Spatial Data Mining Research by the Spatial Database Research Group, University of Minnesota

Shashi Shekhar and Ranga Raju Vatsavai
Spatial Database Research Group
Department of Computer Science and Engineering
EE/CS 4-192, 200 Union Street, SE., Minneapolis, MN 55455.
[$shekhar|vatsavai$]@cs.umn.edu
http://www.cs.umn.edu/research/shashi-group/

Abstract

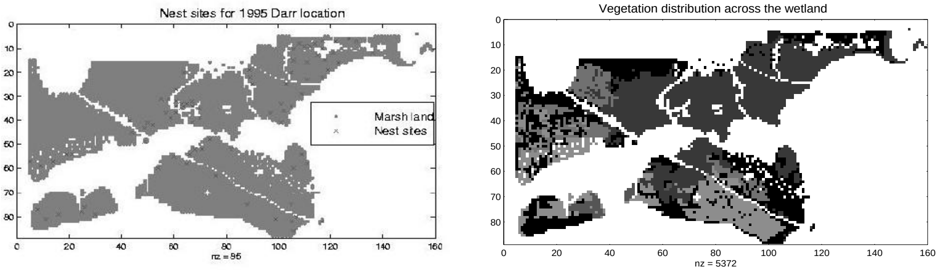
Explosive growth in geospatial data and the emergence of new spatial technologies emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. In this paper we describe the ongoing spatial data mining research by the Spatial Database Research Group, University of Minnesota. We discuss several computationally efficient and scalable techniques for analyzing large geospatial data sets and their applications in location prediction, spatial outliers detection and co-location association rules mining.

Keywords: co-location mining, spatial outliers, spatial context, SAR, MRF

1 Introduction

Researchers in the Spatial Database Research Group [22], University of Minnesota, have recently focussed their reserach in the field of spatial data mining, a field whose importance is growing with the increasing incidence and importance of large geo-spatial datasets such as maps, repositories of remote-sensing images, and the decennial census. Applications of spatial data mining can be found in location-based services in the M(mobile)-commerce industry, in the military (inferring enemy tactics such as Flank attack), at NASA (studying the climatological effects of El Nino, land-use classification and global change using satellite imagery), at the National Institute of Health (predicting the spread of disease), at the National Imagery and Mapping Agency (creating high resolution three-dimensional maps from satellite imagery), at the National Institute of Justice (finding crime hot spots), and in transportation agencies (detecting local instability in traffic).

The differences between classical and spatial data mining are similar to the differences between classical and spatial statistics. First, spatial data is embedded in a continuous space, whereas classical datasets are often discrete. Second, spatial patterns are often local whereas classical data mining techniques often focus on global patterns. Finally, one of the common assumptions in classical statistical analysis is that data samples are independently generated. When it comes to the analysis of spatial data, however, the assumption about the independence of samples is generally false because spatial data tends to be highly self correlated. For example, people with similar


Nest sites for 1995 Darr location
Marshland
Nest sites
nz = 95
Vegetation distribution across the wetland
nz = 5372
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
120
140
160

(a) Nest Locations

(b) Vegetation Durability

Figure 1: (a) Learning dataset: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland

characteristics, occupation, and background tend to cluster together in the same neighborhoods. In spatial statistics this tendency is called spatial autocorrelation. Ignoring spatial autocorrelation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the dataset. Thus classical data mining algorithms often perform poorly when applied to spatial datasets. Thus new methods are needed to analyze spatial data to detect spatial patterns.

The roots of spatial data mining lie in spatial statistics, spatial analysis, geographic information systems, machine learning, image analysis, and data mining. The main contributions made by computer science researchers to this area include algorithms and data-structures that can scale up to massive (terabytes to petabytes) datasets as well as the formalization of newer spatio-temporal patterns (e.g. colocations) which were not explored by other research communities due to computational complexity. Spatial data mining projects in our group at the Department of Computer Science include location prediction, detection of spatial outliers, and discovery of spatial co-location patterns.

Location prediction is concerned with the discovery of a model to infer locations of a spatial phenomenon from the maps of other spatial features. For example, ecologists build models to predict habitats for endangered species using maps of vegetation, water bodies, climate, and other related species. Figure 1 shows maps of nest location and vegetation durability to build a location prediction model for red-winged blackbirds in the Darr and Stubble wetlands on the shores of Lake Eries in Ohio. Classical data mining techniques yield weak prediction models as they do not capture the auto-correlation in spatial datasets. We provided a formal comparison of diverse techniques from spatial statistics (e.g. spatial autoregression) as well as image classification (e.g. Markov Random Field-based Bayesian classifiers) and developed scalable algorithms for these techniques [28].

