Chapter 1

Spatial and Spatiotemporal Data Mining: Recent Advances

Shashi Shekhar
Department of Computer Science and Engineering, University of Minnesota, 4-192, EE/CS Bldg., 200 Union St. SE, Minneapolis, MN 55455.

Ranga Raju Vatsavai
Oak Ridge National Laboratory, MS-6017, Bldg. 5600, One Bethel Valley Rd, Oak Ridge, TN 37831

Mete Celik
Department of Computer Science and Engineering, University of Minnesota, 4-192, EE/CS Bldg., 200 Union St. SE, Minneapolis, MN 55455.

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. Spatiotemporal data mining is also briefly discussed.

1.1 Introduction

The explosive growth of spatial data and widespread use of spatial databases have heightened the need for the automated discovery of spatial knowledge. Spatial data mining [47, 39] 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 (NGA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology [37].

General purpose data mining tools like Clementine from 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, 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.

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) 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 it is often observed that these algorithms perform more poorly on spatial data. Many supportive examples can be found in the literature; for instance, parametric classifiers like maximum likelihood classifier(MLC) perform more poorly than non-parametric classifiers when the assumptions about the parameters (e.g., normal distribution) are violated, and the per-pixel based classifiers perform worse than Markov Random Fields (MRFs) when the features are auto-correlated.

In this chapter we present major accomplishments in the emerging field of spatial data mining, especially in the areas of prediction and classification, outlier detection, spatial co-location rules, and clustering techniques. Spatiotemporal data mining along with research needs are also briefly discussed.

1.2 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. 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 [8]. The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape. Relationships among non-spatial objects are explicit in data inputs, e.g., arithmetic relation, ordering, isinstance-of, subclass-of, and membership-of. In contrast, relationships among spatial objects are often implicit, such as overlap, intersect, and behind. 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 [33, 51, 1, 2, 21]. However, the materialization can result in loss of information. Another way to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the spatial data mining process.

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

TABLE 1.1: Relationships among Non-spatial Data and Spatial Data

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 are going to describe now favor one or more of these data types.

Statistical models [11] 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 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. For a random pattern, this average distance is expected to be $1/(2 \times \sqrt{density})$. where *density* is the average number of points per unit area. If for a real process, the computed distance falls within a certain limit, then we conclude that the pattern is generated by a random process; otherwise it is a non-random process.

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 [11], 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.

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 to cluster 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" [49]. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this property is called spatial autocorrelation.

Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Often the spatial dependencies arise due to the inherent characteristics of the phenomena under study, but in particular they arise due to the fact that the spatial resolution of imaging sensors are finer than the size of the object being observed. For example, remote sensing satellites have resolutions ranging from 30 meters (e.g.,

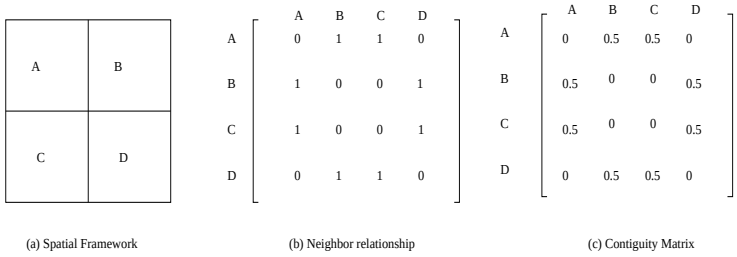


FIGURE 1.1: A spatial framework and its four-neighborhood contiguity matrix.

the Enhanced Thematic Mapper of the Landsat 7 satellite of NASA) to one meter (e.g., the IKONOS satellite from SpaceImaging), while the objects under study (e.g., Urban, Forest, Water) are often much larger than 30 meters. As a result, per-pixel-based classifiers, which do not take spatial context into account, often produce classified images with salt and pepper noise. These classifiers also suffer in terms of classification accuracy.

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, Euclidean distance, etc. Example definitions of neighborhood using adjacency include a four-neighborhood and an eight-neighborhood. Given a gridded spatial framework, a four-neighborhood assumes that a pair of locations influence each other if they share an edge. An eight-neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

Figure 1.1(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 1.1(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 1.1(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 [52]. In spatial statistics, spatial autocorrelation is quantified using measures such as Ripley's K-function and Moran's I [11].

1.3 Prediction and Classification

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

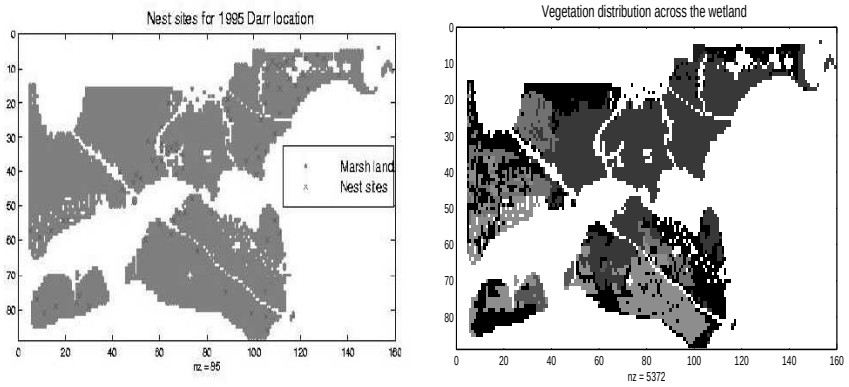
predicts an output for any unseen input (but assumed to be generated from the same distribution), such that the predicted output is as close as possible to the desired output. The name "supervised" comes from the fact that the input-output example pairs are given by an expert (teacher). Examples of the supervised learning include thematic map generation (classification) from satellite images, tumor or other organ recognition from medical images, recognition of hand written characters from the scanned documents, prediction of stock market indexes, and speech recognition. The input-output pairs, also called training samples, or training dataset, is denoted by (x_i, y_i), where x_i's are often vectors of measurements over the attribute space. 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.), and 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:

- **Classification:** In classification, the input vectors x_i are assigned to a few discrete numbers of classes y_i.

- **Regression:** In regression, also known as function approximation, 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 prediction and estimation, for example, share value prediction, 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.

Classification can be viewed as a special case of regression. In this section we specifically consider the problem of multi-spectral remote sensing image classification. Image classification can be formally defined as finding a function $g(x)$ which maps the input patterns x onto output classes y_i (sometimes y_i's are also denoted as ω_i). The main objective is to assign a label (e.g. Water, Forest, Urban) to each pixel in the classified image, given corresponding feature vector x_j in the input image.

The prediction of events occurring at particular geographic locations is very important in several application domains. Crime analysis, cellular networks, and natural disasters such as fires, floods, droughts, vegetation diseases, and earthquakes are all examples of problems which require location prediction. In this section we present two spatial data mining techniques, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF). Before explaining the techniques, we introduce an example application domain to illustrate different concepts in spatial data mining.



