Chapter 3

Spatial Data Mining

Shashi Shekhar[*], *Pusheng Zhang*[*], *Yan Huang*[*], *Ranga Raju Vatsavai*[*]

[*]Department of Computer Science and Engineering, University of Minnesota
4-192, 200 Union ST SE, Minneapolis, MN 55455

Abstract:
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 will discuss some of the accomplishments and research needs of spatial data mining in the following categories: location prediction, spatial outlier detection, co-location mining, and clustering.

Keywords:
Spatial data mining, Spatial autocorrelation, Location prediction, Spatial outliers, Spatial co-location, Clustering

3.1 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 [Stolorz *et al.*1995, Shekhar & Chawla2002] 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 Imagery and Mapping Agency (NIMA), 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 [Roddick & Spiliopoulou1999].

General purpose data mining tools like Clementine, See5/C5.0, and Enterprise Miner are designed for the purpose of analyzing large commercial databases. 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. 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, v) spatial autocorrelation among the features, and vi) non-linear interaction 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.

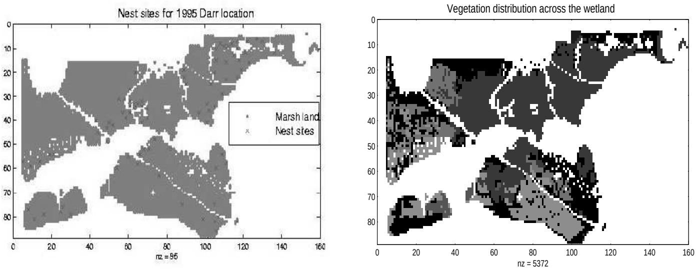
Now the question arises whether we really need to invent new algorithms or extend the existing approaches to explicitly model spatial properties and relationships. Although it is difficult to tell the direction of future research, for now it seems both approaches are gaining momentum. In this chapter we present major accomplishments in the emerging field of spatial data mining and applications, especially in the areas of outlier detection, spatial co-location rules, classification/prediction, and clustering techniques. The research needs for spatial data mining are also identified.

This chapter is organized as follows. In Section 3.2, we review major accomplishments in spatial data mining in the following four categories: location prediction, spatial outlier detection, spatial co-location rules, and spatial clustering. Section 3.2.1 presents extensions of classification and prediction techniques that model spatial context. Section 3.2.2 introduces spatial outlier detection techniques. In Section 3.2.3, we present a new approach, called co-location mining, which finds the subsets of features frequently-located together in spatial databases. Spatial clustering techniques are introduced in Section 3.2.4. Section 3.3 concludes the chapter with a discussion of research needs in spatial data mining.

3.2 Accomplishments

3.2.1 Location Prediction

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 provide two spatial data mining techniques, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).



(a) Nest Locations

(b) Vegetation Durability

Figure 3.1: (a) Learning data set: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland.

An Illustrative Application Domain

We now introduce an example to illustrate the different concepts 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, 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.

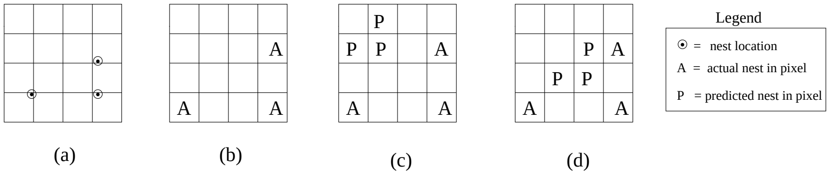
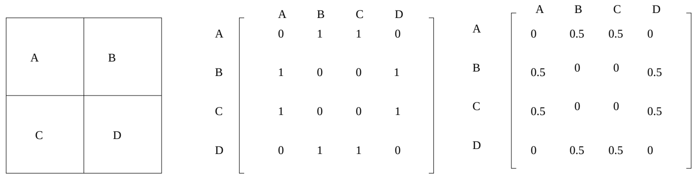


Figure 3.2: (a)The actual locations of nests, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another model. Prediction(d) is spatially more accurate than (c).

The fact that classical data mining techniques ignore spatial autocorrelation and spatial heterogeneity in the model-building process is one reason why these techniques do a poor job. A second, more subtle but equally important reason is related to the choice of the objective function to measure classification accuracy. For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. However, this measure may not be the most suitable in a spatial context. *Spatial accuracy*–how far the predictions are from the actuals–is as important in this application domain due to the effects of the discretizations of a continuous wetland into discrete pixels, as shown in Figure 3.2. Figure 3.2(a) shows the actual locations of nests and 3.2(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 3.2(c) and 3.2(d). Domain scientists prefer prediction 3.2(d) over 3.2(c), since the predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 3.2(c) and 3.2(d), and a measure of spatial accuracy is needed to capture this preference.