Spatial outliers are significantly different from their neighborhood even though they may not be significantly different from the entire population. For example, a brand new house in an old neighborhood of a growing metropolitan area is a spatial outlier. Figure 7 shows another use of spatial outliers in traffic measurements for sensors on I-35W (north bound) for a 24 hour time period. Sensor 9 seems to be a spatial outlier and may be a bad sensor. Note that the figure also shows three clusters of sensor behaviors namely, morning rush hour, evening rush hour, and busy day-time. Spatial statistics tests for detecting spatial outliers do not scale up to massive datasets, such as the Twin Cities traffic dataset measured at thousands of locations in 30-second intervals and archived for years. We generalized spatial statistics tests to spatio-temporal datasets and developed scalable algorithms [29] for detecting spatial ouliers in massive traffic datasets.

The co-location pattern discovery process finds frequently co-located subsets of spatial event types given a map (see Figure 2) of their locations. For example, the analysis of the habitats of animals and plants may identify the co-locations of predator-prey species, symbiotic species, and fire events with fuel, ignition sources etc. Readers may find it interesting to analyze the map in Figure 2 to find the co-location patterns. (There are two co-location patters of size 2 in this map.) Our group has provided one of the most natural formulations as well as the first algorithms [26] for discovering co-location patterns from large spatial datasets and applying them to climatology data from NASA.

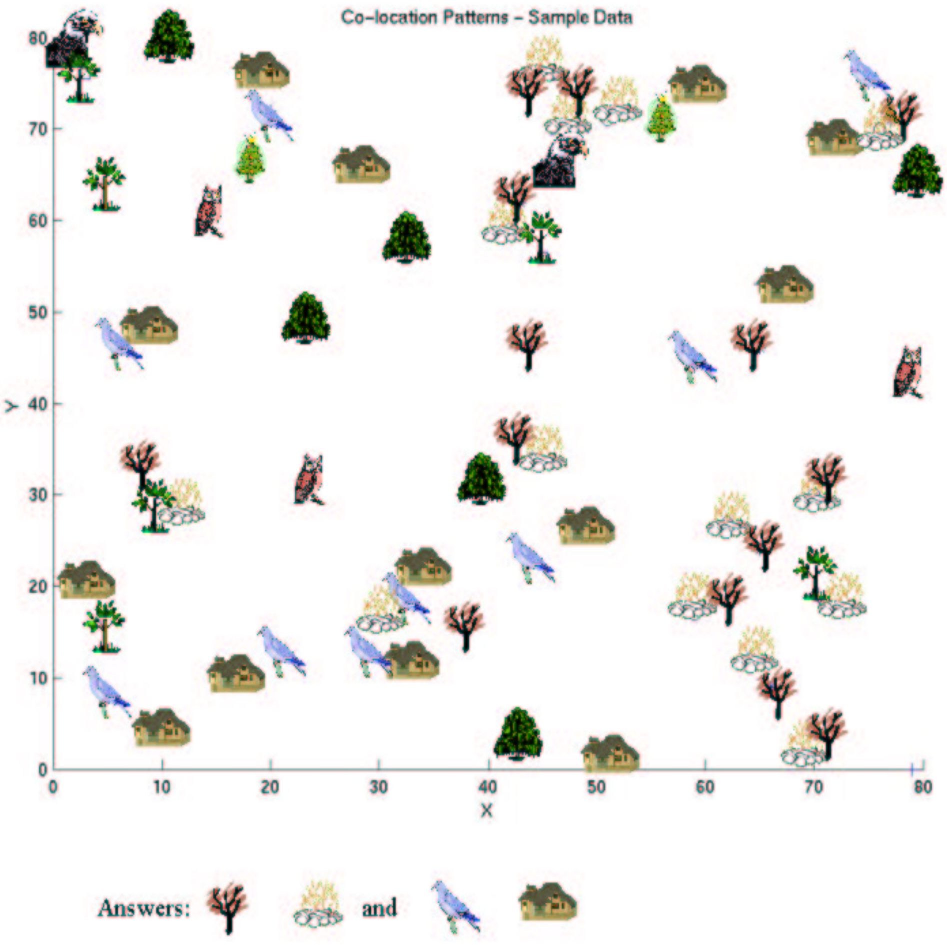
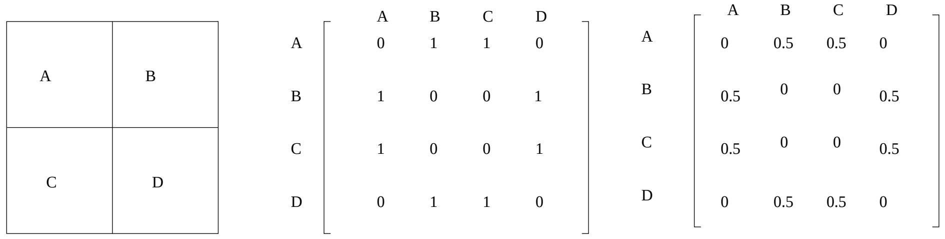


Figure 2: Sample co-location patterns

Paper Organization

We describe each of these techniques in the following sections. In Section 2, we present SAR and MRF techniques for predicting bird nest location using wetland datasets. In Section 3, we introduce spatial outlier detection techniques and their use in finding spatio-temporal outliers in traffic data. Section 4 presents a new approach called co-location mining, which finds the subsets



(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: A spatial framework and its four-neighborhood contiguity matrix

of features frequently-located together in spatial databases. Finally, we conclude with a summary of techniques and results.