(a) Nest Locations (b) Vegetation Durability

FIGURE 1.2: (a) Learning dataset: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland.

1.3.1 An Illustrative Application Domain

We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio USA in order to *predict* the spatial distribution of a marsh-breeding bird, the red-winged blackbird (*Agelaius phoeniceus*). The data was collected from April to June in two successive years, 1995 and 1996.

A uniform grid was imposed on the two wetlands and different types of measurements were recorded at each cell or pixel. In total, values of seven attributes were recorded at each cell. Domain knowledge is crucial in deciding which attributes are important and which are not. For example, *Vegetation Durability* was chosen over *Vegetation Species* because specialized knowledge about the bird-nesting habits of the red-winged blackbird suggested that the choice of nest location is more dependent on plant structure, plant resistance to wind, and wave action than on the plant species.

Our goal is to build a model for predicting the location of bird nests in the wetlands. Typically the model is built using a portion of the data, called the **Learning** or **Training** data, and then tested on the remainder of the data, called the **Testing** data. In the learning data, all the attributes are used to build the model and in the testing data, one value is *hidden*, in our case the location of the nests.

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

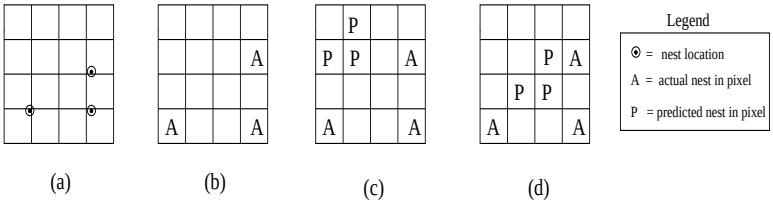


FIGURE 1.3: (a)The actual locations of nests, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another model. Prediction(d) is spatially more accurate than (c).

classification 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 1.3. Figure 1.3(a) shows the actual locations of nests and 1.3(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 1.3(c) and 1.3(d). Domain scientists prefer prediction 1.3(d) over 1.3(c), since the predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 1.3(c) and 1.3(d), and a measure of spatial accuracy is needed to capture this preference.

1.3.2 Modeling Spatial Dependencies Using the SAR and MRF Models

Several previous studies [22], [45] 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 1.1.

1.3.3 Logistic Spatial Autoregressive Model(SAR)

Logistic SAR decomposes a classifier $\hat{f}_C$ into two parts, namely spatial autoregression and logistic transformation. We first show how 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[4]. If

the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{1.1}$$

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. After the correction term $\rho W y$ is introduced, the components of the residual error vector ϵ are then assumed to be generated from independent and identical standard normal distributions. As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables.

We refer to this equation as the Spatial Autoregressive Model (SAR). Notice that when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: First, the residual error will have much lower spatial autocorrelation (i.e., systematic variation). With the proper choice of W, the residual error should, at least theoretically, have no systematic variation. In addition, if the spatial autocorrelation coefficient is statistically significant, then SAR will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. Finally, the model will have a better fit, (i.e., a higher R-squared statistic).

1.3.4 Maximum Likelihood Classifier (MLC)

Maximum likelihood classification is one of the most widely used parametric and supervised classification technique in remote sensing field [18], [48]. Assuming that sufficient ground truth (training) data is available for each thematic class, we can estimate the probability distribution $p(x|y_i)$ for a class (y_i) that describes the chance of finding a pixel from that class at the position $\mathbf{x}$. This estimated $p(y_i|x)$ can be related with the desired $p(x|y_i)$ using Bayes' theorem:

$$p(y_i|x) = \frac{p(x|y_i)p(y_i)}{p(x)} \tag{1.2}$$

where $p(y_i)$ is the probability that class y_i occurs in the image, also know as 'a priori' probability, and $p(\mathbf{x})$ is the probability of finding a pixel from any class at location $\mathbf{x}$. Since $p(\mathbf{x})$ is constant, we can omit it from computation and write the discriminant function $g(\mathbf{x})$ by simplify eq 1.2 and taking logarithm as follows:

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|y_i) + \ln p(y_i), \tag{1.3}$$

where ln is the natural logarithm. By assuming a multivariate normal model

for class probability distributions, the discriminant function $g_i(x)$ for the maximum likelihood classification can be written as the following.

$$g_i(\mathbf{x}) = \ln p(\omega_i) - \frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^t \Sigma_i^{-1}(\mathbf{x} - \mathbf{m}_i) \tag{1.4}$$

MLC is an example of a Bayesian classifier; for more details refer to [35, 12].

1.3.5 Markov Random Field-based Bayesian Classifiers

The MLC presented above 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. We now present Markov Random Field (MRF) based Bayesian classifiers that model spatial context via the *a priori* term in Bayes' rule. A set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov Random Field [27]. The Markov property specifies that a variable depends only on its neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $l_i = f_C(s_i)$, of different locations, s_i, constitutes an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

The Bayesian rule can be used to predict l_i from feature value vector X and neighborhood class label vector L_i as follows:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{1.5}$$

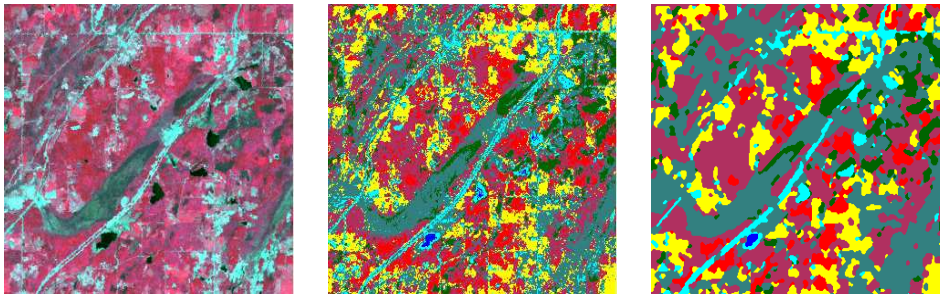
The solution procedure can estimate $Pr(l_i|L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X|l_i, L_i)$ can be estimated using kernel functions from the observed values in the training dataset. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. An assumption on $Pr(X|l_i, L_i)$ may be useful if the training dataset available is not large enough. A common assumption is the uniformity of influence from all neighbors of a location. For computational efficiency it can be assumed that only local explanatory data $X(s_i)$ and neighborhood label L_i are relevant in predicting class label $l_i = f_C(s_i)$. It is common to assume that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with a Gibbs distribution specialized by the locally defined MRF through the Hammersley-Clifford theorem [7].

A more detailed theoretical and experimental comparison of these methods can be found in [42]. Although MRF and SAR classification have different formulations, they share a common goal, estimating the posterior probability distribution: $p(l_i|X)$. However, the posterior for the two models is computed differently with different assumptions. For MRF the posterior is computed using Bayes' rule. In logistic regression, the posterior distribution is directly fit to the data. One important difference between logistic regression and MRF is that logistic regression assumes no dependence on neighboring classes. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by $Pr(u|v) = e^{A(\theta_v)+B(u,\pi)+\theta_v^T u}$, where u and v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases.

