\rho\mathbf{W}\mathbf{y}$ is added to the linear regression model in order to model the strength of the spatial dependencies among the elements of the dependent variable, $\mathbf{y}$.

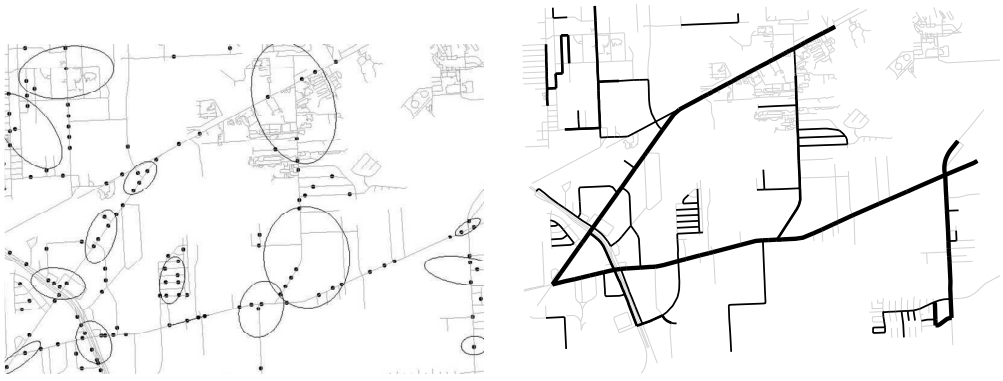
6 Future Directions and Research Needs

This section presents the future directions and research needs in spatial data mining. There are several new areas of research, but the two we will focus on are network-based data mining and spatio-temporal data mining.

6.1 Network Patterns

One of the main challenges in spatial data mining is to account for the network structure in the dataset. For example, in anomaly detection, spatial techniques do not consider the spatial network structure of the dataset, that is, they may not be able to model graph properties such as one-ways, connectivities, left-turns, etc. In this section, we present "Mean Streets", an interesting spatial data mining problem that has a spatial network as part of its input.

Mean Streets. The problem of identifying *Mean Streets* is to discover those connected subsets of a spatial network whose attribute values are significantly higher than expected (Figure 11b). Finding *mean streets* is particularly important for crime analysis (high-crime-density street discovery) and police work (planning effective and efficient patrolling strategies). In urban areas, many human activities are centered about spatio-temporal (ST) infrastructure networks, such as roads and highways, oil/gas pipelines, and utilities (e.g., water, electricity, telephone). Thus, activity reports such as crime logs may often use network based location references (e.g., street addresses). In addition, spatial interaction among activities at nearby locations may be constrained by network connectivity and network



(a) Input and *k-means clustering using Crimestat*

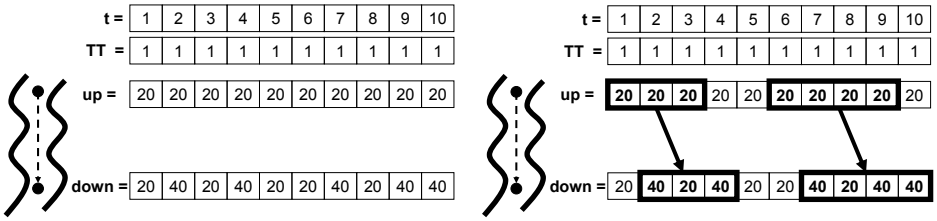
(b) Output using *Mean Streets*

Fig. 11: Crimestat results vs. *Mean Streets* identified in a US city

distances (e.g., shortest paths along roads or train networks) rather than the geometric distances used in traditional spatial analysis. Crime prevention may focus on identifying subsets of ST networks with high activity levels, understanding underlying causes in terms of network properties, and designing network control policies. Identifying and quantifying *mean streets* is a challenging task due to the need to choose the correct statistical model. In addition, the discovery process in large spatial networks is computationally very expensive due to the difficulty of characterizing and enumerating the population of streets to define a normal or expected activity level. Preliminary exploration of descriptive and explanatory models for ST network patterns in [16]. However, further challenges and research is needed to identify other interesting patterns within network datasets, such as partial segments of roads that are more interesting than other parts.

