43

Spatial Data Mining

Shashi Shekhar[1], Pusheng Zhang[1], and Yan Huang[1]

University of Minnesota

Summary. Spatial Data Mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. This chapter provides an overview on the unique features that distinguish spatial data mining from classical Data Mining, and presents major accomplishments of spatial Data Mining research.

Key words: Spatial Data Mining, Spatial Autocorrelation, Location Prediction, Spatial Outliers, Co-location, Spatial Clustering

43.1 Introduction

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial Data Mining (Roddick and Spiliopoulou, 1999, Shekhar and Chawla, 2003) 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 the National Aeronautics and Space Administration (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.

General purpose Data Mining tools, such as SPSS Clementine, Statistica Data Miner, IBM Intelligent Miner, and SAS Enterprise Miner, are designed to analyze

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_43,

large commercial databases. However, extracting interesting and useful patterns from spatial data sets 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 spatial data that preclude the use of general purpose Data Mining algorithms are: rich data types(e.g., extended spatial objects), implicit spatial relationships among the variables, observations that are not independent, and spatial autocorrelation among the features. In this chapter we focus on the unique features that distinguish spatial Data Mining from classical Data Mining, and present major accomplishments of spatial data mining research, especially regarding predictive modeling, spatial outlier detection, spatial co-location rule mining, and spatial clustering.

43.2 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 attribute and spatial attribute. 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 (Bolstad, 2002). 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, is_instance_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 (Quinlan, 1993, Barnett and Lewis, 1994, Agrawal and Srikant, 1994, Jain and Dubes, 1988). 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.

Statistical models (Cressie, 1993) 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

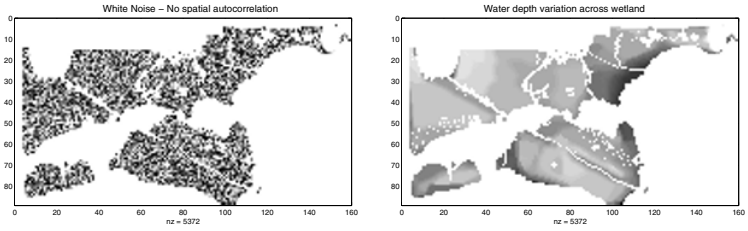
Table 43.1. Relationships among Non-spatial Data and Spatial Data

Non-spatial Relationship	*Spatial Relationship*
Arithmetic	Set-oriented: union, intersection, membership, $\cdots$
Ordering	Topological: meet, within, overlap, $\cdots$
Is_instance_of	Directional: North, NE, left, above, behind, $\cdots$
Subclass_of	Metric: e.g., distance, area, perimeter, $\cdots$
Part_of	Dynamic: update, create, destroy, $\cdots$
Membership_of	Shape-based and visibility

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 $\frac{1}{2*\sqrt{density}}$, 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 analyses, 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 (Cressie, 1993), which is an inherent characteristics of spatial data sets. Geostatistics provides a set of statistics tools, such as kriging (Cressie, 1993) to the interpolation of attributes at unsampled locations.



(a) Attribute with an Independent Identical Distribution

(b) Attribute with Spatial Autocorrelation

Fig. 43.1. Attribute Values in Space with Independent Identical Distribution and 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 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"* (Tobler, 1979). In spatial statistics, an area within statistics devoted to the analysis of spatial data, this property is called **spatial autocorrelation**. For example, Figure 43.1 shows the value distributions of an attribute in a spatial framework for an independent identical distribution and a distribution with 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., 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.

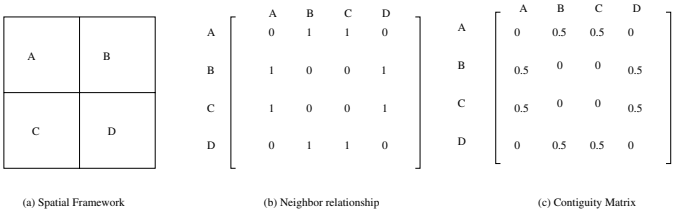


Fig. 43.2. A Spatial Framework and Its Four-neighborhood Contiguity Matrix.