	CID	C.Nane	MLC	MRF
1	1	Hardwood.1	79.82	95.34
2	2	Hardwood.2	82.96	87.18
3	3	Conifer	94.02	96.60
4	4	Agriculture	90.60	93.03
5	5	Urban	53.57	64.29
6	6	Wetlands	93.51	95.15
7	7	Water	100.00	100.00
8	O	Overall	87.05	91.82

TABLE 1.2: MLC vs. MRF Classification Accuracy

Experiments were carried out on the Darr and Stubble wetlands to compare the classical regression, SAR, and the MRF-based Bayesian classifiers. The results showed that the MRF models yield better spatial and classification accuracies over SAR in the prediction of the locations of bird nests. We also observed that SAR predications are extremely localized, missing actual nests over a large part of the marsh lands. We also compared performance of MRF against MLC in a multi-class satellite image classification setting. We used a spring Landsat 7 image, taken May 31, 2000, and clipped to the study region (Carlton County, Minnesota). The final rectified and clipped image size is 1343 lines x 2019 columns x 6 bands. We trained MLC and MRF classifiers using 60 labeled training plots and tested performance using an independent test dataset consisting of 205 labeled plots. The accuracies are summarized in Table 1.2 and Figure 1.4 shows small windows from the classified images.



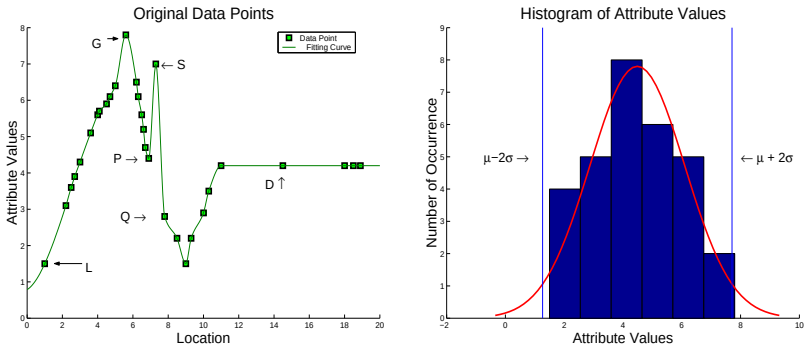
(a) RGB (b) MLC (c) MRF

FIGURE 1.4: Sample RGB image and corresponding MLC and MRF Classified output

1.4 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 [6], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism [17]. 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, athlete performance analysis, and severe weather prediction. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases. These application domains include transportation, ecology, public safety, public health, climatology, and location-based services.

We model a spatial dataset to be a collection of spatially referenced objects, such as houses, roads, and traffic sensors. Spatial objects have two distinct categories of dimensions along which attributes may be measured. Categories of dimensions of interest are spatial and non-spatial. Spatial attributes of a spatially referenced object include location, shape, and other geometric or topological properties. Non-spatial attributes of a spatially referenced object include traffic-sensor identifiers, manufacturer, owner, age, and measurement readings. A spatial neighborhood of a spatially referenced object is a subset of the spatial data based on a spatial dimension, e.g., location. Spatial neighborhoods may be defined based on spatial attributes, e.g., location, using spatial relationships such as distance or adjacency. Comparisons between spatially referenced objects are based on non-spatial attributes.



(a) An Example Dataset (b) Histogram

FIGURE 1.5: A Dataset for Outlier Detection.

A spatial outlier [41] 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.

1.4.1 Illustrative Examples and Application Domains

We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 1.5(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. The outlier detected using this approach is the data point G, which has an extremely high attribute value 7.9, exceeding the threshold of $\mu + 2\sigma = 4.49 + 2 * 1.61 = 7.71$, as shown in Figure 1.5(b). 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.

As another example, we use a spatial database consisting of measurements from the Minneapolis-St. Paul freeway traffic sensor network. The sensor network includes about nine hundred stations, each of which contains one to four loop detectors, depending on the number of lanes. Sensors embedded in the freeways and interstate monitor the occupancy and volume of traffic on

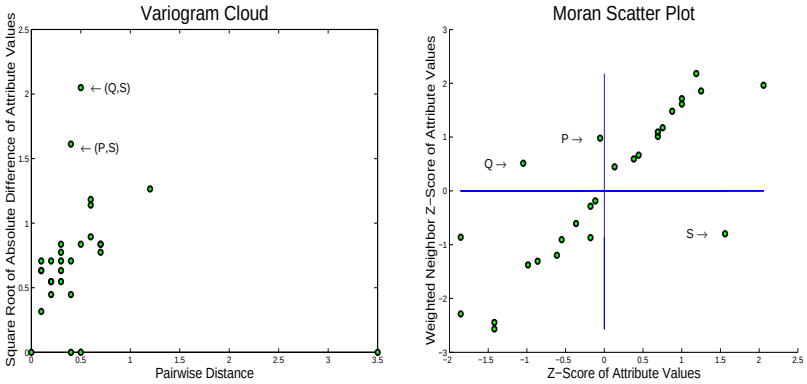
the road. At regular intervals, this information is sent to the Traffic Management Center for operational purposes, e.g., ramp meter control, as well as for experiments and research on traffic modeling. In this application, we are interested in discovering the location of stations whose measurements are inconsistent with those of their spatial neighbors and the time periods when those abnormalities arise.

1.4.2 Tests for Detecting Spatial Outliers

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 and Moran scatterplots. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data. Scatterplots [29] are a representative technique from the quantitative family.

A variogram cloud displays data points related by neighborhood relationships. For each pair of locations, the square-root of the absolute difference between attribute values at the locations versus the Euclidean distance between the locations are plotted. In datasets exhibiting strong spatial dependence, the variance in the attribute differences will increase with increasing distance between locations. Locations that are near to one another, but with large attribute differences, might indicate a spatial outlier, even though the values at both locations may appear to be reasonable when examining the dataset non-spatially. Figure 1.6(a) shows a variogram cloud for the example dataset shown in Figure 1.5(a). This plot shows that two pairs (P, S) and (Q, S) on the left hand side lie above the main group of pairs and are possibly related to spatial outliers. The point S may be identified as a spatial outlier since it occurs in both pairs (Q, S) and (P, S). However, graphical tests of spatial outlier detection are limited by the lack of precise criteria to distinguish spatial outliers. In addition, a variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present, or density varies greatly.

