Overview

Identifying patterns in spatial information: a survey of methods

Shashi Shekhar, Michael R. Evans, James M. Kang and Pradeep Mohan

Explosive growth in geospatial data and the emergence of new spatial technologies emphasize the need for automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial databases. The complexity of spatial data and implicit spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. In this paper, we explore the emerging field of spatial data mining, focusing on different methods to extract patterns from spatial information. We conclude with a look at future research needs. © 2011 John Wiley & Sons, Inc. *WIREs Data Mining Knowl Discov* 2011 1 193–214
DOI: 10.1002/widm.25

INTRODUCTION

The significant growth of spatial data collection and widespread use of spatial databases[1–4] have heightened the need for the automated discovery of spatial knowledge. Spatial data mining[2,5] is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and implicit spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns.

Specific features of geographical data that preclude the use of general purpose data mining algorithms are: (1) the spatial relationships among the variables; (2) the spatial structure of errors; (3) the presence of mixed distributions as opposed to commonly assumed normal distributions; (4) observations that are not independent and identically distributed (*i.i.d.*); (5) spatial autocorrelation among the features; and (6) nonlinear interactions in feature space. Although conventional data mining algorithms can be applied under assumptions such as *i.i.d.*, these algorithms often perform poorly on spatial data due to their self-correlated nature. To illustrate, we use an example from ecology where domain scientists are interested in studying the habitats of birds based on the attributes of locations in the study area, such as water depth. Figure 1 shows two different attributes, one with an assumption of *i.i.d.* (Figure 1(a)) and one that has spatial autocorrelation (Figure 1(b)), that is, water depth in this case. Making use of water depth as an explanatory variable by accounting for spatial autocorrelation was found to model the ground truth better (i.e., predicting the habitat of birds).[6]

Efficient tools for extracting information from geospatial data are crucial to organizations that make decisions based on large spatial data sets. These application domains include public health,[7–9] mapping and analysis for public safety,[10] transportation,[11–14] environmental science and management,[15–19] economics,[20] climatology,[5,21,22] public policy,[23,24] earth science,[25] market research and analytics,[26–28] public utilities and distribution, etc.[29–31] Many government and private agencies that are likely beneficiaries of spatial data mining include the National Institute of Health (NIH), National Institute of Justice (NIJ), US Department of Transportation, (USDOT), US Department of Agriculture (USDA), National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), IBM, and SIEMENS.

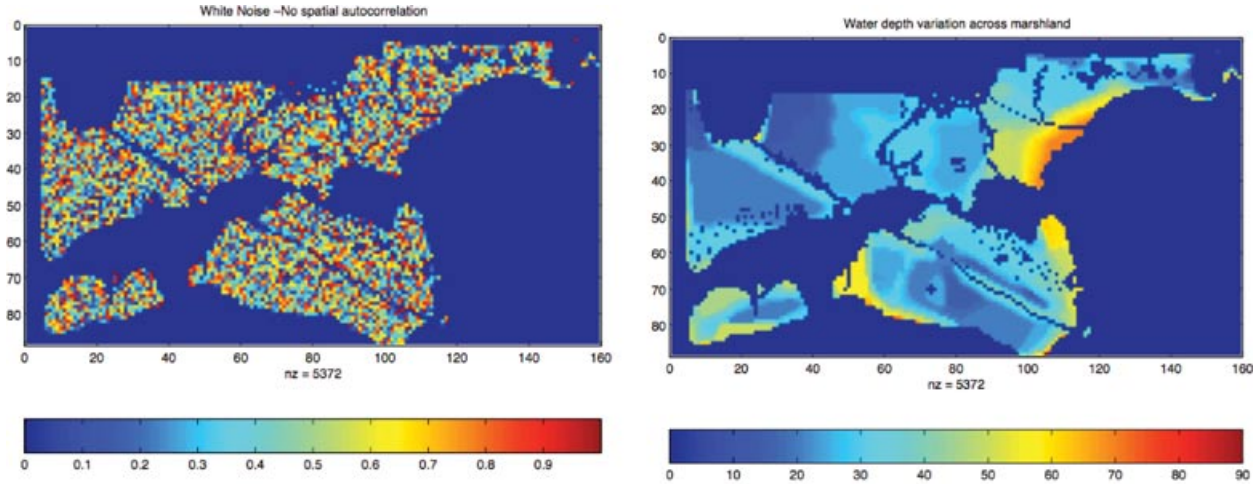
The challenges inherent in the management and analysis of spatial data sets have made spatial databases a particularly active area of research for several decades. The impacts of this research extend far and wide. To cite a few examples, the filter-and-refine technique used in spatial query processing has been applied to subsequence mining; multidimensional-index structures are used in computer graphics and image processing; and space-filling

*Correspondence to: shekhar@cs.umn.edu

Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, USA

DOI: 10.1002/widm.25

 © 2011 John Wiley & Sons, Inc.