6.2 Spatio-temporal Data Mining

Spatio-temporal data is often modeled using events and processes, both of which generally represent change of some kind. Processes refer to ongoing phenomena that represent activities of one or more types without a specified endpoint [92, 7, 108]. Events refer to individual occurrences of a process with a specified beginning and end. Event-types and event-instances are distinguished. For example, a hurricane event-type may occur at many different locations and times e.g.,Katrina(New Orleans, 2005) and Rita(Houston, 2005). Each event-instance is associated with a particular occurrence time and location. The ordering may be total if event-instances have disjoint occurrence times. Otherwise, ordering is based on spatio-temporal semantics such as partial order, and spatio-temporal patterns can be modeled as partially ordered subsets. These unique characteristics create new and interesting challenges to discover spatio-temporal patterns. For example, in contrast to spatial outliers, a spatio-temporal outlier is a spatio-temporal object whose thematic (non-spatial and non-temporal) attributes are significantly different from those of other objects in its spatial and temporal neighborhoods. A spatio-temporal object is defined as a time-evolving spatial object whose evolution or history is represented by a set of instances (EQ), where the space stamp is the location of the object o id at timestamp t. In the remainder of this section, we present research trends in various areas of spatio-temporal data mining.



(a) Example Input

(b) Example Output

Fig. 12: Flow Anomaly Example

Flow Anomalies: Given a percentage-threshold and a set of observations across multiple spatial locations, flow anomaly discovery aims to identify dominant time intervals where the fraction of time instants of significantly mis-

matched sensor readings exceed the given percentage-threshold. Figure 12 gives a simple example of flow anomalies (FA). In figure 12(a), the input to the flow anomaly problem consists of two spatial locations (i.e., an upstream (up) and downstream (down) sensor), 10 time instants, and the notion of travel time (TT) or flow between the locations. For simplicity, the TT is set to a constant of 1, but it can be variable. The output contains two FAs; using the time instants at the upstream sensor, periods 1-3 and 6-9, where the majority of time-points show significant differences in-between (Figure 12b). Discovering flow anomalies is an important problem in several applications such as environmental systems, transportation networks, and video surveillance systems. However, mining flow anomalies is computationally expensive due to the large (potentially infinite) number of time instants across a spatial network of locations. Traditional outlier detection methods (e.g. t-test) are suited for detecting transient FAs (i.e., time instants of significant mis-matches across consecutive sensors) but cannot detect persistent FAs (i.e., long variable time-windows with a high fraction of time instant transient FAs) due to a lack of a predetermined window size. Spatial outlier detection techniques do not consider the flow (i.e., travel time) between spatial locations and cannot detect any type of flow anomalies. Preliminary analysis introduced a time-scalable technique called SWEET (Smart Window Enumeration and Evaluation of persistent-Thresholds) that utilizes several algebraic properties in the flow anomaly problem to discover these patterns efficiently [49, 30, 26]. However, further research is needed to discover other types of patterns within this environment.

Teleconnected Flow Anomalies An additional pattern that utilizes flow-anomalies is teleconnected patterns [48]. A teleconnection represents a strong interaction between paired events that are spatially distant from each other. Identifying teleconnected flow events is computationally hard due to the large number of time instants of measurement, sensors, and locations. For example, a well-known teleconnected event pair involves the warming of the eastern pacific region (i.e., El Nino) and unusual weather patterns throughout the world [77]. Recently, a RAD (Relationship Analysis of Dynamic-neighborhoods) technique has been proposed that models flow networks to identify teleconnected events [48]. Further research is needed to explore new and interesting patterns that may lie within the RAD model.

Mixed Drove Co-Occurrence Patterns Another type of dynamic behavior of spatial datasets which might affect colocation patterns is changing the specification of zone of interest and measure values according to user preferences. Mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) represent subsets of two or more different object-types whose instances are often located in spatial and temporal proximity. Discovering MDCOPs is potentially useful in identifying tactics in battlefields and games, understanding predator-prey interactions, and in transportation (road and network) planning [35, 54]. However, mining MDCOPs is computationally very expensive because the interest measures are computationally complex, datasets are larger due to the archival history, and the set of candidate patterns is exponential in the number of object-types. Preliminary work has produced a monotonic composite interest measure for discovering MDCOPs and novel MDCOP mining algorithms are presented in [18].