Figure 43.2(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 43.2(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 43.2(c). Other contiguity matrices can be designed to model neighborhood relationships 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 (Warrender and Augusteijn, 1999). In spatial statistics, spatial autocorrelation is quantified using measures such as Ripley's K-function and Moran's I (Cressie, 1993).

In the rest of the chapter, we present case studies of the discovering four important patterns for spatial Data Mining: spatial outliers, spatial co-location rules, predictive models, and spatial clusters.

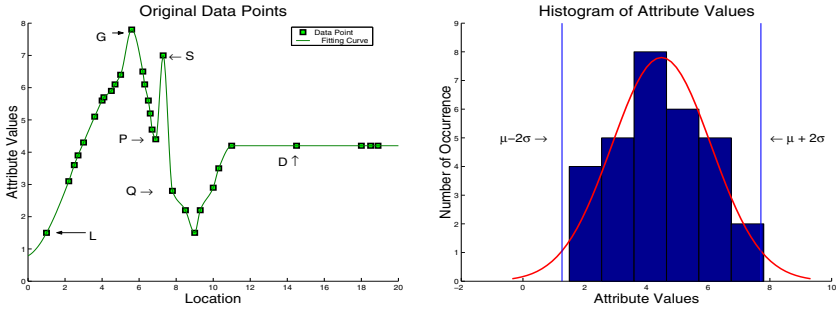
43.3 Spatial Outliers

Outliers have been informally defined as observations in a dataset which appear to be inconsistent with the remainder of that set of data (Barnett and Lewis, 1994), or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism (Hawkins, 1980). The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud, athlete performance analysis, voting irregularity, 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, including transportation, ecology, public safety, public health, climatology, and location-based services.

A spatial outlier 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 We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 43.3(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 43.3(b). This test assumes a normal distribution for attribute val-

ues. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.



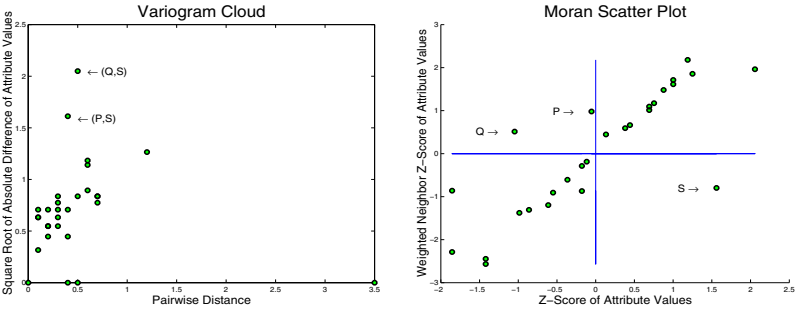
(a) An Example Dataset (b) Histogram

Fig. 43.3. A Dataset for Outlier Detection.

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 (Anselin, 1994) are a representative technique from the quantitative family.

A variogram-cloud (Cressie, 1993) 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 43.4(a) shows a variogram cloud for the example dataset shown in Figure 43.3(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) Variogram cloud (b) Moran scatterplot

Fig. 43.4. Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers.

A Moran scatterplot (Anselin, 1995) 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 43.4(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 can be treated as spatial outliers.

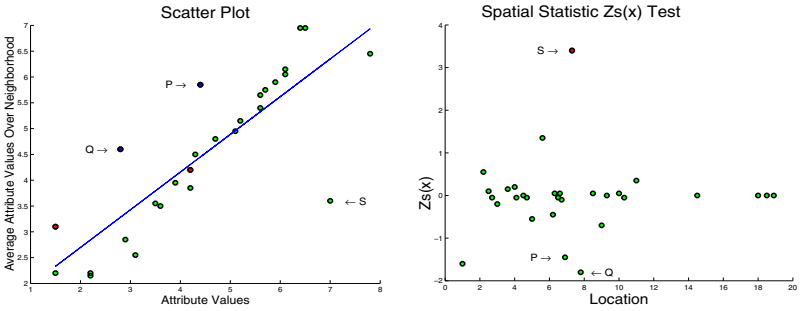
A scatterplot (Anselin, 1994) shows attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. A least square regression line is used to identify spatial outliers. A scatter sloping upward to the right indicates a positive spatial autocorrelation (adjacent 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 $\varepsilon = Y_p - (mX_p + b)$. Cases with standardized residuals, $\varepsilon_{standard} = \frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}$, 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 43.5(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the dataset in Figure 43.3(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 (Shekhar et al., 2003). 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 $Z_{s(x)} = |\frac{S(x)-\mu_s}{\sigma_s}| > \theta$. 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 43.5(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.