(a) Attribute with an independent identical distribution (b) Attribute with spatial autocorrelation

FIGURE 1 | Attribute values in space with independent identical distribution and spatial autocorrelation.

curves used in spatial query processing and data storage are applied in dimension reduction problems. The value of its contributions no longer in doubt, current research in spatial databases aims to improve its functionality, extensibility, and performance. The impetus for improving functionality comes from the needs of numerous existing applications such as geographic information systems, location-based services,[32] and sensor networks.[33]

These research advances coupled with the growing need for spatial information awareness have given rise to many commercial spatial database management systems (SDBMS). Some examples of SDBMS include ESRI's ArcGIS Geodatabase,[34] Oracle Spatial,[35] IBM's DB2 Spatial Extender and Spatial Datablade, and systems such as Microsoft's SQL Server 2008.[36] Spatial databases have played a major role in popular applications such as Google Earth[37] and Microsoft's Virtual Earth.[38] Research prototype examples of SDBMS include spatial datablades with PostGIS,[39] MySQL's Spatial Extensions,[40] Sky Server,[41] and spatial extensions. The functionalities provided by these systems include use of spatial data types such as points, line segments and polygons, and spatial operations such as inside, intersection, and distance. Spatial types and operations may be integrated into query languages such as SQL, which allows spatial querying to be combined with object-relational database management systems.[42,43] The performance enhancement provided by these systems includes a multidimensional spatial index and algorithms for spatial database modeling such as OGIS[44] and 3D topological modeling; spatial query processing including point, regional, range, and nearest neighbor queries; and spatial data methods using a variety of indexes such as quad trees and grid cells.

In addition, there has been a growth in general purpose data mining tools such as Clementine from Statistical Package for the Social Sciences (SPSS), Enterprise Miner from SAS, Data Mining extensions from relational database vendors such as Oracle and IBM, public domain data mining packages such as Weka,[45] See5/C5.0, etc., which are designed for the purpose of analyzing data archived as transactions or other forms such as semi-structured data. Although these tools were primarily designed to identify customer-buying patterns in market basket data, they have also been used in analyzing scientific and engineering data, astronomical data, multi-media data, genomic data, and web data.[46,47] However, extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, spatial autocorrelation, and nonlinearity.

The remainder of this paper is organized as follows: *Spatial Data Inputs* begins with a description of the data input characteristics of several tasks in spatial data mining. *Statistical Foundations* provides an overview of the statistical foundation of spatial data mining (SDM). *Spatial data mining tasks, explains in detail* four main output patterns and methods of SDM related to anomalies, clustering, co-location, and prediction. Computational issues regarding these patterns are discussed in *Computational Issues*. We survey some available spatial analysis tools for different SDM techniques in *Spatial Analysis Tools*. *Future Directions and*

 © 2011 John Wiley & Sons, Inc.

Research Needs concludes this paper with an examination of research needs and future directions.

SPATIAL DATA INPUTS

The data inputs of SDM are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons in vector representation and field data in regular or irregular tessellation such as raster data. The data inputs of SDM have two distinct types of attributes: nonspatial attributes and spatial attributes. Nonspatial attributes are used to characterize nonspatial features of objects such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used to define the spatial location and extent of spatial objects.[48,49] The spatial attributes of a spatial object most often include information related to spatial locations, for example, longitude, latitude, and elevation, defined in a spatial reference frame, as well as shape.

In some applications, spatial data sets include discrete representations of continuous phenomena (e.g., ecology). Discretization of continuous space is necessitated by the nature of the digital representation or semantics associated with the underlying phenomenon under study by an application domain. There are two basic models to represent spatial data, namely, raster (grid) and vector. Satellite images are good examples of raster data, while vector consists of points, lines, polygons, and their aggregate (or multi-) counterparts. This distinction is important because many of the techniques that we describe later favor one or more of these data types. Vector data over a space is a framework to formalize specific relationships among a set of objects. Depending on the relationships of interest, the space can be modeled in many different ways, that is, as set-based space, topological space, Euclidean space, metric space, and network space.[4] These models of space are described briefly in this paper.