A Moran scatterplot [30] is a plot of normalized attribute value ($Z[f(i)] = \frac{f(i)-\mu_f}{\sigma_f}$) against the neighborhood average of normalized attribute values ($W \cdot Z$), where W is the row-normalized (i.e., $\sum_j W_{ij} = 1$) neighborhood matrix, (i.e., $W_{ij} > 0$ iff neighbor(i, j)). The upper left and lower right quadrants of Figure 1.6(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). Thus we can identify points(nodes) that are surrounded by unusually high or low value neighbors. These points



(a) Variogram cloud (b) Moran scatterplot

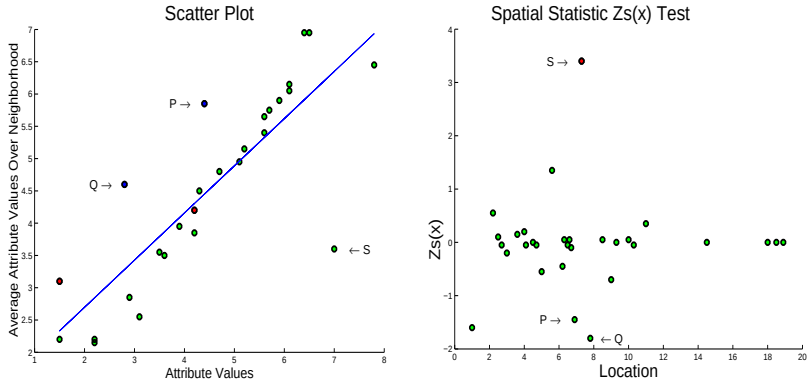
FIGURE 1.6: Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers.

can be treated as spatial outliers.

A scatterplot [29] 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 1.7(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the dataset in Figure 1.5(a). The point S turns out to be the farthest from the regression line and may be identified as a spatial outlier.

A location (sensor) is compared to its neighborhood using the function $S(x) = [f(x) - E_{y \in N(x)}(f(y))]$, where $f(x)$ is the attribute value for a location x, $N(x)$ is the set of neighbors of x, and $E_{y \in N(x)}(f(y))$ is the average attribute value for the neighbors of x. The statistic function $S(x)$ denotes the difference of the attribute value of a sensor located at x and the average attribute value of $x's$ neighbors.

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 $Z_{s(x)} =$



(a) Scatterplot

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

FIGURE 1.7: Scatterplot and Spatial Statistic $Z_{s(x)}$ to Detect Spatial Outliers.

$|\frac{S(x)-\mu_s}{\sigma_s}| > \theta$. 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 1.7(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 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.

1.4.3 Outliers in Spatial Networks

The outlier detection techniques presented so far are most suitable for general spatial databases and time series databases. However, these approaches do not consider the spatial network structure of the dataset, and may not model graph properties such as one-ways, connectivities etc. Recently methods are proposed to discover graph-based hotspots, e.g. *mean streets*, which represent those connected subsets of a spatial network whose attribute values are significantly higher than expected. Finding *mean streets* is very im-

portant for many application domains, including 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 spatiotemporal (ST) infrastructure networks, such as transportation, oil/gas pipelines, and utilities (e.g., water, electricity, telephone). Thus, activity reports such as crime reports 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 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 ST network properties, and designing ST network control policies.

However, identifying and quantifying *mean streets* is a challenging task for several reasons. One large challenge is choosing the correct statistical model. Many existing ST models assume data normality and either spatial and temporal homogeneity or a well-defined autocorrelation in these domains. A major limitation is the inadequacy of descriptive and explanatory models for activity around ST networks such as train and road networks. Another challenge is that the discovery process of *mean streets* 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.

Public safety professionals may be interested in analyzing the ST network factors to explain high activity levels or changes in activity levels at certain highway segments, or to compare prevention options such as check points. Such analysis is not only difficult using existing methods, but it may not be statistically meaningful, since common methods such as spatial regression do not adequately model ST network constraints such as connectivity and directions.

Previous studies on discovering high-density regions (i.e. hotspots) can be classified into two main categories based on their statistical interpretability. For example, [28] defined the hot routes discovery problem in road networks using moving object trajectories. However, discovered patterns in this approach do not have a statistical interpretation such as statistical significance. In addition, this algorithm is designed to process tracks (e.g., GPS tracks) rather than point or aggregate datasets referencing street networks. A distance based technique to detect outliers in spatial networks is presented in [23].

Statistics-based methods to identify hotspots can be classified into two categories based on the nature of the dataset: point-based methods [5, 13, 14, 26, 34, 36, 44, 32] and aggregate-based methods. *Mean streets* problem belongs to the latter one. The aim of the point-based approaches is to discover high-density regions from point datasets which show the actual locations of the crimes (Figure 1.9). The point-based approaches focus on the discovery

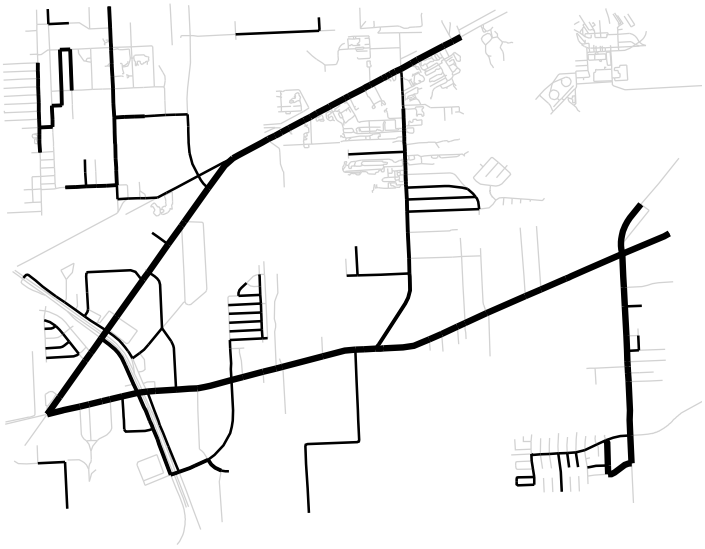


FIGURE 1.8: *Mean streets* of a metropolitan city of the United States

of the geometry (e.g. circle, ellipse, etc.) of the high-density regions [13]. The Spatial and Temporal Analysis of Crime (STAC) tool in the CrimeStat software, nearest neighbor hierarchical clustering techniques, and K-means clustering techniques are among the methods that use the ellipse method to identify hotspots [26]. Figure 1.9 shows the result of CrimeStat using the K-means clustering method for 15 clusters [26]. Kernel estimation methods have been developed to identify isodensity hotspot surfaces because hotspots may not have crisp ellipsoid boundaries. Local indicators of spatial association (LISA) statistics were proposed to eliminate the limitations of ellipsoid-based and kernel-based estimation techniques [5, 14]. The clumping method was proposed by Roach to discover clumped points (e.g. hotspots) from a point dataset [36]. However, these approaches will not be able to discover and quantify high-crime-density regions (e.g. streets) for given aggregate crime data. They also do not consider the spatial network structure of the urban dataset, and may not model graph properties such as one-way streets or connectivity. For example, if all crime events occur along a street of a city, these approaches may tend to divide the street into several ellipsoid clusters or may tend to discover a big ellipse where most of the inside of the area has no activity. The clumping method for analyzing point patterns on a spatial network [36, 44, 32] was extended in [44]. In this extended approach, if crime point locations on an edge are close enough, they form a clump. A user-defined distance threshold (or clump radius) is used to check if the points are close enough or not. However, their approach will not be able to discover and quantify patterns for aggregate crime data.