(a) Spatial Framework

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 1 \\ C & 1 & 0 & 0 & 1 \\ D & 0 & 1 & 1 & 0 \end{array}$$

(b) Neighbor relationship

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0.5 & 0.5 & 0 \\ B & 0.5 & 0 & 0 & 0.5 \\ C & 0.5 & 0 & 0 & 0.5 \\ D & 0 & 0.5 & 0.5 & 0 \end{array}$$

(c) Contiguity Matrix

Figure 3.3: A spatial framework and its four-neighborhood contiguity matrix.

Modeling Spatial Dependencies Using the SAR and MRF Models

Several previous studies [Jhung & Swain1996], [Solberg, Taxt, & Jain1996] 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. 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 3.3(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 3.3(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 3.3(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 [Warrender & Augusteijn1999].

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[Anselin1988]. If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{3.1}$$

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).

Markov Random Field-based Bayesian Classifiers

Markov Random Field (MRF) based Bayesian classifiers estimate 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 [Li1995]. 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:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{3.2}$$

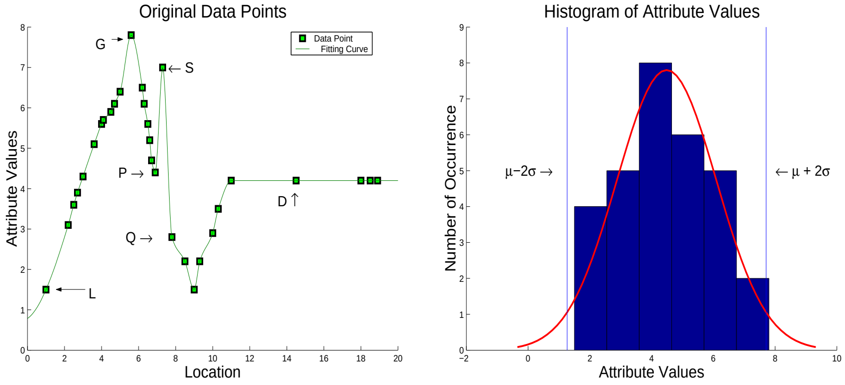
The solution procedure can estimate $Pr(l_i|L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X|l_i, L_i)$ can be estimated using kernel functions from the observed values in the training data set. 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 [Besag1974].

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. 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 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.

3.2.2 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 [Barnett & Lewis1994], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism [Hawkins1980]. 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 data set 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



(a) An Example Data Set

(b) Histogram

Figure 3.4: A Data Set for Outlier Detection.

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 spatial outlier [Shekhar, Lu, & Zhang2001] is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age.

Illustrative Examples and Application Domains

We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 3.4(a), the X-axis is the location of data points in one-dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point and fit the distribution model to the values of the non-spatial attribute. 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 3.4(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.

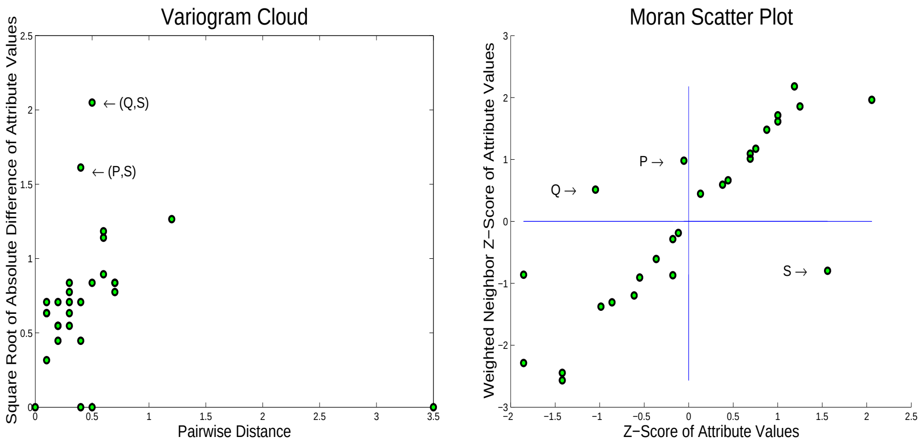
As another example, we use a spatial database consisting of measurements from the Minneapolis-St. Paul freeway traffic sensor network. The sensor network includes about nine hundred stations, each of which contains one to four loop detectors, depending on the number of lanes. Sensors embedded in the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