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

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

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

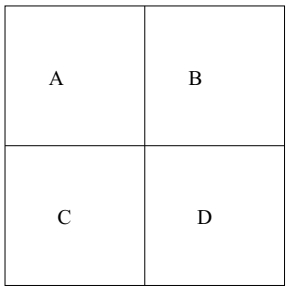
Apart from different concepts of space, many gazetteers employ spatial referencing with identifiers of a location that can be transformed into coordinates, such as a postal code (street addresses) or geoname that is more natural to human understanding. Time is usually included in the spatial data as a time stamp.

During data input, relationships among nonspatial objects are made explicit through arithmetic relation, ordering, instance-of, subclass-of, and membership-of. In contrast, relationships among spatial objects are often implicit, such as overlap, intersect, and behind. Table 1 gives examples of spatial and nonspatial relationships. One possible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques such as those described in Refs 50–54. However, the materialization can result in loss of information. Usually, spatial and temporal vagueness, which naturally exists in data and relationships, creates further modeling and processing difficulty in SDM. Another way

TABLE 1 | Common Relationships among Nonspatial and Spatial Data

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

 © 2011 John Wiley & Sons, Inc.



(a) Spatial framework

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

(b) Neighbor relationship

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

(c) Contiguity matrix

FIGURE 2 | A spatial framework and its four-neighborhood contiguity matrix.

to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the SDM process.

STATISTICAL FOUNDATIONS

Spatial statistics is a branch of statistics concerned with the analysis and modeling of spatial data.[55] The field classifies spatial data into three basic types for ease of interpretation: (1) point referenced data, which is modeled as a fixed collection of spatial locations, S, in a two-dimensional framework D (e.g., set of police stations in a metropolitan city); (2) areal data, modeled as a finite set of irregular shaped polygons in a two-dimensional framework D (e.g., set of police districts in a metropolitan city); and (3) point process data, which is modeled as a random collection of spatial events, collectively referred to as the spatial point pattern over a two-dimensional framework D (e.g., home locations of patients infected by a disease). In this section, three important statistical foundations are reviewed. They are: (1) spatial statistical interpretation models, (2) spatial neighborhood models, and (3) special properties of spatial data analysis.

Statistical interpretation models[56] are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on probability theory. Spatial data can be thought of as resulting from observations on the stochastic process $Z(s) : s \in D$, where s is a spatial location and D is possibly a random set of points in a spatial framework. Three types of spatial statistical interpretation models that one might encounter are a point process, lattice, and geostatistics.

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

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

Geostatistics: Geostatistics deals with the analysis of spatial continuity and weak stationarity,[56] which are inherent characteristics of spatial data sets. Geostatistics provides a set of statistics tools, such as kriging, to the interpolation of attributes at unsampled locations.

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

Figure 2(a) shows a gridded spatial framework with four locations, A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 2(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 2(c). Other contiguity matrices can be designed to model neighborhood relationship based on distance, different forms of connectivity (e.g., rook, queen), etc. The uniqueness of SDM originates from two central concepts in spatial statistics: spatial autocorrelation and spatial heterogeneity(or nonstationarity).[56–58]

 © 2011 John Wiley & Sons, Inc.

Spatial autocorrelation: One of the fundamental assumptions of traditional statistical analysis is that the data samples are independently generated: like successive tosses of coin or the rolling of a die. However, in the analysis of spatial data, the assumption about the independence of samples is generally false. In fact, spatial data tend to be highly self-correlated. For example, people with similar characteristics, occupation, and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife, and temperature vary gradually over space. The property of like things clustering in space is so fundamental that geographers have elevated it to the status of the first law of geography: 'Everything is related to everything else, but nearby things are more related than distant things'.[59] For example, Figure 1 shows the value distributions of an attribute in a spatial framework for an independent identical distribution and a distribution with spatial autocorrelation.