2 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, 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) and analyze their performance in an example case, the prediction of the location of bird nests in the Darr and Stubble wetlands.

2.1 Modeling Spatial Dependencies Using the SAR and MRF Models

Several previous studies [13], [30] 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(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(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 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 [31].

2.2 Logistic Spatial Autoregression 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[2]. If the dependent values y_i are related to each other, then the regression equation can be modified as

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

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable. 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: 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). We compare SAR with linear regression for predicting nest location in Section 4.

Solution Procedures

The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package[1], which implements a Bayesian approach using sampling-based Markov Chain Monte Carlo (MCMC) methods[21]. Without any optimization, likelihood-based estimation would require $O(n^3)$ operations. Recently [24], [25], and [15] have proposed several efficient techniques to solve SAR. The techniques studied include divide and conquer, and sparse matrix algorithms. Improved performance is obtained by using LU decompositions to compute the log-determinant over a grid of values for the parameter ρ by restricting it to $[0, 1]$.

2.3 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 [16]. 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, constitute an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

[1] We would like to thank James Lesage (http://www.spatial-econometrics.com/) for making the matlab toolbox available on the web.

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

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{2}$$

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 [5].

Solution Procedures

Solution procedures for the MRF Bayesian classifier include stochastic relaxation [9], iterated conditional modes [4], dynamic programming [8], highest confidence first [7] and graph cut [6]. We followed the approach suggested in[6], where it is shown that the maximum a posteriori estimate of a particular configuration of an MRF can be obtained by solving a suitable min-cut multiway graph partitioning problem. Here we briefly provide theoretical and experimental comparisions; more details can be found in [28].

2.4 Comparison of SAR and MRF Using a Probabilistic Framework

We use a simple probabilistic framework to compare SAR and MRF in this section. We will assume that classes $l_i \in (c_1, c_2, \ldots, c_M)$ are discrete and that the class label estimate $\hat{f}_C(s_i)$ for location s_i is a random variable. We also assume that feature values (X) are constant since there is no specified generative model. Model parameters for SAR are assumed to be constant, (i.e., β is a constant vector and ρ is a constant number). Finally, we assume that the spatial framework is a regular grid.

We first note that the basic SAR model can be rewritten as follows:

$$y = X\beta + \rho W y + \epsilon$$

$$(I - \rho W)y = X\beta + \epsilon$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon = (QX)\beta + Q\epsilon \tag{3}$$

where $Q = (I - \rho W)^{-1}$ and $\beta,\ \rho$ are constants (because we are modeling a particular problem). The effect of transforming feature vector X to QX can be viewed as a spatial smoothing operation.

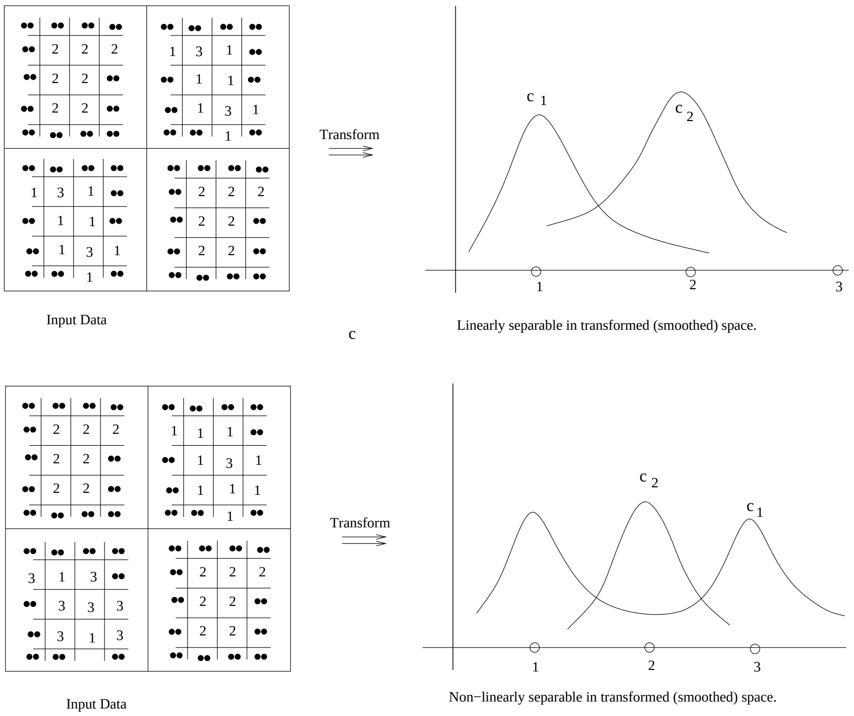

Transform
c 1
c 2
1
2
3
Input Data
c
Linearly separable in transformed (smoothed) space.
Transform
c 2
c 1
1
2
3
Input Data
Non−linearly separable in transformed (smoothed) space.