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 [Luc1994] 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 data sets 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 data set non-spatially. Figure 3.5(a) shows a variogram cloud for the example data set shown in Figure 3.4(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 [Luc1995] 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 3.5(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values sur-



(a) Variogram cloud

(b) Moran scatterplot

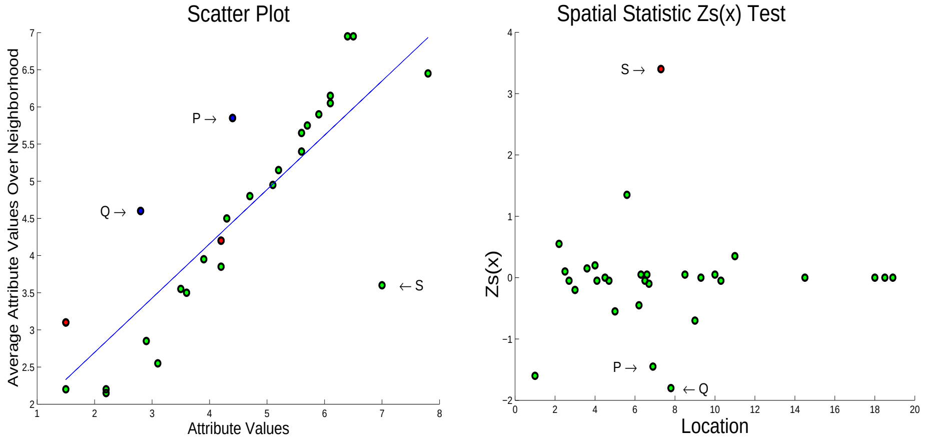
Figure 3.5: Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers.

rounded 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 [Luc1994] 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 3.6(a), a scatter plot shows the attribute values plotted against the average of the attribute values in neighboring areas for the data set in Figure 3.4(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 $f(x)$



(a) Scatterplot

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

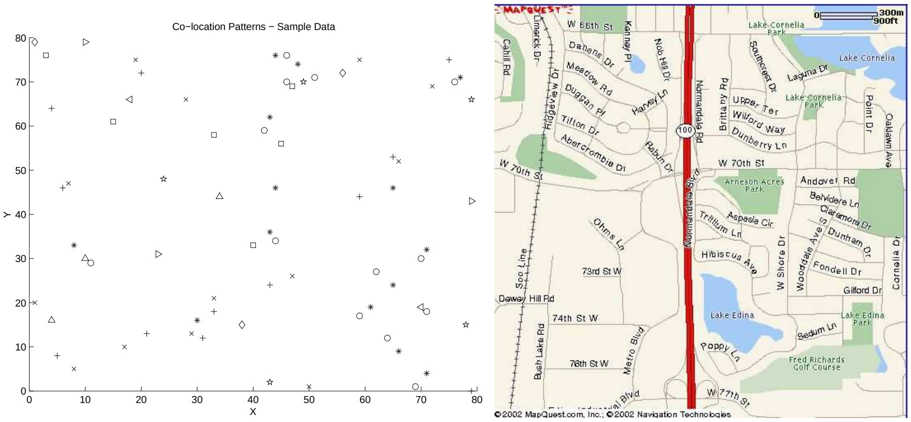
Figure 3.6: Scatterplot and Spatial Statistic $Z_{s(x)}$ to Detect Spatial Outliers.

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

3.2.3 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 3.7: 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 3.7(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 [Agrawal & Srikant1994] because of the lack of transactions. In market basket data sets, 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.

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 [Cressie1993]. 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.

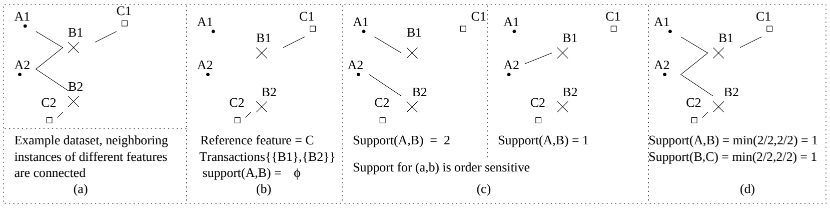

A1
A2
B1
B2
C1
C2
Example dataset, neighboring instances of different features are connected
(a)
Reference feature = C
Transactions{{B1},{B2}}
support(A,B) = ϕ
(b)
Support(A,B) = 2
Support(A,B) = 1
Support for (a,b) is order sensitive
(c)
Support(A,B) = min(2/2,2/2) = 1
Support(B,C) = min(2/2,2/2) = 1
(d)