(a) Scatterplot

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

Fig. 43.5. Scatterplot and Spatial Statistic $Z_{s(x)}$ to Detect Spatial Outliers.

43.4 Spatial Co-location Rules

Spatial 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., surface of the Earth. Examples of boolean spatial features include plant species, animal species, disease, crime, business types, climate disturbances, etc.

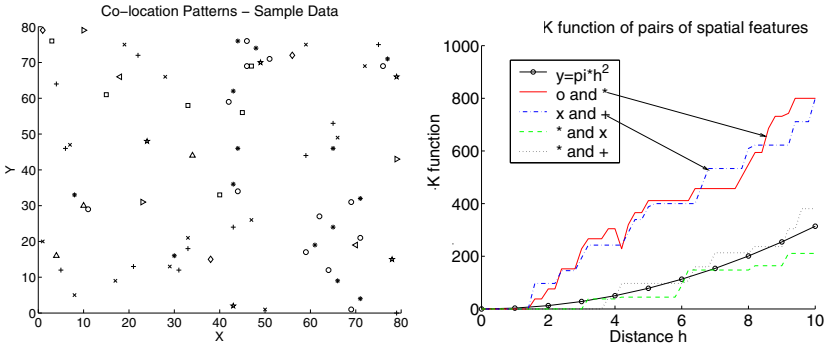


Fig. 43.6. a) A spatial dataset. Shapes represent different spatial feature types. b) Spatial features in sets {'+', '×'} and {'o', '*'} are co-located in (a) as shown by Ripley's K function

Spatial 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 43.6(a) shows a dataset consisting of instances of several boolean spatial features, each represented by a distinct shape. A careful visual review reveals two prevalent co-location patterns, i.e., ('+','×') and ('o','*'). These co-location patterns are also identified via a spatial statistical interest measure, namely Ripley's K function (Ripley, 1977). This interest measure has a value of πh^2 for a co-location pattern with a pair of spatial independent features for a given distance h. The co-location patterns ('+','×') and ('o','*') have much higher values of this interest measure relative to that of an independent pair illustrated in Figure 43.6 (b). Also note that we will refer to Ripley's K function as the K function in the rest of the chapter for simplicity.

Spatial co-location rule discovery is a process to identify co-location patterns from spatial datasets of instances of a number of boolean features. It is not trivial to adapt association rule mining algorithms to mine co-location patterns since instances of spatial features are embedded in a continuous space and share a variety of spatial

relations. Reusing association rule algorithm may require transactionizing spatial datasets, which is challenging due to the risk of transaction boundaries splitting co-location pattern instances across distinct transactions as illustrated in Figure 43.7, which uses cells of a rectangular grid to define transactions. Transaction boundaries split many instances of ('+','×') and ('o','*'), which are highlighted using ellipses. Transaction-based association rule mining algorithms need to be extended to correctly and completely identify co-locations defined by interest measures, such as the K function, whose values may be adversely affected by the split instances.