Figure 4: Spatial datasets with *salt and pepper* spatial patterns

The SAR model is similar to the linear logistic model in terms of the transformed feature space. In other words, the SAR model assumes the linear separability of classes in transformed feature space.

Figure 4 shows two datasets with a *salt and pepper* spatial distribution of the feature values. There are two classes, c_1 and c_2, defined on this feature. Feature values close to 2 map to class c_2 and feature values close to 1 or 3 will map to c_1. These classes are not linearly separable in the original feature space. Local spatial smoothing can eliminate the *salt and pepper* spatial pattern in the feature values to transform the distribution of the feature values. In the top part of Figure 4, there are few values of 3 and smoothing revises them close to 1 since most neighbors have values of 1. SAR can perform well with this dataset since classes are linearly separable in the transformed space. However, the bottom part of Figure 4 shows a different spatial dataset where local smoothing does not make the classes linearly separable. Linear classifiers cannot separate these classes even in the transformed feature space, assuming that $Q = (I - \rho W)^{-1}$ does not make the classes linearly separable.

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. For logistic regression, the probability of the set of labels L is given by:

$$Pr(L|X) = \prod_{i=1}^{N} p(l_i|X) \tag{4}$$

One important difference between logistic regression and MRF is that logistic regression assumes no dependence on neighboring classes. Given the logistic model, the probability that the binary label takes its first value c_1 at a location s_i is:

$$Pr(l_i|X) = \frac{1}{1 + \exp(-Q_i X\beta)} \tag{5}$$

where the dependence on the neighboring labels exerts itself through the W matrix, and subscript i (in Q_i) denotes the i^{th} row of the matrix Q. Here we have used the fact that y can be rewritten as in equation 3.

To find the local relationship between the MRF formulation and the logistic regression formulation (for the two class case $c_1 = 1$ and $c_2 = 0$), at point s_i

$$\begin{aligned} Pr((l_i = 1)|X, L_i) &= \frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i) + Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)} \\ &= \frac{1}{1 + \exp(-Q_i X\beta)} \end{aligned} \tag{6}$$

which implies

$$Q_i X\beta = \ln(\frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)}) \tag{7}$$

This last equation shows that the spatial dependence is introduced by the W term through Q_i. More importantly, it also shows that in fitting β we are trying to simultaneously fit the relative importance of the features and the relative frequency $(\frac{Pr(l_i=1,L_i)}{Pr(l_i=0,L_i)})$ of the labels. In contrast, in the MRF formulation, we explicitly *model* the relative frequencies in the class prior term. Finally, the relationship shows that we are making distributional assumptions about the class conditional distributions in logistic regression. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by

$$Pr(u|v) = e^{A(\theta_v) + B(u,\pi) + \theta_v^T u} \tag{8}$$

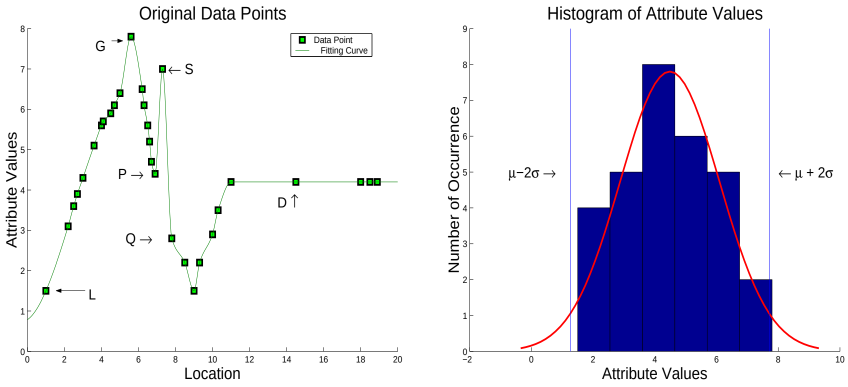
where u, v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases. The parameters θ_v and π control the form of the distribution. Equation 7 implies that the class conditional distributions are from the exponential family. Moreover, the distributions $Pr(X|l_i = 1, L_i)$ and $Pr(X|l_i = 0, L_i)$ are matched in all moments higher than the mean (e.g., covariance, skew, kurtosis, etc.), such that in the difference $ln(Pr(X|l_i = 1, L_i)) - ln(Pr(X|l_i = 0, L_i))$, the higher order terms cancel out, leaving the linear term $(\theta_v^T u)$ in equation 8 on the left hand-side of equation 7.