Figure 3.8: 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 [Agrawal & Srikant1994] can be used. Transactions in space can use a reference-feature centric [Koperski & Han1995] approach or a data-partition [Morimoto2001] approach. The **reference feature centric model** is based on the choice of a reference spatial feature [Koperski & Han1995] 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 [Koperski & Han1995]. 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 3.8(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 3.8(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 [Morimoto2001] 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 3.8 c) shows two possible partitions for the dataset of Figure 3.8 a), along with the supports for co-location (A, B).

Table 3.1: Interest Measures for Different Models

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 $C_1)$

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 3.8 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 3.8 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 3.1. 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 [Shekhar & Huang2001] for details).

3.2.4 Spatial Clustering

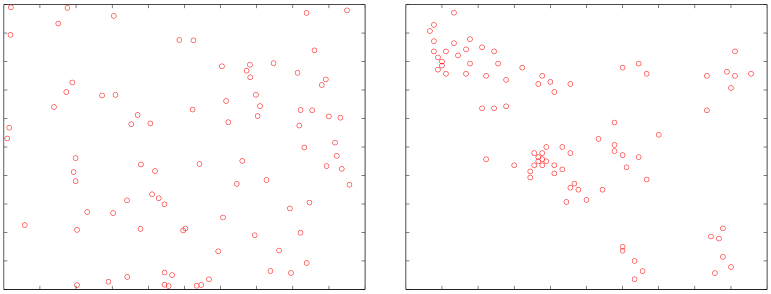
Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. For example, clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hotspot 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 hotspots in order to take preventive strategies such as deploying saturation patrols in hotspot areas.

3.2.5 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 any serious clustering analyses.

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) [Cressie1993] 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 3.9, which shows realizations from a completely spatially random process and from a spatial cluster process respectively (each conditioned to have 85 points in a unit square).



(a) CSR Pattern

(b) Clustered Pattern

Figure 3.9: Complete Spatial Random (CSR) and Spatially Clustered Patterns

Notice from Figure 3.9 (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 [Cressie1993]. True clustering is shown in Figure 3.9 (b), which should be compared with Figure 3.9 (a).

Several statistical methods [Cressie1993] 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.

3.2.6 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 interconnectivity (CHAMELEON), clustering using representatives (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 data sets. 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 [Han, Kamber, & Tung2001].

3.3 Research Needs

In this chapter we have presented the major research achievements and techniques which have emerged from spatial data mining, especially for finding location prediction, spatial outliers, co-location rules, and spatial clusters. We conclude by identifying areas of research in spatial data mining that require further investigation.

3.3.1 Location Prediction

Further exploration of measures of location predictive accuracy is needed. In traditional data mining, precision and recall are two major measures for predictive accuracy. However, predictive models in spatial data mining do not incorporate measures of spatial accuracy. There is a research need to investigate proper measures for location prediction.

The determination of contiguity matrix is an expensive task for large spatial datasets. Parallel processing of predictive models such as SAR and MRF could be explored to further improve performance.

3.3.2 Spatial Outlier Detection

Most spatial outlier detection algorithms are designed to detect spatial outliers using a single non-spatial attribute from a data set. Spatial outlier detection with multiple non-spatial attributes are still under investigation. For multiple attributes, the definition of spatial neighborhood will be the same, but the neighborhood aggregate function, comparison function, and statistic test function need to be redefined. The key challenge is to define a general distance function in a multi-attribute data space.

Many visual representations have been proposed for spatial outliers. However, there is a research need for effective representations to facilitate the visualization of spatial relationships while highlighting spatial outliers. For instance, in variogram cloud and scatterplot visualizations, the spatial relationship between a single spatial outlier and its neighbors is not obvious. It is necessary to transfer the information back to the original map to check the neighbor relationships. Since a single spatial outlier tends to flag not only the spatial location of local instability but also its neighboring locations, it is important to group flagged locations and identify real spatial outliers from the group in the post-processing step.