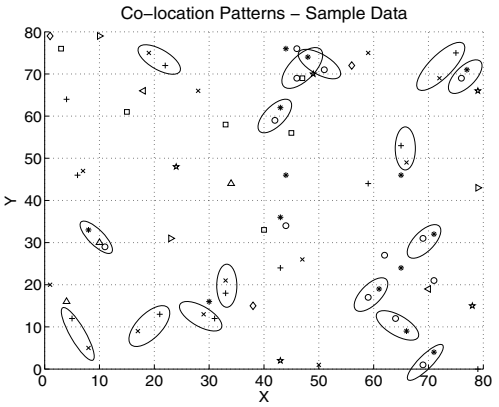


Fig. 43.7. Transactions split circled instances of co-location patterns

Approaches to discovering spatial co-location rules in the literature can be categorized into two classes, namely spatial statistics and association rules. In spatial statistics, interest measures such as the K function (Ripley, 1977) (and variations such as the L function (Cressie, 1993) and G function (Cressie, 1993)), mean nearest-neighbor distance, and quadrat count analysis (Cressie, 1993) are used to identify co-located spatial feature types. The K function for a pair of spatial features is defined as follows: $K_{ij}(h) = \lambda_j^{-1}$ E [number of type j event within distance h of a randomly chosen type i event], where λ_j is the density (number per unit area) of event j and h is the distance. Without edge effects, the K-function could be estimated by: $\hat{K}_{ij}(h) = \frac{1}{\lambda_i \lambda_j W} \sum_k \sum_l I_h(d(i_k, j_l))$, where $d(i_k, j_l)$ is the distance between the k'th location of type i and the l'th location of type j, I_h is the indicator function assuming value 1 if $d(i, j) \leq h$, and value 0 otherwise, and W is the area of the study region. $\lambda_j \times \hat{K}_{ij}(h)$ estimates the expected number of type j event instances within distance h of a type i event. The value of πh^2 is expected for a pair of independent pair of spatial features. The variance of the K function can be estimated by Monte Carlo simulation (Cressie, 1993) in general and by a close form equation under special circumstances (Cressie, 1993). Pointwise confidence intervals, e.g., 95%, can be estimated by simulating many realizations of the spatial patterns. The critical values for a test of independence could be calculated accordingly. In Figure 43.6 (b),

the K functions of the two pairs of spatial features, i.e., {'+','x'} and {'o','*'}, are well above the $y = \pi * h^2$ while the K functions of the other random two pairs of spatial features, i.e., {'*','x'} and {'*','+'}, are very close to complete spatial independence. This figure does not show the confidence band. We are not aware of the definition of the K function for subsets with 3 or more spatial features. Even if the definition is generalized, 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.

Data Mining approaches to spatial co-location mining can be broadly divided into transaction-based and spatial join-based approaches. The transaction based approaches focus on the creation of transactions over space so that an association rule mining algorithm (Agrawal and Srikant, 1994) can be used. Transactions over space have been defined by a reference-feature centric model (Koperski and Han, 1995) or a data-partition (Morimoto, 2001) approach. In the reference feature centric model (Koperski and Han, 1995), transactions are created around instances of a special user-specified spatial feature. The association rules are derived using the *apriori* (Agrawal and Srikant, 1994) algorithm. The rules found are all related to the reference feature. Generalizing this paradigm to the case where no reference feature is specified is non-trivial. Defining transactions around locations of instances of all features may yield duplicate counts for many candidate associations. Transactions in the data-partition approach (Morimoto, 2001) are formulated via grouping the spatial instances into disjoint partitions using different partitioning methods, which may yield distinct sets of transactions, which in turn yields different values of support of the co-location. Occasionally, imposing artificial disjoint transactions via space partitioning may undercount instances of tuples intersecting the boundaries of artificial transactions. In addition, to the best of our knowledge, no previous study has identified the relationship between transaction-based interest measures(e.g., support and confidence) (Agrawal and Srikant, 1994) and commonly used spatial interest measures(e.g., K function).

Spatial join-based approaches work directly with spatial data and include the cluster-then-overlay approaches (Estivill-Castro and Murray, 1998, Estivill-Castro and Lee, 2001) and instance join-based approach (Shekhar and Huang, 2001). The former treats every spatial attribute as a map layer and first identifies spatial clusters of instance data in each layer. Given X and Y as sets of layers, a clustered spatial association rule is defined as $X \Rightarrow Y(CS, CC\%)$, for $X \bigcap Y = \emptyset$, where CS is the clustered support, defined as the ratio of the area of the cluster (region) that satisfies both X and Y to the total area of the study region S, and $CC\%$ is the clustered confidence, which can be interpreted as $CC\%$ of areas of clusters (regions) of X intersect with areas of clusters(regions) of Y. The value of interest measures, e.g., clustered support and clustered confidence, depend on the choice of clustering algorithms from a large collection of choices (Han et al., 2001). To our knowledge, the relationship between these interest measures and commonly used spatial statistical measures(e.g., K function) is not yet established. In recent work (Huang et al., 2004), an instance join-based approach was proposed that uses join selectivity as the prevalence inter-

est measures and provided interpretation models by relating those to other interest measures, e.g., K function.