Experimental Results: 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 MRF models yield better spatial and classification accuracies over SAR in the prediction of the locations of bird nets. We also observed that SAR predications are extremely localized, missing actual nests over a large part of the marsh lands.

3 Spatial Outlier Detection Techniques

Global outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [3], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [11]. 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. These application domains include transportation, ecology, public safety, public health, climatology, and location based services.

We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 5(a), the X-axis is the location of data points in one dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point, and fit the distribution model to the values of the non-spatial attribute. The outlier detected using a this approach is the data point G. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.



(a) An Example Data Set

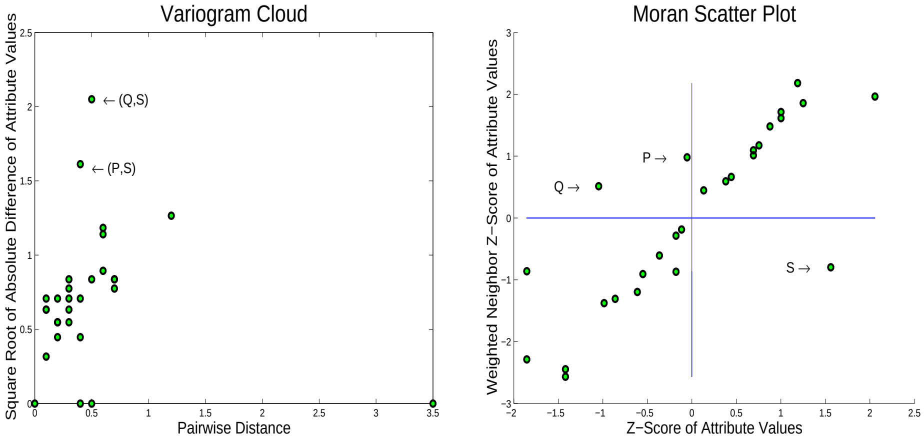
(b) Histogram

Figure 5: A Data Set for Outlier Detection

3.1 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 are based on the visualization of spatial data which highlights 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 [18] are a representative technique from the quantitative family. Figure 6(a) shows a variogram cloud for the example data set shown in Figure 5(a). This plot shows that two pairs (P, S) and (Q, S)

on the left hand side lie above the main group of pairs, and are possibly related to spatial outliers. The point S may be identified as a spatial outlier since it occurs in both pairs (Q, S) and (P, S). However, graphical tests of spatial outlier detection are limited by the lack of precise criteria to distinguish spatial outliers. In addition, a variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present or density varies greatly.



(a) Variogram cloud

(b) Moran scatterplot

Figure 6: Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers

A Moran scatterplot [19] 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 6(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values surrounded by low values (e.g,. point S). Thus we can identify points(nodes) that are surrounded by unusually high or low value neighbors. These points can be treated as spatial outliers.

3.2 Definition of S-Outliers

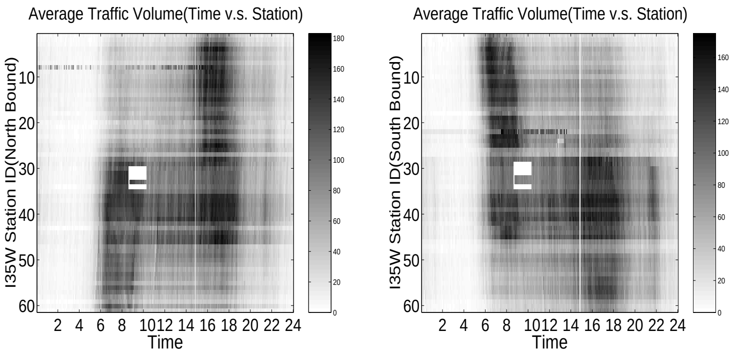
Consider a spatial framework $SF =< S, NB >$, where S is a set of locations $\{s_1, s_2, \ldots, s_n\}$ and $NB : S \times S \rightarrow \{True, False\}$ is a neighbor relation over S. We define a neighborhood $N(x)$ of a location x in S using NB, specifically $N(x) = \{y \mid y \in S\,, NB(x, y) = True\}$.

Definition: An object O is an S-outlier(f, f^N_{aggr}, F_{diff}, ST) if $ST\{F_{diff}[f(x), f^N_{aggr}(f(x),N(x))]\}$ is true, where $f : S \rightarrow R$ is an attribute function, $f^N_{aggr} : R^N \rightarrow R$ is an aggregation function for the values of f over neighborhood, R is a set of real numbers, $F_{diff} : R \times R \rightarrow R$ is a difference function, and $ST : R \rightarrow \{True, False\}$ is a statistic test procedure for determining statistical significance.