Overall, point-based approaches mainly focus on discovering and quantifying hotspots using point crime data. However, due to the type of crime or concerns for victim security, crime location information may not be released by the authorities and only aggregated crime values may be released for spatial regions, e.g. streets. In that case, point-based approaches, whether they consider the spatial network structure or not, will fail to discover and quantify hotspots since these approaches are dependent on knowing the locations of the crimes. In contrast, statistics-based methods are proposed to discover hotspots (e.g. *mean streets*) from aggregated datasets referencing urban street networks and taking graph semantics into account.

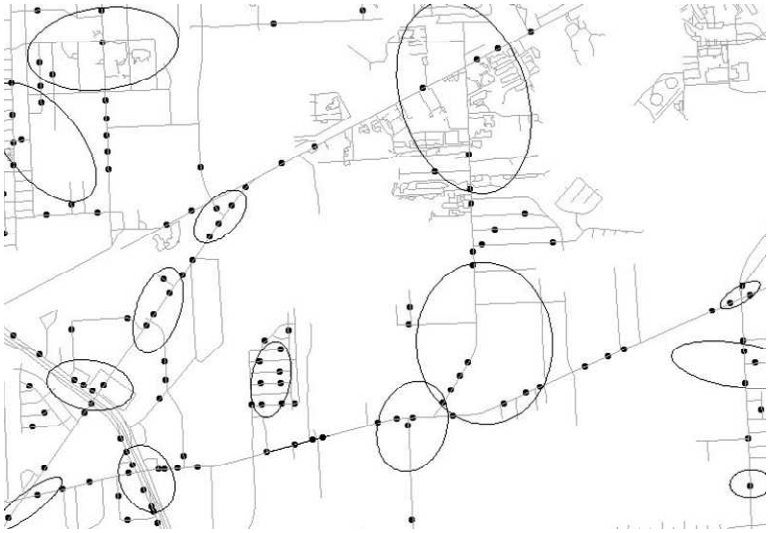


FIGURE 1.9: Point data and output of K-means clustering using Crime Stat

A novel ST network analysis method is explored to study descriptive and explanatory models for ST network patterns in [9]. Formally, given a road network $G = (V, E)$ and a set of aggregated crime values on edges E, *mean street* mining algorithm aims to discover and quantify correct and complete sets of connected subsets of the road network. For example, Figure 1.8 shows " *mean streets*" of a part of a metropolitan city in the United States. Each line represents a street and the thickness of it represents the aggregated crime value of the street. In this figure, the thicker the street is, the higher the crime density is. Two algorithms are developed: An apriori-based *mean street* miner and a graph-based *mean street* miner. The key idea behind the apriori-based method is to discover *size* $k + 1$ *mean streets* using *size* k *mean streets*. This approach has two pruning strategies: i) to eliminate unconnected edge

combinations, and ii) to eliminate edge combinations that do not satisfy the crime thresholds. This approach will generate $size\ k+1$ streets using $size\ k$ "*mean streets*" until there are no more candidate streets. The inputs of the algorithm are a road network $G = (V, E)$, a set of aggregated crime values C_{real}, and a user-defined confidence threshold α. The output is connected sets of streets whose aggregated crime values are no less than their crime thresholds $C_{treshold}$.

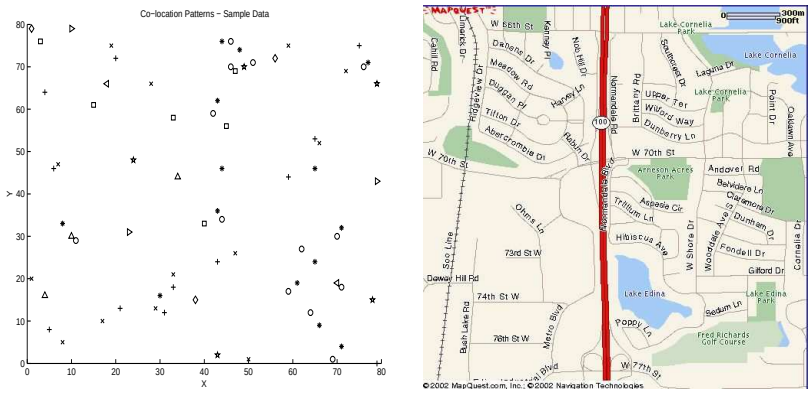
On the other hand, in the graph-based approach, the key idea is to generate all possible street sets in a spatial network using path generation algorithms and prune the streets that do not satisfy the criteria. Road networks are often represented as graphs and one method to generate *mean streets* is to find all possible paths in the graph and then use an appropriate filtering technique to eliminate the connected street sets that are irrelevant. The constraints that need to be satisfied while computing street sets would depend on the users' preferences. For example, in some scenarios, it might be required to generate connected street sets that traverse every edge in the graph at least once. It is also possible that some locations in the road network are designated as start points and end points and the connected street set generation needs to incorporate this requirement.

For the Apriori-based approach, a significant part of the computation time would be spent in generating candidate *mean streets* without looking at the connectivity of the edges. Experimental results showed that the execution time variation in the graph-based approach is less pronounced. This is more computationally efficient than the apriori-based approach since only the connected paths are generated. The apriori-based algorithm generates candidates without checking the graph connectivity, thus increasing the size of the search space. The execution time of the graph-based approach decreases as the confidence threshold increases. The apriori-based approach is computationally more expensive as the confidence threshold decreases because of the increase in the number of *mean streets* to be discovered. These two methods were evaluated and results were summarized in [9].

1.5 Co-location Rules

Co-location patterns represent subsets of boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species, e.g. the Nile Crocodile and Egyptian Plover in ecology and frontage-roads and highways in metropolitan road maps. 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, animal species, road types, cancers, crime, and business types.



(a) (b)

FIGURE 1.10: a) Illustration of Point Spatial Co-location Patterns. Shapes represent different spatial feature types. Spatial features in sets {'+', '×'} and {'o', '*'} tend to be located together. b) Illustration of Line String Co-location Patterns. Highways, e.g. Hwy100, and frontage roads, e.g. Normandale Road, are co-located.

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 1.10(a) 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 [2] 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. The purpose of mining association rules is to identify frequent item sets for planning store layouts or marketing campaigns. 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.

1.5.1 Co-location Rule Approaches

Approaches to discovering co-location rules can be divided into three categories: those based on spatial statistics, those based on association rules, and those based on the event centric model. Spatial statistics-based approaches use measures of spatial correlation to characterize the relationship between different types of spatial features using the cross K function with Monte Carlo simulation and quadrat count analysis [11]. Computing spatial correlation measures for all possible co-location patterns can be computationally expensive due to the exponential number of candidate subsets given a large collection of spatial boolean features.