43.5 Predictive Models

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as fires, floods, droughts, vegetation diseases, and earthquakes. In this section we provide two spatial Data Mining techniques for predicting locations, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

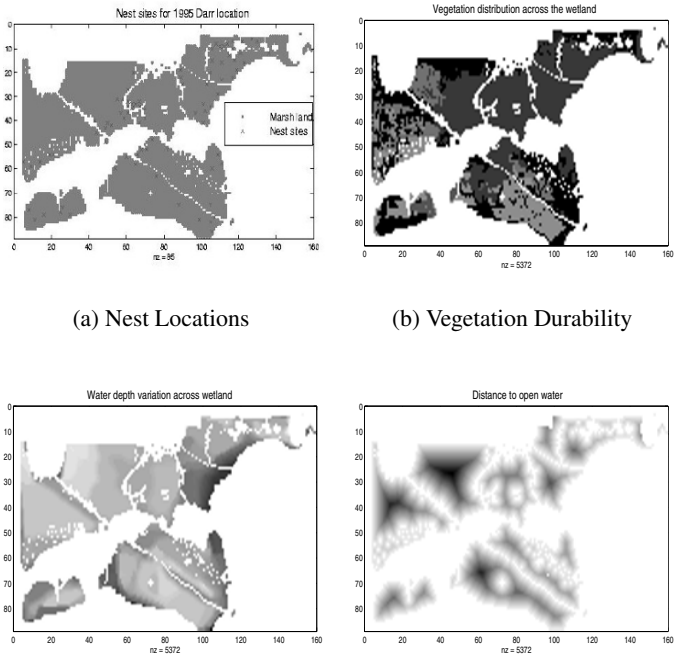
An Application Domain We begin by introducing an example to illustrate the different concepts related to location prediction in spatial Data Mining. We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio USA in order to *predict* the spatial distribution of a 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, the 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.

An important 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 this study we build a model using the 1995 Darr wetland data and then tested it 1995 Stubble wetland data. In the learning data, all the attributes are used to build the model and in the training data, one value is hidden, in our case the location of the nests. Using knowledge gained from the 1995 Darr data and the value of the independent attributes in the test data, we want to predict the location of the nests in 1995 Stubble data.

Modeling Spatial Dependencies Using the SAR and MRF Models Several previous studies (Jhung and Swain, 1996), (Solberg et al., 1996) 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. In this section, we present two models to model spatial dependency: the spatial autoregressive model(SAR) and Markov random field(MRF)-based Bayesian classifiers.

Spatial Autoregressive Model The spatial autoregressive model 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



(a) Nest Locations (b) Vegetation Durability

(c) Water Depth (d) Distance to Open Water

Fig. 43.8. (a) Learning dataset: The geometry of the Darr wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water.*

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 (Anselin, 1988). If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho Wy + X\beta + \varepsilon. \tag{43.1}$$

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 ρWy 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: 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. 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).

Markov Random Field-based Bayesian Classifiers Markov random field-based Bayesian classifiers estimate the classification model $\hat{f}_C$ using MRF and 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 (Li, 1995). 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:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{43.2}$$

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 (Besag, 1974).

A more detailed theoretical and experimental comparison of these methods can be found in (Shekhar et al., 2002). 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. On the other hand, 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, 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.

Experiments were carried out on the Darr and Stubble wetlands to compare 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 predictions are extremely localized, missing actual nests over a large part of the marsh lands.

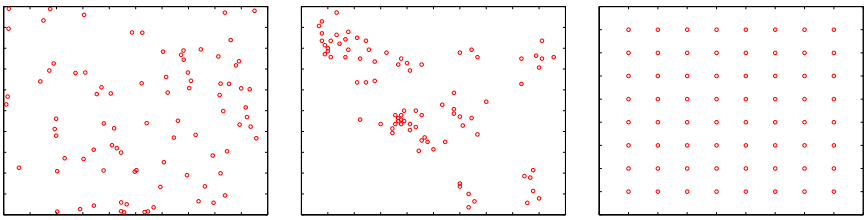
43.6 Spatial Clusters

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. For example, clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hot spot analysis is the process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas.

Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. The test is critical before proceeding with any serious clustering analyses.