3.3 Solution Procedures

Given the components of the S-outlier definition, the objective is to design a computationally efficient algorithm to detect S-outliers. We presented scalable algorithms for spatial outlier detetection in [29], where we showed that almost all statistical tests are "algebraic" aggregate functions over a neighborhood join. The spatial outlier detection algorithm has two distinct tasks: the first task deals with model building and the second task involves a comparison (test statistic) with spatial neighbors. During model building, algrebraic aggregate functions (e.g., mean and standard deviation) are computed in a single scan of a spatial-join using a neighbor relationship. In the second step, a neighborhood aggregate function is computed by retrieving the neighboring nodes and then a difference function is applied over the neighborhood aggregates and algebraic aggregates. This study showed that the computational cost of outlier detection algorithms are dominated by the disk page access time (i.e., the time spent on accessing neighbors of each point). In this study we utilized three different data page clustering schemes: the Connectivity-Clustered Access Method (CCAM) [27], Z-ordering [23], and Cell-tree [10] and found that CCAM produced the lowest number of data page accesses for outlier detection.

The effectiveness of the $Z_{s(x)}$ method on a Minneapolis-St. Paul traffic data set is illustrated in the following example. Figure 7 shows one example of traffic flow outliers. Figures 7(a) and (b) are the traffic volume maps for I-35W north bound and south bound, respectively, on January 21, 1997. The X-axis is a 5-minute time slot for the whole day and the Y-axis is the label of the stations installed on the highway, starting from 1 on the north end to 61 on the south end. The abnormal white line at 2:45PM and the white rectangle from 8:20AM to 10:00AM on the X-axis and between stations 29 to 34 on the Y-axis can be easily observed from both (a) and (b). The white line at 2:45PM is an instance of temporal outliers, where the white rectangle is a spatial-temporal outlier. Both represent missing data. Moreover, station 9 in Figure 7(a) exhibits inconsistent traffic flow compared with its neighboring stations, and was detected as a spatial outlier. Station 9 may be a malfunctioning sensor.



(a) I-35W North Bound

(b) I-35W South Bound

Figure 7: Spatial outliers in traffic volume data

4 Spatial Co-location Rules

Association rule finding [12] is an important data mining technique which has helped retailers interested in finding items frequently bought together to make store arrangements, plan catalogs, and promote products together. In market basket data, a transaction consists of a collection of item types purchased together by a customer. Association rule mining algorithms [1] assume that a finite set of disjoint transactions are given as input to the algorithms. Algorithms like *apriori* [1] can efficiently find the frequent itemsets from all the transactions and association rules can be found from these frequent itemsets. Many spatial datasets consist of instances of a collection of boolean spatial features (e.g. drought, needle leaf vegetation). While boolean spatial features can be thought of as item types, there may not be an explicit finite set of transactions due to the continuity of underlying spaces. In this section we define co-location rules, a generalization of association rules to spatial datasets.

4.1 Illustrative Application Domains

Many ecological datasets [17, 20] consist of raster maps of the Earth at different times. Measurement values for a number of variables (e.g., temperature, pressure, and precipitation) are collected for different locations on Earth. A set of events, i.e., boolean spatial features, are defined on these spatial variables. Example events include drought, flood, fire, and smoke. Ecologists are interested in a variety of spatio-temporal patterns including co-location rules. Co-location patterns represent frequent co-occurrences of a subset of boolean spatial features.

4.2 Co-location Rule Approaches

Given the difficulty in creating explicit disjoint transactions from continuous spatial data, this section defines several approaches to model co-location rules. We use Figure 8 as an example spatial dataset to illustrate the different models. In this figure, a uniform grid is imposed on the underlying spatial framework. For each grid l, its neighbors are defined to be the nine adjacent grids (including l). Spatial feature types are labeled beside their instances. We define the following basic concepts to facilitate the description of different models.

Definition 1 *A* **co-location** *is a subset of boolean spatial features or spatial events.*

Definition 2 *A* **co-location rule** *is of the form* $C_1 \rightarrow C_2(p, cp)$ *where* C_1 *and* C_2 *are co-locations,* p *is a number representing the prevalence measure, and* cp *is a number measuring conditional probability.*

The prevalence measure and the conditional probability measure are called interest measures and are defined differently in different models which will be described shortly.

The **reference feature centric model** 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, other substances) to the reference feature. This model enumerates neighborhoods to "materialize" a set of transactions around instances of the reference spatial feature. A specific example is provided by the spatial association rule [14].