3.3.3 Co-location

Preliminary results show that the co-location pattern mining problem can be formulated using neighborhoods instead of transactions. Interest measures, e.g., participation index, and the conditional probability of a co-location can be defined using neighborhoods. The participation index is a monotonically non-decreasing interest measure and can be used to reduce computation cost by filtering out co-location patterns based on user-defined thresholds. The co-location miner algorithm can exploit the participation index-based filter to mine co-location patterns in small spatial datasets, e.g. a metropolitan road map. However, to enable wider use, co-location mining techniques need to address the following two significant issues.

First, there is a need for an independent measure of the quality of co-location patterns due to the unsupervised learning. In order to achieve this, we need to compare co-location modelx with dedicated spatial statistical measures, such as Ripley's K function; characterize the distribution of the participation index interest measure under spatial complete randomness using Monte Carlo simulation; and develop a classical statistical interpretation of co-location rules to compare the rules with other patterns in unsupervised learning. Second, the co-location miner may not be able to discover high-confidence, low-prevalence rules due to its reliance on prevalence-based pruning. High-confidence, low-prevalence ($HCLP$) co-locations are useful in many applications where the numbers of occurrences of various boolean spatial features vary by orders of magnitude. The prevalence based pruning used in the co-location miner makes it hard to mine $HCLP$ co-location rules. Models and efficient mining algorithms are needed to retain and mine $HCLP$ co-location rules.

3.3.4 Other Research Needs

Other research needs include detecting the shapes of spatial phenomena and performance tuning. Shape detection identifies changes in shapes of a spatial phenomenon, e.g., variations in the shape and extent of the hot water area in the eastern tropical region of the Pacific during El Nino years. Algorithms are needed to explore shape detection for spatial phenomena. Due to the large-scale and spatial autocorrelation nature of spatial data, efficient and scalable algorithms are needed to improve performance of spatial data mining processes.

3.4 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. 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, and Baris Kazar for their valuable feedback on early versions of this chapter. We would also like to express our thanks to Kim Koffolt for improving the readability of this chapter.

Bibliography

[Agrawal & Srikant1994] Agrawal, R., and Srikant, R. 1994. Fast algorithms for Mining Association Rules. In *Proc. of Very Large Databases.*

[Anselin1988] Anselin, L. 1988. *Spatial Econometrics: methods and models.* Dordrecht, Netherlands: Kluwer.

[Barnett & Lewis1994] Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data*. John Wiley, 3rd edition edition.

[Besag1974] Besag, J. 1974. Spatial Interaction and Statistical Analysis of Latice Systems. *Journal of Royal Statistical Society, Ser. B (Publisher: Blackwell Publishers)* 36:192–236.

[Cressie1993] Cressie, N. 1993. *Statistics for Spatial Data (Revised Edition).* New York: Wiley.

[Han, Kamber, & Tung2001] Han, J.; Kamber, M.; and Tung, A. 2001. Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, J., eds., *Geographic Data Mining and Knowledge Discovery.* Taylor and Francis.

[Hawkins1980] Hawkins, D. 1980. *Identification of Outliers.* Chapman and Hall.

[Jhung & Swain1996] 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 & Han1995] 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.*

[Li1995] Li, S. 1995. A Markov Random Field Modeling. *Computer Vision (Publisher: Springer Verlag.*

[Luc1994] Luc, A. 1994. Exploratory Spatial Data Analysis and Geographic Information Systems. In Painho, M., ed., *New Tools for Spatial Analysis*, 45–54.

[Luc1995] Luc, A. 1995. Local Indicators of Spatial Association: LISA. *Geographical Analysis* 27(2):93–115.

[Morimoto2001] Morimoto, Y. 2001. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Roddick & Spiliopoulou1999] 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 *et al.*2002] 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).

[Shekhar, Lu, & Zhang2001] Shekhar, S.; Lu, C.; and Zhang, P. 2001. 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*.

[Shekhar & Chawla2002] Shekhar, S., and Chawla, S. 2002. A Tour of Spatial Databases. *Prentice Hall (ISBN 0-7484-0064-6).*

[Shekhar & Huang2001] Shekhar, S., and Huang, Y. 2001. Co-location Rules Mining: A Summary of Results. *Proc. of Spatio-temporal Symposium on Databases*.

[Solberg, Taxt, & Jain1996] 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.

[Stolorz *et al.*1995] Stolorz, P.; Nakamura, H.; Mesrobian, E.; Muntz, R.; Shek, E.; Santos, J.; Yi, J.; Ng, K.; Chien, S.; Mechoso, R.; and Farrara, J. 1995. 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.*

[Warrender & Augusteijn1999] 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.