Complete Spatial Randomness, Cluster, and Decluster In spatial statistics, the standard against which spatial point patterns are often compared is a completely spatially random point process, and departures indicate that the pattern is not distributed randomly in space. *Complete spatial randomness (CSR)* (Cressie, 1993) 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. However, patterns generated by a non-random process can be either cluster patterns(aggregated patterns) or decluster patterns(uniformly spaced patterns).

To illustrate, Figure 43.9 shows realizations from a completely spatially random process, a spatial cluster process, and a spatial decluster process (each conditioned to have 80 points) in a square. Notice in Figure 43.9 (a) that the complete spatial randomness pattern seems to exhibit some clustering. This is not an unrepresentive 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 (Cressie, 1993). Spatial clustering is more statistically significant when the data exhibit a cluster pattern rather than a CSR pattern or decluster pattern.



(a) CSR Pattern (b) Cluster Pattern (c) Decluster Pattern

Fig. 43.9. Illustration of CSR, Cluster, and Decluster Patterns

Several statistical methods can be applied to quantify deviations of patterns from a complete spatial randomness point pattern (Cressie, 1993). One type of descriptive statistics is based on quadrats (i.e., well defined area, often rectangle in shape). Usually quadrats of random location 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 (Cressie, 1993).

After the verification of the statistical significance of the spatial clustering, classical clustering algorithms (Han et al., 2001) can be used to discover interesting clusters.

43.7 Summary

In this chapter, we have focused on the features of spatial data mining that distinguish it from classical Data Mining. We have discussed major research accomplishments and techniques in spatial Data Mining, especially those related to four important output patterns: predictive models, spatial outliers, spatial co-location rules, and spatial clusters.

Acknowledgments

This work was supported in part by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

We are particularly grateful to our collaborators Prof. Vipin Kumar, Prof. Paul Schrater, Dr. Sanjay Chawla, Dr. Chang-Tien Lu, Dr. Weili Wu, and Prof. Uygar Ozesmi for their various contributions. We also thank Xiaobin Ma, Hui Xiong, Jin Soung Yoo, Qingsong Lu, Baris Kazar, and anonymous reviewers for their valuable feedbacks on early versions of this chapter.

References

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In *Proc. of Very Large Databases*.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer, Dordrecht, Netherlands.

Anselin, L. (1994). Exploratory Spatial Data Analysis and Geographic Information Systems. In Painho, M., editor, *New Tools for Spatial Analysis*, pages 45–54.

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

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley, 3rd edition edition.

Besag, J. (1974). Spatial Interaction and Statistical Analysis of Lattice Systems. *Journal of Royal Statistical Society: Series B*, 36:192–236.

Bolstad, P. (2002). *GIS Foundamentals: A Fisrt Text on GIS*. Eider Press.

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

Estivill-Castro, V. and Lee, I. (2001). Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. In *Proc. of the 6th International Conference on Geocomputation*.

Estivill-Castro, V. and Murray, A. (1998). Discovering Associations in Spatial Data - An Efficient Medoid Based Approach. In *Proc. of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Han, J., Kamber, M., and Tung, A. (2001). Spatial Clustering Methods in Data Mining: A Survey. In Miller, H. and Han, J., editors, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Huang, Y., Shekhar, S., and Xiong, H. (2004). Discovering Co-location Patterns from Spatial Datasets:A General Approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12).

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Jhung, Y. and Swain, P. H. (1996). Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75.

Koperski, K. and Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66*.

Li, S. (1995). A Markov Random Field Modeling. *Computer Vision*.

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

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Ripley, B. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society*, Series B 39:172–192.

Roddick, J.-F. and Spiliopoulou, M. (1999). A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. *SIGKDD Explorations 1(1): 34-38 (1999)*.

Shekhar, S. and Chawla, S. (2003). Spatial Databases: A Tour. *Prentice Hall (ISBN 0-7484-0064-6)*.

Shekhar, S. and Huang, Y. (2001). Co-location Rules Mining: A Summary of Results. In *Proc. of the 7th Int'l Symp. on Spatial and Temporal Databases*.

Shekhar, S., Lu, C., and Zhang, P. (2003). A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7(2).

Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., and Chawla, S. (2002). Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transaction on Multimedia*, 4(2).

Solberg, A. H., Taxt, T., and Jain, A. K. (1996). A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113.

Tobler, W. (1979). *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel.

Warrender, C. E. and Augusteijn, M. F. (1999). Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002.