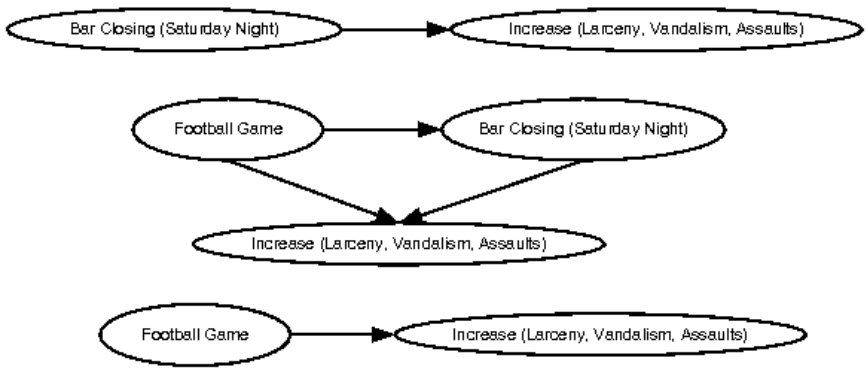


Fig. 13: Cascading spatio-temporal patterns from Public Safety

Cascading spatio-temporal patterns Partially ordered subsets of event-types whose instances are located together and occur in stages are called cascading spatio-temporal patterns (CSTP). Figure 13 shows some interesting partially-ordered patterns that were discovered from real spatio-temporal crime datasets from the city of Lincoln, Nebraska [25]. In the domain of public safety, events such as bar closings and football games are considered generators of crime. Preliminary analysis revealed that football games and bar closing events do indeed generate CSTPs. CSTP discovery can play an important role in disaster planning, climate change science [1, 31](e.g. understanding the effects of climate change and global warming) and public health (e.g. tracking the emergence, spread and re-emergence of multiple infectious diseases [68]). Further research is needed, however, to deal with challenges such as the lack of computationally efficient, statistically meaningful metrics to quantify interestingness, and the large cardinality of

candidate pattern sets that are exponential in the number of event types. Existing literature for spatio-temporal data mining focuses on mining totally ordered sequences or unordered subsets [42, 104, 17].

6.3 Broader Future Directions

In this chapter, we have presented the major research achievements and techniques which have emerged from spatial data mining, especially for predicting locations and discovering spatial outliers, co-location rules, and spatial clusters. Current research is mostly concentrated on developing algorithms that model spatial and spatio-temporal autocorrelations and constraints. Spatio-temporal data mining remains, however, still largely an unexplored territory; thus we conclude by noting other areas of research that require further investigation, such as the mining of trajectory and streaming data. New algorithms must be able to scale better to large datasets. Finally, and most urgently, methods are needed to validate the hypotheses generated by spatial data mining algorithms as well as to ensure that the knowledge generated is actionable.

7 Acknowledgments

We are particularly grateful to our collaborators Prof. Vipin Kumar, Prof. Paul Schrater, Prof. Sanjay Chawla, Dr. Chang-Tien Lu, Dr. Weili Wu, and Prof. Uygar Ozesmi, Prof. Yan Huang, and Dr. Pusheng Zhang for their various contributions. We also thank Pradeep Mohan, Dev Oliver, Xun Zhou, Abdulvahit Torun, Abdussalam Milad, Prof. Hui Xiong, Prof. Jin Soung Yoo, Dr. Qingsong Lu, Dr. Baris Kazar, and anonymous reviewers for their valuable feedback on early versions of this chapter. We would like to thank Kim Koffolt for improving the readability of this chapter.

References

[1] Strategic plan for the climate change science program. http://www.climatescience.gov/Library/stratplan2003/final/ccspstratplan2003-chap9.htm, 2003.

[2] Sky Server, 2007. http://skyserver.sdss.org/.

[3] W. W. 0010, J. Yang, and R. R. Muntz. Sting: A statistical information grid approach to spatial data mining. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 186–195. Morgan Kaufmann, 1997.

[4] T. Agarwal, R;Imielinski and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C.*, may 1993.

[5] R. Agrawal and R. Srikant. Fast algorithms for Mining Association Rules. In *Proc. of Very Large Databases*, may 1994.

[6] P. Albert and L. McShane. A generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics (Publisher: Washington, Biometric Society, etc.)*, 51:627–638, 1995.

[7] J. F. Allen. Towards a general theory of action and time. *Artif. Intell.*, 23(2):123–154, 1984.

[8] L. Anselin. *Spatial Econometrics: methods and models*. Kluwer, Dordrecht, Netherlands, 1988.

[9] L. Anselin. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dorddrecht, 1988.

[10] L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27(2):93–155, 1995.

[11] D. K. Arctur and M. Zeiler. *Designing Geodatabases*. ESRI Press, 2004. ISBN: 158948021X.

[12] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman & Hall, 2004.

[13] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, 3rd edition edition, 1994.