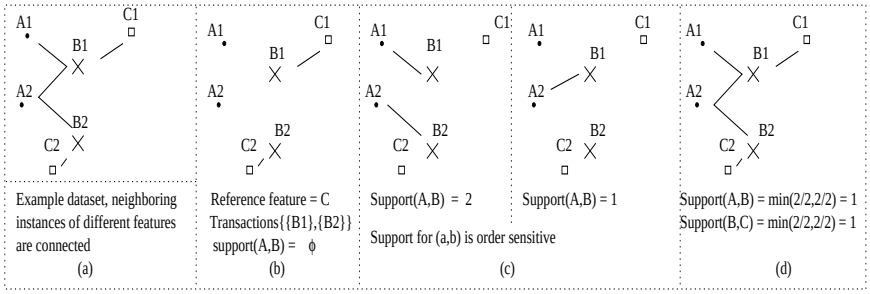


FIGURE 1.11: Example to Illustrate Different Approaches to Discovering Co-location Patterns a) Example dataset. b) Data partition approach. Support measure is ill-defined and order sensitive c) Reference feature centric model d) Event centric model

Association rule-based approaches focus on the creation of transactions over space so that an *apriori* like algorithm [2] can be used. Transactions in space can use a reference-feature centric [24] approach or a data-partition [31] approach. The **reference feature centric model** is based on the choice of a reference spatial feature [24] and is relevant to application domains focusing on a specific boolean spatial feature, e.g. cancer. Domain scientists are interested in finding the co-locations of other task relevant features (e.g. asbestos) to the reference feature. A specific example is provided by the spatial association rule [24]. Transactions are created around instances of one user-specified reference spatial feature. The association rules are derived using the *apriori*

algorithm. The rules found are all related to the reference feature. For example, consider the spatial dataset in Figure 1.11(a) with three feature types, A, B and C. Each feature type has two instances. The neighbor relationships between instances are shown as edges. Co-locations (A, B) and (B, C) may be considered to be frequent in this example. Figure 1.11(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.

Defining transactions by a data-partition approach [31] defines transactions by dividing spatial datasets into disjoint partitions. There may be many distinct ways of partitioning the data, each yielding a distinct set of transactions, which in turn yields different values of support of a given co-location. Figure 1.11 c) shows two possible partitions for the dataset of Figure 1.11 a), along with the supports for co-location (A, B).

Model	Items	Transactions defined by	Interest measures for $C_1 \rightarrow C_2$	
			Prevalence	Conditional probability
reference feature centric	predicates on reference and relevant features	instances of reference feature C_1 and C_2 involved with	fraction of instance of reference feature with $C_1 \cup C_2$	$Pr(C_2$ is true for an instance of reference features given C_1 is true for that instance of reference feature)
data partitioning	boolean feature types	a partitioning of spatial dataset	fraction of partitions with $C_1 \cup C_2$	$Pr(C_2$ in a partition given C_1 in that partition)
event centric	boolean feature types	neighborhoods of instances of feature types	participation index of $C_1 \cup C_2$	$Pr(C_2$ in a neighborhood of $C_1)$

TABLE 1.3: Interest Measures for Different Models

The event centric model finds subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types. For example, let us determine the probability of finding at least one instance of feature type B in the neighborhood of an instance of feature type A in Figure 1.11 a). There are two instances of type A and both have some instance(s) of type B in their neighborhoods. The conditional probability for the co-location rule is: *spatial feature A at location $l \rightarrow$ spatial feature type B in neighborhood is 100%*. This yields a well-defined prevalence measure(i.e. support) without

the need for transactions. Figure 1.11 d) illustrates that our approach will identify both (A, B) and (B, C) as frequent patterns.

Prevalence measures and conditional probability measures, called interest measures, are defined differently in different models, as summarized in Table 1.3. The reference feature centric and data partitioning models "materialize" transactions and thus can use traditional support and confidence measures. The event centric model-based approach defined new transaction free measures, such as the participation index (please refer to [40] for details).

1.6 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. Cluster analysis is used in many spatial and spatiotemporal application domains. For example, clustering is used in remote sensing data analysis as a first step to determine the number and distribution of spectral classes. Cluster analysis is used in epidemiology for finding unusual groups of health-related events. Cluster analysis is also used in detection of crime hot spots.

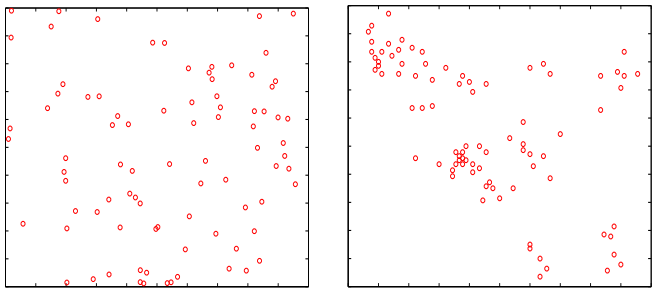
1.6.1 Complete Spatial Randomness and Clustering

Spatial clustering can be applied to group similar spatial objects together, and its implicit assumption is that patterns tend to be grouped in space rather than in a random pattern. The statistical significance of spatial clustering can be measured by testing the assumption in the data. The test is critical for proceeding to any serious clustering analysis.

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) [11] 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 1.12, 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).

Notice from Figure 1.12 (a) that the complete spatial randomness pattern seems to exhibit some clustering. This is not an unrepresentive realization, but



(a) CSR Pattern (b) Clustered Pattern

FIGURE 1.12: Complete Spatial Random (CSR) and Spatially Clustered Patterns

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 [11]. In contrast to Figure 1.12 (a), true clustering is shown in Figure 1.12 (b).

Several statistical methods [11] can be applied to quantify deviations of patterns from complete spatial randomness point pattern. One type of descriptive statistics is based on quadrats (i.e., well defined area, often rectangle in shape). Usually quadrats of random locations and orientations in the quadrats are counted, and statistics derived from the counters are computed. Another type of statistics is based on distances between patterns. One such type is Ripley's K function.

1.6.2 Categories of Clustering Algorithms

After verification of the statistical significance of spatial clustering, clustering algorithms are used to discover interesting clusters. Because of the multitude of clustering algorithms that have been developed, it is useful to categorize them into groups. Based on the technique adopted to define clusters, the clustering algorithms can be divided into four broad categories:

1. *Hierarchical* clustering methods, which 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 iterative reducing and clustering using hierarchies (BIRCH), clustering using inter-connectivity (CHAMELEON), clustering using representa-

tives (CURE), and robust clustering using links (ROCK).

2. *Partitional* clustering algorithms, which 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).

3. *Density-based* clustering algorithms, which 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 (DENCLUE).