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

There is a strong relationship between measures of spatial autocorrelation and the contiguity matrix. This is because the contiguity matrix represents the relationship between a spatial unit and its neighbors. This neighborhood interaction is quantified by common measures of spatial autocorrelation. However, the contiguity matrix for a particular spatial relationship may vary depending upon the definition of the spatial neighborhood. Such sensitivity in turn affects the robustness common measures of spatial autocorrelation (e.g., Moran's I) and many spatial statistical models. This is a challenging problem due to the many possible methods of defining a spatial neighborhood, namely, graph based, grid based, etc. A detailed study of these challenges is beyond the scope of this paper.

Spatial heterogeneity: Apart from spatial autocorrelation, an important feature of spatial data sets is the variability of observed process over space. Spatial heterogeneity refers to the inherent variation in measurements of relationships over space. The influence of spatial context on spatial relationships can be seen in the variation of human behavior over space (e.g., differing cultures). Different jurisdictions tend to produce different laws (e.g., speed limit differences between Minnesota and Wisconsin). The term spatial heterogeneity is most often used interchangeably with spatial nonstationarity, which is defined as the change in the parameters of a statistical model or change in the ranking of candidate models over space.[57]

SPATIAL DATA MINING TASKS

Important tasks in SDM are spatial outlier detection, co-location pattern discovery, spatial classification and regression modeling, spatial clustering, and spatial hotspot analysis. This section elaborates these techniques by briefly describing their computational structure, applications, and related methods.

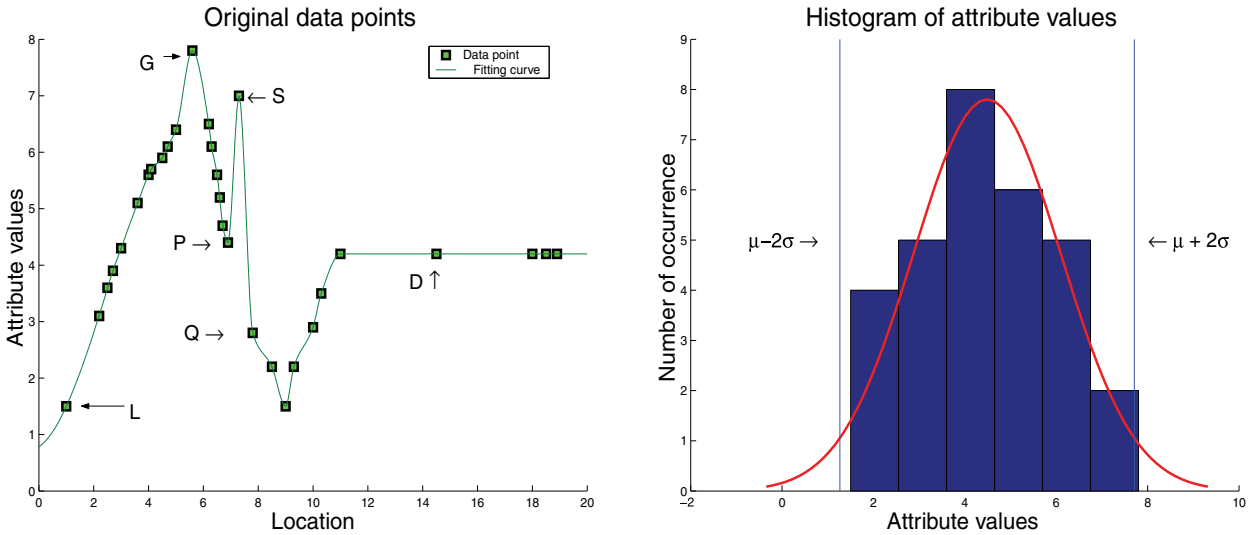
Spatial Outlier Detection

Outliers have been informally defined as observations in a data set that appear to be inconsistent with the remainder of that set of data,[60] or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism.[61] The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as detection of credit card fraud and voting irregularities. This section focuses on spatial outliers, that is, observations that appear to be inconsistent with their neighborhoods.[62–64] Detecting spatial outliers is useful in many geographic information systems and spatial databases applications such as transportation, ecology, homeland security, public health, climatology, and location-based services.

A spatial outlier[65] is a spatially referenced object whose nonspatial 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 nonspatial 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 nonspatial 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(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 nonspatial attribute. As shown in Figure 3(b), the outlier detected using this approach is the data point G, which has an extremely high attribute value 7.9, exceeding the threshold of $\mu + 2\sigma = 4.49 + 2 *$

 © 2011 John Wiley & Sons, Inc.