[14] E. Beinat, A. Godfrind, and R. V. Kothuri. *Pro Oracle Spatial*. Apress, 2004. ISBN: 978-1590593837.

[15] P. Bolstad. *GIS Fundamentals: A First Text on GIS*. Eider Press, 2002.

[16] M. Celik, S. Shekhar, B. George, J. P. Rogers, and J. A. Shine. Discovering and quantifying mean streets: A summary of results. Technical Report 025, University of Minnesota, 07 2007.

[17] M. Celik, S. Shekhar, J. P. Rogers, and J. A. Shine. Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):1322–1335, 2008.

[18] M. Celik, S. Shekhar, J. P. Rogers, J. A. Shine, and J. S. Yoo. Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 119–128, Washington, DC, USA, 2006. IEEE Computer Society.

[19] G. Cervone, P. Franzese, Y. Ezber, and Z. Boybeyi. Risk assessment of atmospheric hazard releases using k-means clustering. In *ICDM Workshops*, pages 342–348, 2008.

[20] D. Chamberlin. Using the New DB2: IBM's Object Relational System. *Morgan Kaufmann*, 1997. ISBN: 978-1558603738.

[21] S. Chawla, S. Shekhar, W. Wu, and U. Ozesmi. Modeling spatial dependencies for mining geospatial data. *1st SIAM International Conference on Data Mining*, 2001.

[22] N. Cressie. *Statistics for Spatial Data (Revised Edition)*. Wiley, New York, 1993.

[23] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[24] C. M. de Almeida, I. M. Souza, C. D. Alves, C. M. D. Pinho, M. N. Pereira, and R. Q. Feitosa. Multilevel object-oriented classification of quickbird images for urban population estimates. In *GIS*, page 12, 2007.

[25] L. C. P. Department. Lincoln city crime records. http://www.lincoln.ne.gov/city/police/, 2008.

[26] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid. Stagger: Periodicity mining of data streams using expanding sliding windows. In *ICDM*, pages 188–199, 2006.

[27] P. Elliott, J. Wakefield, N. Best, and D. Briggs. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2000. ISBN: 978-0192629418.

[28] A. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, 2002.

[29] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, and B. Pfahringer. Weka - a machine learning workbench for data mining. In *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314. 2005.

[30] C. Franke and M. Gertz. Detection and exploration of outlier regions in sensor data streams. In *ICDM Workshops*, pages 375–384, 2008.

[31] L. E. Frelich and P. B. Reich. Will environmental changes reinforce the impact of global warming on the prairie-forest border of central north america? *Frontiers in Ecology and the Environment*, 2009.

[32] A. R. Ganguly and K. Steinhaeuser. Data mining for climate change and impacts. In *ICDM Workshops*, pages 385–394, 2008.

[33] Google Earth, 2006. http://earth.google.com.

[34] R. Guting. An Introduction to Spatial Database Systems. In *Very Large Data Bases Journal(Publisher: Springer Verlag)*, October 1994.

[35] R. Guting and M. Schneider. *Moving Object Databases*. Morgan Kaufmann, 2005.

[36] J. Han, M. Kamber, and A. K. H. Tung. Spatial Clustering Methods in Data Mining: A Survey. In *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[37] J. Haslett, R. Bradley, P. Craig, A. Unwin, and G. Wills. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician*, pages 234–242, 1991.