4. *Grid-based* clustering algorithms, which first quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. Cells that contain more than a certain number of points are treated as dense. The dense cells are connected to form the clusters. Grid-based clustering algorithms are primarily developed for analyzing large spatial datasets. The grid-based clustering algorithms include the statistical information grid-based method (STING), WaveCluster, BANG-clustering, and clustering-in-quest (CLIQUE).

Sometimes the distinction among these categories diminishes, and some algorithms can even be classified into more than one group. For example, clustering-in-quest (CLIQUE) can be considered as both a density-based and grid-based clustering method. More details on various clustering methods can be found in a recent survey paper [16]. Many of the clustering algorithms discussed here do not take into account the spatial autocorrelation and spatial constraints. Limited studies can be found in the literature to model spatial neighborhood relationships in clustering process. For example, in [3, 19] the conventional expectation maximization approach has been extended by incorporating a spatial penalty term in estimating the likelihood function. Likewise, algorithms for spatial clustering in the presence of obstacles have been proposed in [50, 53]. These approaches shows improved clustering results and stress the importance of modeling neighborhood relationships in clustering.

1.7 Spatiotemporal data mining

So far we have discussed techniques that are applicable to spatial data. Like spatial data, which requires consideration of spatial autocorrelation and spatial relationships and constraints in the model building, spatiotemporal data mining also requires explicit or implicit modeling of spatiotemporal autocorrelation and constraints. Several spatiotemporal extensions of classification, clustering, and outlier detection can be found in the literature [38]. In this section we consider the problem of spatiotemporal co-occurrence pattern mining and briefly discuss the algorithm recently presented in [10].

Mixed-drove spatiotemporal 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 an important problem with many applications such as identifying tactics in battlefields, games, and predator-prey interactions. 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. A monotonic composite interest measure for discovering MDCOPs and novel MDCOP mining algorithms are presented in [10].

As the volume of spatiotemporal data continues to increase significantly due to both the growth of database archives and the increasing number and resolution of spatiotemporal sensors, automated and semi-automated pattern analysis becomes more essential. As a result, spatiotemporal co-occurrence pattern mining has been the subject of recent research. Given a moving object database, the aim in [10] was to discover mixed-drove spatiotemporal co-occurrence patterns (MDCOPs) representing subsets of different object-types whose instances are located close together in geographic space for a significant fraction of time. Unlike the objectives of some other spatiotemporal co-occurrence pattern identification approaches where the pattern is the primary interest, in MDCOPs both the pattern and the nature of the different *object-types* are of interest.

A simple example of an MDCOP is in ecological predator-prey relationships. Patterns of movements of rabbits and foxes, for example, will tend to be co-located in many time-frames which may or may not be consecutive. Rabbits may attempt to move away from foxes, and the foxes may attempt to stay with the rabbits. Other factors such as available food and water may also affect the patterns.

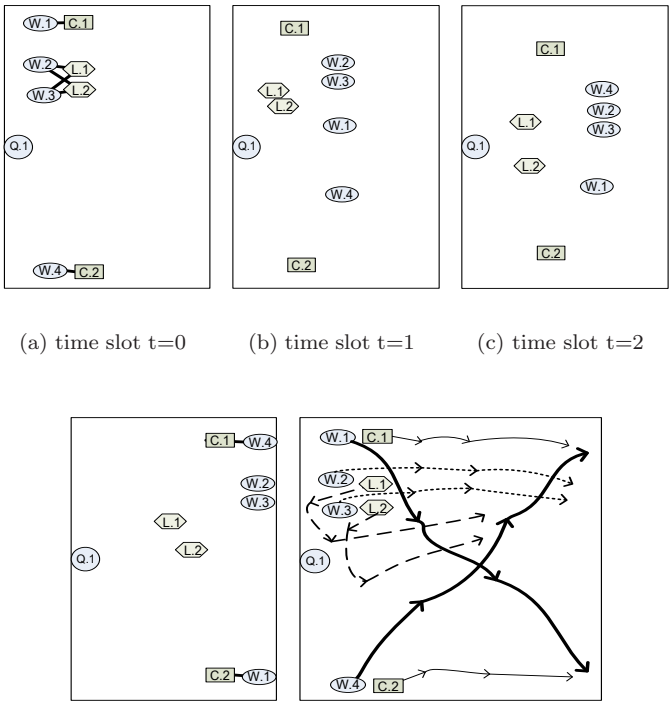
A detailed example: More example MDCOPs may be illustrated in American football where two teams try to outscore each other by moving a football to the opponent's end of the field. Various complex interactions occur within one team and across teams to achieve this goal. These interactions involve intentional and accidental MDCOPs, the identification of which may

help teams to study their opponent's tactics. In American football, object-types may be defined by the roles of the offensive and defensive players, such as quarterback, running back, wide receiver, kicker, holder, linebacker, and cornerback. An MDCOP is a subset of these different object-types (such as {kicker, holder} or {wide_receiver, cornerback}) that occur frequently. One example MDCOP involves offensive wide receivers, defensive linebackers, and defensive cornerbacks, and is called a Hail Mary play. In this play, the objective of the offensive wide receivers is to outrun any linebackers and defensive backs and get behind them, catching an undefended pass while running untouched for a touchdown. This interaction creates an MDCOP between wide receivers and cornerbacks. An example Hail Mary play is given in Figure 1.13. It shows the positions of four offensive wide receivers (W.1, W.2, W.3, and W.4), two defensive cornerbacks (C.1 and C.2), two defensive linebackers (L.1 and L.2), and a quarterback (Q.1) in four time slots. The solid lines between the players show the neighboring players. The wide receivers W.1 and W.4 cross over each other and the wide receivers W.2 and W.3 run directly to the end zone of the field. Initially, the wide receivers W.1 and W.4 are co-located with cornerbacks C.1 and C.2 respectively and the wide receivers W2. and W.3 are co-located with linebackers L.1 and L.2 at time slot t=0 (Figure 1.13 (a)). In time slot t=1, the four wide receivers begin to run, while the linebackers run towards the quarterback and the cornerbacks remain in their original position, possibly due to a fake handoff from the quarterback to the running back (Figure 1.13 (b)). In time slot t=2, the wide receivers W.1 and W.4 cross over each other and try to drift further away from their respective cornerbacks (Figure 1.13 (c)). When the quarterback shows signs of throwing the football, both cornerbacks and linebackers run to their respective wide receivers (Figure 1.13(d)). The overall sketch of the game tactics can be seen in Figure 1.13(e). In this example, wide receivers and cornerbacks form an MDCOP since they are persistent over time and they occur 2 out of 4 time slots. However, wide receivers and linebackers do not form an MDCOP due to the lack of temporal persistence.

There are many applications for which discovering co-occurring patterns of specific combinations of object-types is important. Some of these include military (battlefield planning and strategy), ecology (tracking species and pollutant movements), homeland defense (looking for significant "events"), and transportation (road and network planning) [15, 25].