(a) An example dataset (b) Histogram

FIGURE 3 | A data set for outlier detection.

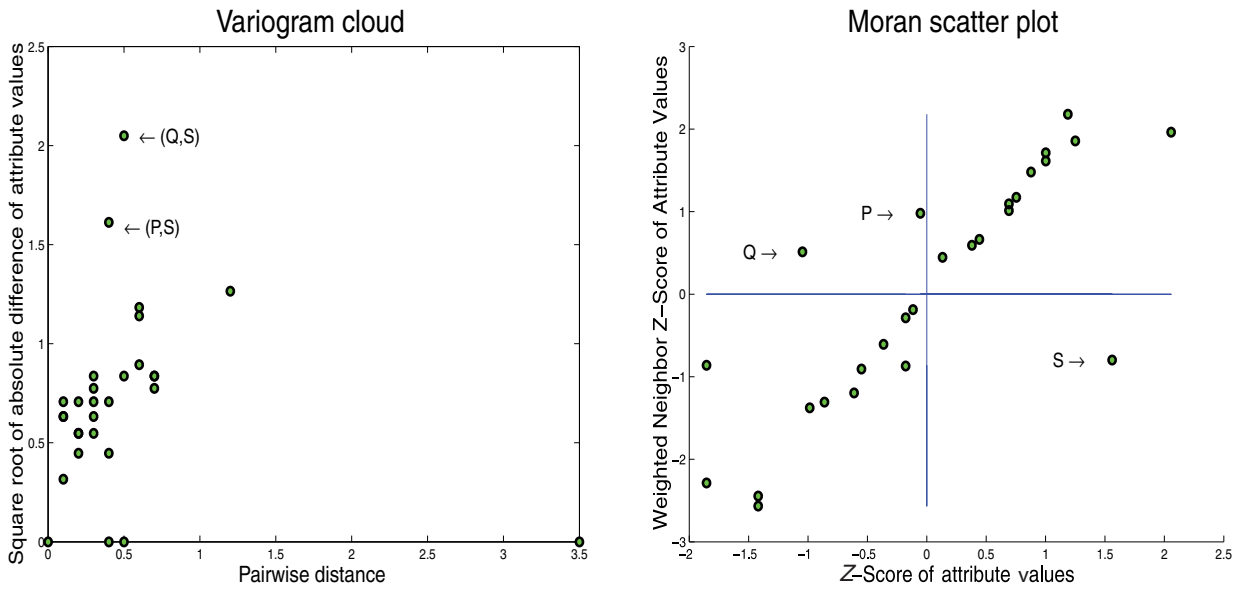
$1.61 = 7.71$. This test assumes a normal distribution for attribute values. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.

Common methods: Tests to detect spatial outliers separate spatial attributes from nonspatial attributes. Spatial attributes are used to characterize location, neighborhood, and distance. Nonspatial 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[66] and Moran scatterplots.[56,67] A variogram cloud displays data points related by neighborhood relationships. Figure 4(a) shows a variogram cloud for the example data set shown in Figure 3(a). This plot shows that two pairs (P, S) and (Q, S) on the left hand side lie above the main group of pairs and are possibly related to spatial outliers. A Moran scatterplot shows the spatial association or disassociation of spatially close objects. The upper left and lower right quadrants of Figure 4(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors (e.g., points P and Q) and high values surrounded by low values (e.g., point S). Figure 4(b) indicates a spatial association of dissimilar values: low values surrounded by high value neighbors (e.g., points P and Q) and high values surrounded by low values (e.g., point S).

A scatterplot[68] 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, respectively. In Figure 5(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the data set in Figure 3(a). Point S turns out to be the farthest from the regression line and may be identified as a spatial outlier.