[38] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.
[39] M. Hixson, D. Scholz, and N. Funs. Evaluation of several schemes for classification of remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 46:1547–1553, 1980.
[40] M. Hohn and L. G. A.E. Liebhold. A Geostatistical model for Forecasting the Spatial Dynamics of Defoliation caused by the Gypsy Moth, Lymantria dispar (Lepidoptera:Lymantriidae). *Environmental Entomology (Publisher: Entomological Society of America)*, 22:1066–1075, 1993.
[41] T. Hu, H. Xiong, X. Gong, and S. Y. Sung. Anemi: An adaptive neighborhood expectation-maximization algorithm with spatial augmented initialization. In *PAKDD*, pages 160–171, 2008.
[42] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):433–448, 2008.
[43] Issaks, Edward, and M. Svivastava. Applied Geostatistics. In *Oxford University Press, Oxford*, 1989.
[44] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
[45] A. Jain, M. Murty, and P. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
[46] V. P. Janeja and N. R. Adam. Homeland security and spatial data mining. In *Encyclopedia of GIS*, pages 434–440. 2008.
[47] Y. Jhung and P. H. Swain. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75, 1996.
[48] J. M. Kang, S. Shekhar, M. Henjum, P. Novak, and W. Arnold. Discovering Teleconnected Flow Anomalies: A Relationship Analysis of spatio-temporal Dynamic (RAD) neighborhoods. In *Symposium on Spatial and Temporal Databases*, 2009.
[49] J. M. Kang, S. Shekhar, C. Wennen, and P. Novak. Discovering Flow Anomalies: A SWEET Approach. In *International Conference on Data Mining*, 2008.
[50] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.
[51] SQL Server 2008 (Code-name Katmai), 2007. http://www.microsoft.com/sql/prodinfo/futureversion/default.mspx.
[52] B. Kazar, S. Shekhar, D. Lilja, and D. Boley. A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets. *SIAM International Conf. on Data Mining Workshop on High Performance and Distributed Mining (HPDM2004)*, April 2004.
[53] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66*, 1995.
[54] M. Koubarakis, T. Sellis, A. Frank, S. Grumbach, R. Guting, C. Jensen, N. Lorentzos, H. J. Schek, and M. Scholl. *Spatio-Temporal Databases: The Chorochronos Approach, LNCS 2520*, volume 9. Springer Verlag, 2003.
[55] P. Krugman. *Development, geography, and economic theory*. MIT Press, Cambridge, MA, 1995.
[56] C. Lai and N. T. Nguyen. Predicting density-based spatial clusters over time. In *ICDM*, pages 443–446, 2004.
[57] L. Lang. *Transportation GIS*. ESRI Press, 1999. ISBN: 978-1879102471.
[58] M. R. Leipnik and D. P. Albert. *GIS in Law Enforcement: Implementation Issues and Case Studies*. CRC, 2002. ISBN: 978-0415286107.
[59] N. Levine. *CrimeStat 3.0: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Ned Levine & Associatiates: Houston, TX / National Institute of Justice: Washington, DC, 2004.
[60] B. Li. Implementing spatial statistics on parallel computers. *Practical Handbook of Spatial Statistics, CRC Press*, pages 107–148, 1996.
[61] S. Li. A Markov Random Field Modeling. *Computer Vision (Publisher: Springer Verlag*, 1995.
[62] B. Little, M. Schucking, B. Gartrell, B. Chen, K. Ross, and R. McKellip. High granularity remote sensing and crop production over space and time: Ndvi over the growing season and prediction of cotton yields at the farm field level in texas. In *ICDM Workshops*, pages 426–435, 2008.
[63] C.-T. Lu, D. Chen, and Y. Kou. Detecting spatial outliers with multiple attributes. In *ICTAI '03: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, page 122, Washington, DC, USA, 2003. IEEE Computer Society.
[64] A. Luc. Exploratory Spatial Data Analysis and Geographic Information Systems. In M. Painho, editor, *New Tools for Spatial Analysis*, pages 45–54, 1994.
[65] A. Luc. Local Indicators of Spatial Association: LISA. *Geographical Analysis*, 27(2):93–115, 1995.
[66] D. Ma and A. Zhang. An adaptive density-based clustering algorithm for spatial database with noise. In *ICDM*, pages 467–470, 2004.
[67] Microsoft Virtual Earth, 2006. http://www.microsoft.com/virtualearth.
[68] D. M. Morens, G. K. Folkers, and A. S. Fauci. The challenge of emerging and re-emerging infectious diseases. *Nature*, 430:242–249, July 2004.
[69] Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.
[70] MySQL Spatial Extensions, 2007. http://dev.mysql.com/doc/refman/5.0/en/spatial-extensions.html.
[71] D. Neill and A. Moore. Rapid detection of significant spatial clusters. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265. ACM New York, NY, USA, 2004.