For example, in Figure 8 (a), let the reference feature be A, the set of task relevant features be B and C, and the set of spatial predicates include one predicate named "*close_to*". Let us define

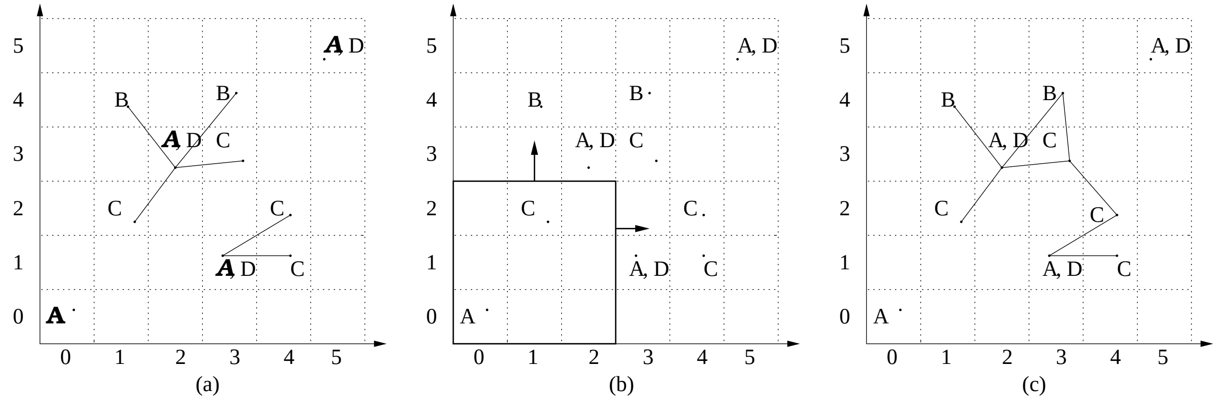


Figure 8: Spatial dataset to illustrate different co-location models. Spatial feature types are labeled besides their instances. The 9 adjacent grids of a grid l (including l) are defined to be l's neighbors. a) Reference feature-centric model. The instances of A are connected with their neighboring instances of B and C by edges. b) Window-centric model. Each 3 X 3 window corresponds to a transaction. c) Event-centric model. Neighboring instances are joined by edges.

$close\_to(a, b)$ to be true if and only if b is a's neighbor. Then for each instance of spatial feature A, a transaction which is a subset of relevant features $\{B, C\}$ is defined. For example, for the instance of A at (2,3), transaction $\{B, C\}$ is defined because the instance of B at (1,4) (and at (3,4)) and instance of C at (1,2) (and at (3,3)) are *close_to* (2,3). The transactions defined around instances of feature A are summarized in Table 1.

Table 1: Reference feature centric view: transactions are defined around instances of feature A relevant to B and C in figure 8(a)

Instance of A	Transaction
(0,0)	ø
(2,3)	$\{B, C\}$
(3,1)	$\{C\}$
(5,5)	ø

With "materialized" transactions, the support and confidence of the traditional association rule problem [1] may be used as prevalence and conditional probability measures as summarized in Table 2. Since 1 out of 2 non-empty transactions contains instances of both B and C and 1 out of 2 non-empty transactions contain C in Table 1, an association rule example is: $is\_type(i, A) \wedge \exists j\, is\_type(j, B) \wedge close\_to(j, i) \rightarrow \exists k\, is\_type(k, C) \wedge close\_to(k, i)$ with $\frac{1}{1} * 100\% = 100\%$ probability.

The **window centric model** is relevant to applications like mining, surveying and geology, which focus on land-parcels. A goal is to predict sets of spatial features likely to be discovered in a land parcel given that some other features have been found there. The window centric model enumerates all possible windows as transactions. In a space discretized by a uniform grid, windows of size kXk can be enumerated and materialized, ignoring the boundary effect. Each transaction contains a subset of spatial features of which at least one instance occurs in the corresponding window. The support and confidence of the traditional association rule problem

Table 2: Interest measures for different models

Model	Items	transactions defined by	Interest measures for $C_1 \rightarrow C_2$	
			Prevalence	Conditional probability
local	boolean feature types	partitions of space	fraction of partitions with $C_1 \cup C_2$	$Pr(C_2$ in a partition given C_1 in the partition)
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)
window centric	boolean feature types	possibly infinite set of distinct overlapping windows	fraction of windows with $C_1 \cup C_2$	$Pr(C_2$ in a window given C_1 in that window)
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 $C_1)$

may again be used as prevalence and conditional probability measures as summarized in Table 2. There are 16 3X3 windows corresponding to 16 transactions in Figure 8 b). All of them contain A and 15 of them contain both A and B. An example of an association rule of this model is: $an\, instance\, of\, type\, A\, in\, a\, window \rightarrow an\, instance\, of\, type\, B\, in\, this\, window$ with $\frac{15}{16} = 93.75\%$ probability. A special case of the window centric model relates to the case when windows are spatially disjoint and form a partition of space. This case is relevant when analyzing spatial datasets related to the units of political or administrative boundaries (e.g. country, state, zip-code). In some sense this is a local model since we treat each arbitrary partition as a transaction to derive co-location patterns without considering any patterns cross partition boundaries. The window centric model "materializes" transactions in a different way from the reference feature centric model.