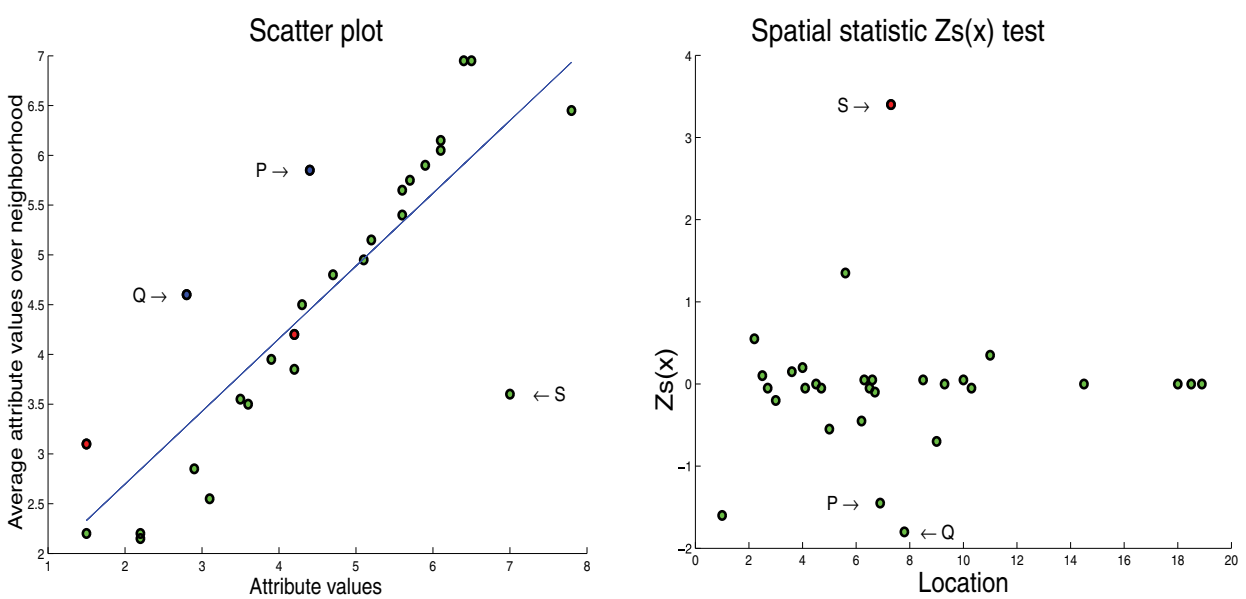
Spatial statistic $S(x)$ is normally distributed if the attribute value $f(x)$ is normally distributed. A popular test for detecting spatial outliers for normally distributed $f(x)$ can be described as follows: spatial statistic $Z_{s(x)} = |\frac{S(x) - \mu_s}{\sigma_s}| > \theta$. For each location,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

 © 2011 John Wiley & Sons, Inc.



(a) Variogram cloud (b) Moran scatterplot

FIGURE 4 | Variogram cloud and Moran scatterplot to detect spatial outliers.



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

FIGURE 5 | Scatterplot and spatial statistic $Z_{s(x)}$ to detect spatial outliers.

Figure 5(b) shows the visualization of the spatial statistic method described above. The X-axis is the location of data points in one-dimensional space; the Y-axis is the value of spatial statistic $Z_{s(x)}$ for each data point. We can easily observe that point S has a $Z_{s(x)}$ value exceeding 3 and will be detected as a spatial outlier. Note that the two neighboring points P and Q of S have $Z_{s(x)}$ values close to -2 due to the presence of spatial outliers in their neighborhoods.

 © 2011 John Wiley & Sons, Inc.