[72] R. T. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. *IEEE Trans. Knowl. Data Eng.*, 14(5):1003–1016, 2002.
[73] OGIS, 2007. http://www.opengeospatial.org/standards.
[74] R. Pace and J. LeSage. Closed-form maximum likelihood estimates for spatial problems (mess). *http://www.spatial-statistics.com*, 2000.
[75] R. Pace and J. LeSage. Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis*, 34(1):76–90, 2002.
[76] R. Pace and J. LeSage. Simple bounds for difficult spatial likelihood problems. *http://www.spatial-statistics.com*, 2003.
[77] R. Pastor. El niño climate pattern forms in pacific ocean, 2006, http://www.usatoday.com/weather/climate/2006-09-13-el-nino_x.htm.
[78] T. Pei, A. Jasra, D. J. Hand, A.-X. Zhu, and C. Zhou. Decode: a new method for discovering clusters of different densities in spatial data. *Data Min. Knowl. Discov.*, 18(3):337–369, 2009.
[79] Y. Pei, O. R. Zaïane, and Y. Gao. An efficient reference-based approach to outlier detection in large datasets. In *ICDM*, pages 478–487, 2006.
[80] PostGIS, 2007. http://postgis.refractions.net/.
[81] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
[82] R.J.Haining. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, U.K., 1989.
[83] J.-F. Roddick and M. Spiliopoulou. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. *SIGKDD Explorations 1(1): 34-38 (1999)*, 1999.
[84] R. Scally. *GIS for Environmental Management*. ESRI Press, 2006. ISBN: 978-589481429.
[85] J. Schiller. *Location-Based Services*. Morgan Kaufmann, 2004. ISBN: 978-1558609296.
[86] S. Shekhar and S. Chawla. Spatial databases: A tour. *Prentice Hall (ISBN 0-7484-0064-6)*, 2002.
[87] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C.-T. Lu. Spatial Databases - Accomplishments and Research Needs. *Trans. on Knowledge and Data Engineering 11(1): 45-55 (1999)*, 1999.
[88] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. *Proc. of Spatio-temporal Symposium on Databases*, 2001.
[89] S. Shekhar, C. Lu, and P. Zhang. Graph-based Outlier Detection : Algorithms and Applications (A Summary of Results). In *Proc. of the Seventh ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2001.
[90] S. Shekhar, P. Schrater, R. Raju, and W. Wu. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2):174–188, 2002.
[91] S. Shekhar, P. R. Schrater, R. R. Vatsavai, W. Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4(2), 2002.
[92] S. Shekhar and H. Xiong, editors. *Encyclopedia of GIS*. Springer, 2008.
[93] S. Shekhar, T. Yang, and P. Hancock. An Intelligent Vehicle Highway Information Management System. *Intl Jr. on Microcomputers in Civil Engineering (Publisher:Blackwell Publishers*, 8, 1993.
[94] A. H. Solberg, T. Taxt, and A. K. Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113, 1996.
[95] A. Stefanidis and S. Nittel. *GeoSensor Networks*. CRC, 2004. ISBN: 978-0415324045.
[96] P. Stolorz, H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S. Chien, R. Mechoso, and J. Farrara. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 300-305*, 1995.
[97] M. Stonebraker and D. Moore. Object Relational DBMSs: The Next Great Wave. *Morgan Kaufmann*, 1997. ISBN: 978-1558603974.
[98] A. Strahler. The use of prior probabilities in maximum likelihood classificaiton of remote sensing data. *Remote Sensing of Environment*, 10:135–163, 1980.
[99] P. Sun and S. Chawla. On local spatial outliers. In *ICDM*, pages 209–216, 2004.
[100] W. Tobler. *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.
[101] A. Tung, J. Hou, and J. Han. Spatial clustering in the presence of obstacles. *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 359–367, 2001.
[102] N. J. van Eck and L. Waltman. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5):625–645, 2007.
[103] V. Varnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, 1994.
[104] J. Wang, W. Hsu, and M. L. Lee. A framework for mining topological patterns in spatio-temporal databases. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 429–436, New York, NY, USA, 2005. ACM.
[105] C. E. Warrender and M. F. Augusteijn. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002, 1999.
[106] M. Worboys. *GIS: A Computing Perspective*. Taylor and Francis, 1995.
[107] M. Worboys and M. Duckham. *GIS: A Computing Perspective. Second Edition*. CRC, 2004. ISBN: 978-0415283755.
[108] M. F. Worboys. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28, 2005.

[109] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng. Localized outlying and boundary data detection in sensor networks. *IEEE Trans. Knowl. Data Eng.*, 19(8):1145–1157, 2007.

[110] Y. Yasui and S. Lele. A Regression Method for Spatial Disease Rates: An Estimating Function Approach. *Journal of the American Statistical Association*, 94:21–32, 1997.

[111] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. In H. V. Jagadish and I. S. Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 103–114. ACM Press, 1996.

[112] X. Zhang, J. Wang, F. Wu, Z. Fan, and X. Li. A novel spatial clustering with obstacles constraints based on genetic algorithms and k-medoids. In *ISDA '06: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*, pages 605–610, Washington, DC, USA, 2006. IEEE Computer Society.