The **event centric model** is relevant to applications like ecology, where there are many types of boolean spatial features. Ecologists are interested in finding 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 8 c). There are four instances of type A and only one of them has some instance(s) of type B in its 9-neighbor adjacent neighborhoods. The conditional probability for the co-location rule is: $spatial\, feature\, A\, at\, location\, l \rightarrow spatial\, feature\, type\, B\, in\, 9 - neighbor\, neighborhood\, is\, 25\%$.

4.3 Solution Procedures

Co-location mining is a complex task. It consits of two tasks, schema level purning and instance level purning. At schema level purning, *apriori* [1] can be used. However instance level purning involves neighborhood (i.e., co-location row instance) enumeration, which is a compute intense task. Shekhar et al [26] developed pure geometric, pure combinatorial, hybrid, and multi-resolution algorithms for instance level purning. Experimental analysis shows that the pure geometric algrothm performs much better than pure combinatorial approach. Hybrid algorithm, which is a combination of geometric and combinatorial methods, performed better than both of these approaches. On the other hand, multi-resolution algorithm out performs all these methods when the data is "clumped". It is also shown that co-lcation miner algorithm is complete and correct.

5 Conclusions and Future Work

In this paper we have provided techniques that are specifically designed to analyze large volumes of spatial data to predict bird nests, to find spatial outliers, and to find co-location association rules. We compared the SAR and MRF models using a common probabilistic framework. Our study shows that the SAR model makes more restrictive assumptions about the distribution of features and class shapes (or decision boundaries) than MRF. We also observed an interesting relationship between classical models that do not consider spatial dependence and modern approaches that explicitly model spatial context. The relationship between SAR and MRF is analogous to the relationship between logistic regression and Bayesian Classifiers. The analysis of spatial outlier detection algorithms showed the need for good clustering of data pages. The CCAM method yielded the best overall performance. We showed that the co-location miner algorithm is complete and correct and performs better than the well know *apriori* algorithm.

6 Acknowledgements

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 would like to thank to our collaborators Prof. Paul Schrater and Prof. Uygar Ozesmi for their various contributions and group members Yan Huang, Chang-Tien Lu, Weili Wu, and Pusheng Zang for their contributions in developing these techniques and software. The comments of Kimberly Koffolt have greatly improved the readability of this paper.

References

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

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

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

[4] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Soc.*, (48):259–302, 1986.

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

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts . *Proc. of International Conference on Computer Vision*, September 1999.

[7] P.B. Chou, P.R. Cooper, M. J. Swain, C.M. Brown, and L.E. Wixson. Probabilistic network inference for cooperative high and low levell vision. In *In Markov Random Field, Theory and Applicaitons*. Academic Press, New York, 1993.

[8] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, (9):39–55, 1987.

[9] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[10] O. Gunther. The Design of the Cell Tree: An Object-Oriented Index Structure for Geometric Databases. In *Proc. 5th Intl. Conference on Data Engineering*, Feb. 1989.

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

[12] J. Hipp, U. Guntzer, and G. Nakaeizadeh. Algorithms for Association Rule Mining - A General Survey and Comparison. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

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

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

[15] J. P. LeSage and R.K. Pace. Spatial dependence in data mining. In *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, forthcoming, 2001.

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

[17] Z. Li, J. Cihlar, L. Moreau, F. Huang, and B. Lee. Monitoring Fire Activities in the Boreal Ecosystem. *Journal Geophys. Res,. 102(29):611-629*, 1997.

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

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

[20] D.C. Nepstad, A. Verissimo, A. Alencar, C. Nobre, E. Lima, P. Lefebvre, P. Schlesinger, C. Potter, P. Moutinho, E. Mendoza, M. Cochrane, and V. Brooks. Large-scale Impoverishment of Amazonian Forests by Logging and Fire. *Nature, 398:505-508*, 1999.

[21] J.P. oeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

[22] University of Minnesota. Spatial database research group. http://www.cs.umn.edu/research/shashi-group/.

[23] A. Orenstein and T. Merrett. A Class of Data Structures for Associative Searching. In *Proc. Symp. on Principles of Database Systens*, pages 181–190, 1984.

[24] R. Pace and R. Barry. Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable. *Geographic Analysis*, 1997.

[25] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.

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

[27] S. Shekhar and D-R. Liu. CCAM: A Connectivity-Clustered Access Method for Aggregate Queries on Transportation Networks. *IEEE Transactions on Knowledge and Data Engineering*, 9(1):102–119, January 1997.

[28] S. Shekhar, Paul R. Schrater, Ranga R. Vatsavai, Weili Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transaction on Multimedia (accepted), http://www.cs.umn.edu/research/shashi-group/paper_list.html*, 2002.

[29] Shashi Shekhar, C.T. Lu, and P. Zhang. A unified approach to spatial outliers detection. *TR01-045 (Also to appear in IEEE TKDE), http://www.cs.umn.edu/research/shashi-group/paper_list.html*, 2001.

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

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