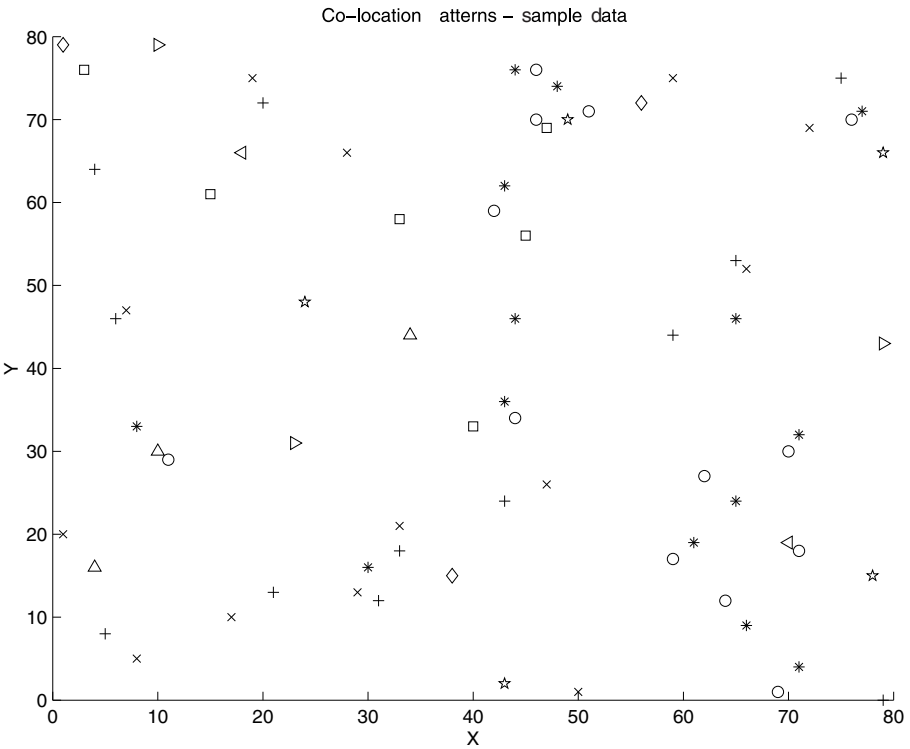


FIGURE 6 | Illustration of point spatial co-location patterns.

The techniques presented above are based on single attributes. However, multi-attribute based spatial outlier detection is also possible, such as with the average and median attribute value-based algorithms presented in Ref 69. Finally, we note that statistical tests used in outlier detection are normally prone to biases resulting from multiple hypothesis testing as spatial data sets are self-correlated. In order to deal with this, spatial statistics has explored several corrections to characterize the statistical significance of spatial outliers.[67]

Co-location Patterns

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

Spatial co-location: Co-location rules are models to infer the presence of boolean spatial features in the neighborhood of instances of other boolean spatial features. For example, 'Nile Crocodiles → Egyptian Plover' predicts the presence of Egyptian Plover birds in areas with Nile Crocodiles. Figure 6 shows a data set consisting of instances of several boolean spatial features, each represented by a distinct shape. The shapes in Figure 6 represent different spatial feature types. Spatial features in sets {'+', '×'} and {'o', '*'} tend to be located together. A careful review reveals two co-location patterns, that is, ('+', '×') and ('o', '*').

Co-location rule discovery is the process of identifying co-location patterns from large spatial data sets 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[51] 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.

Common methods: Spatial co-location rule mining approaches can be grouped into two broad

 © 2011 John Wiley & Sons, Inc.

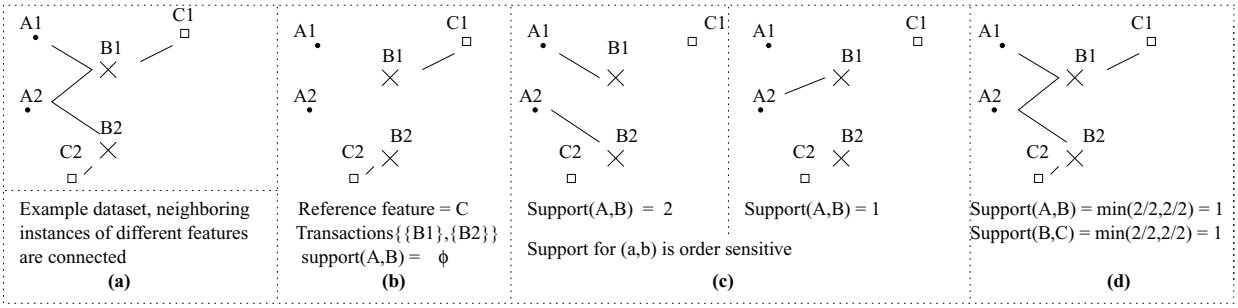


FIGURE 7 | Example to illustrate different approaches to discovering co-location patterns: (a) Example data set. (b) Reference feature-centric model. (c) Data partition approach. Support measure is ill-defined and order sensitive. (d) Event-centric model.