However, discovering MDCOPs poses several non-trivial challenges. First, current interest measures (i.e. the spatial prevalence measure) are not sufficient to quantify such patterns, so new composite interest measures must be created and formalized [20, 43]. Second, the set of candidate patterns grows exponentially with the number of object-types. Finally, since spatiotemporal datasets are huge, computationally efficient algorithms must be developed [46].

In contrast to the approaches proposed in the literature, the proposed interest measure and algorithms in [10] efficiently mine mixed groups of objects



(a) time slot t=0 (b) time slot t=1 (c) time slot t=2

(d) time slot t=3 (e) coach sketch

FIGURE 1.13: An example Hail Mary play in American football

(e.g MDCOPs) which are close in space and persistent (but not necessarily close) in time.

1.8 Summary

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. We conclude by identifying areas of research in spatial and spatiotemporal data mining that require further investigation. The current research focus is mostly concentrated on developing algorithms that model spatial and spatiotemporal autocorrelations and constraints. Spatiotemporal data mining is

still largely an unexplored territory; further research is especially needed for mining trajectory data and streaming data. Further research is also needed to scale these algorithms for large spatiotemporal datasets. Other important issues that need immediate attention include how to validate the hypotheses generated by spatial data mining algorithms and how to generate actionable knowledge.

1.9 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 Xiaobin Ma, Prof. Hui Xiong, Prof. Jin Soung Yoo, Dr. Qingsong Lu, Dr. Baris Kazar, Betsy George, 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] 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.

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

[3] C. Ambroise, V.M. Dang, and G. Govaert. Clustering of spatial data by the em algorithm. In *geoENV I - Geostatistics for Environmental Applications, Quantitative Geology and Geostatistics (Vol. 9)*, pages 493–504, 1997.

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

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

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

[7] J.E. Besag. Spatial Interaction and Statistical Analysis of Latice Systems. *Journal of Royal Statistical Society, Ser. B (Publisher: Blackwell Publishers)*, 36:192–236, 1974.

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

[9] Mete Celik, Shashi Shekhar, Betsy George, James P. Rogers, and James A. Shine. Discovering and quantifying mean streets: A summary of results. Technical Report 025, University of Minnesota, 07 2007.

[10] Mete Celik, Shashi Shekhar, James P. Rogers, James A. Shine, and Jin Soung 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.

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

[12] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2000.

[13] John E. Eck and et al. Mapping crime: Understanding hot spots. *US National Institute of Justice (http://www.ncjrs.gov/pdffiles1/nij/209393.pdf)*, 2005.

[14] A. Getis and J.K. Ord. Local spatial statistics: An overview. In *Spatial Analysis: Modelling in a GIS Environment*, pages 261–277. GeoInformation International, Cambridge, England, 1996.

[15] R. Guting and Markus Schneider. *Moving Object Databases*. Morgan Kaufmann, 2005.

[16] 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.

[17] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[18] 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.

[19] Tianming Hu and Yuan Sung. Clustering spatial data with a hybrid em approach. *Pattern Anal. Appl.*, 8(1):139–148, 2005.

[20] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering co-location patterns from spatial datasets: A general approach. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(12):1472–1485, 2004.

[21] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[22] Yonhong Jhung and Philip 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.

[23] Wen Jin, Yuelong Jiang, Weining Qian, and Anthony K. H. Tung. Mining outliers in spatial networks. In *DASFAA*, pages 156–170, 2006.

[24] 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.

[25] 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.

[26] Ned 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.

[27] S. Li. A Markov Random Field Modeling. *Computer Vision (Publisher: Springer Verlag*, 1995.

[28] Xiaolei Li, Jiawei Han, Jae-Gil Lee, and Hector Gonzalez. Traffic density-based discovery of hot routes in road networks. In *10th International Symposium on Spatial and Temporal Databases*, pages 441–459, Boston, MA, 2007.

[29] Anselin Luc. Exploratory Spatial Data Analysis and Geographic Information Systems. In M. Painho, editor, *New Tools for Spatial Analysis*, pages 45–54, 1994.

[30] Anselin Luc. Local Indicators of Spatial Association: LISA. *Geographical Analysis*, 27(2):93–115, 1995.

[31] Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[32] Atsuyuki Okabe, KeiIchi Okunuki, and Shino Shiode. The sanet toolbox: New methods for network spatial analysis. *Transactions in GIS*, 10(4):535–550, 2006.

[33] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[34] Jerry H. Ratcliffe. The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, 5(1):05–23, 2004.

[35] John A. Richards and Xiuping Jia. *Remote Sensing Digital Image Analysis*. Springer, New York, 1999.

[36] S.A. Roach. *The Theory of Random Clumping*. Methuen, London, 1968.

[37] 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.

[38] John F. Roddick, Kathleen Hornsby, and Myra Spiliopoulou. An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In *TSDM*, volume 2007 of *Lecture Notes in Computer Science*, pages 147–164. Springer, 2000.

[39] S. Shekhar and S. Chawla. Spatial databases: A tour. *Prentice Hall (ISBN 0-7484-0064-6)*, 2002.

[40] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. *Proc. of Spatio-temporal Symposium on Databases*, 2001.

[41] S. Shekhar, C.T. 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.

[42] S. Shekhar, Paul R. Schrater, Ranga R. Vatsavai, Weili Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4(2), 2002.

[43] Shashi Shekhar, Yan Huang, and Hui Xiong. Discovering spatial co-location patterns: A summary of results. In *7th International Symp. on Spatial and Temporal Databases (SSTD)*, L.A., CA, 2001.

[44] S. Shiode and A. Okabe. Network variable clumping method for analyzing point patterns on a network. In *Unpublished paper presented at the Annual Meeting of the Associations of American Geographers*, Philadelphia, Pennsylvania, 2004.

[45] A. H. Solberg, Torfinn Taxt, and Anil 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.

[46] SSTDM06. First international workshop on spatial and spatio-temporal data mining (sstdm). *In Conjunction with the 6th IEEE International Conference on Data Mining (ICDM 2006)*, 2006.

[47] P. Stolorz, H. Nakamura, E. Mesrobian, R.R. Muntz, E.C. Shek, J.R. Santos, J. Yi, K. Ng, S.Y. Chien, R. Mechoso, and J.D. 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.

[48] A.H. Strahler. The use of prior probabilities in maximum likelihood classificaiton of remote sensing data. *Remote Sensing of Environment*, 10:135–163, 1980.

[49] W.R. Tobler. *Cellular Geography, Philosophy in Geography.* Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[50] A.K.H. Tung, J. Hou, and Jiawei Han. Spatial clustering in the presence of obstacles. *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 359–367, 2001.

[51] V. Varnett and T. Lewis. *Outliers in Statistical Data.* John Wiley, 1994.

[52] 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.

[53] Xueping Zhang, Jiayao Wang, Fang Wu, Zhongshan Fan, and Xiaoqing 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.