categories: approaches that use spatial statistics and algorithms that use association rule mining kind of primitives. Spatial statistics based approaches utilize statistical measures such as cross-K function, mean nearest-neighbor distance, and spatial autocorrelation. However, these approaches are computationally expensive. Association rule-based approaches focus on the creation of transactions over space so that an *a priori* like algorithm[51] can be used. Transactions in space can use a reference-feature centric[70] approach or a data-partition[71] approach. The reference-feature centric model is based on the choice of a reference spatial feature[70] and is relevant to application domains focusing on a specific boolean spatial feature, for example, cancer. In the data partitioning approach, transactions are created by making use of a prevalence measure that is order sensitive. In the spatial co-location rule mining problem, however, transactions are often not explicit. Force fitting the notion of transaction in a *continuous spatial framework* will lead to loss of implicit spatial relationships across the boundary of these transactions, as illustrated in Figure 7. In the data set, in Figure 7(a), there are three feature types, A, B, and C, each of which has two instances. The neighbor relationships between instances are shown as edges. Co-locations (A, B) and (B, C) may be considered as frequent in this example. Figure 7(b) shows transactions created by choosing C as the reference feature. As Co-location (A, B) does not involve the reference feature, it will not be found. Figure 7(c) shows two possible partitions for the data set of Figure 7(a), along with the supports for co-location (A, B); in this case, the support measure is order sensitive and may also miss the Co-location (A, B). However, the event-centric model addresses these limitations[72] and finds subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types (see Figure 7(d)).

Spatial Classification and Regression Models

Spatial classification and regression models in data mining have been used to represent relationships between variables in different data sets (e.g., climate). In most of these data sets, there are two sets of variables, namely, independent or explanatory variables and dependent variables. Although classification models deal with discrete values of dependent variables (e.g., class labels), regression models are concerned with continuous valued ones. In most SDM applications, classification and regression models can be learned from data in different ways such as supervised learning, unsupervised learning, and semi-supervised learning. In this paper, we review only supervised learning. Given a sample set of input–output pairs, the objective of supervised learning is to learn a function that matches reasonably well with the input data and predicts an output for any unseen input (but assumed to be generated from the same distribution), such that the predicted output is as close as possible to the desired output. For example, in remote sensing image classification, the input attribute space consists of various spectral bands or channels (e.g., blue, green, red, infra-red, thermal, etc.) The input vectors (x_i's) are reflectance values at the ith location in the image, and the outputs (y_i's) are thematic classes such as forest, urban, water, and agriculture. The type of output attribute determines the supervised learning task; two such tasks are:

- ***Classification:*** Here, the input vectors x_i are assigned to a few discrete numbers of classes, for example, image classification[73] y_i.
- ***Regression:*** In regression, also known as function approximation or prediction, the input–output pairs are generated from an unknown function of the form $y = f(x)$, where y is

 © 2011 John Wiley & Sons, Inc.

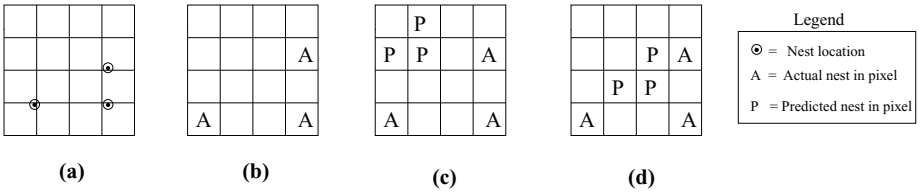


FIGURE 8 | (a) The actual locations of nests. (b) Pixels with actual nests. (c) Location predicted by a model. (d) Location predicted by another model. Prediction (d) is spatially more accurate than (c).

continuous. Typically, regression is used in regression and estimation, for example, crop yield prediction,[74] daily temperature prediction, and market share estimation for a particular product. Regression can also be used in inverse estimation, that is, given that we have an observed value of y, we want to determine the corresponding x value.

However, while performing supervised learning, conventional data mining techniques perform poorly in identifying values of dependent variables due to two reasons. The first reason is because they ignore spatial autocorrelation and heterogeneity in the model building process. 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. This is because the measure of *Spatial accuracy*—how far the predictions are from the actuals—is important in some applications such as ecology due to the effects of the discretization of a continuous wetland into discrete pixels, as shown in Figure 8. Figure 8(a) shows the actual locations of nests and (b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled 'A' and are quite close to other blank pixels, which represent 'no-nest'. Now consider two predictions shown in Figure 8(c) and (d). Domain scientists prefer prediction 8(d) over (c), as the predicted nest locations are closer on average to some actual nest locations. The classification accuracy measure cannot distinguish between 8(c) and (d), and a measure of spatial accuracy is needed to capture this preference.

Common methods: Several previous studies[75,76] have shown that the modeling of spatial dependency (often called context) during the classification or regression process improves overall accuracy. Spatial context can be defined by the relationships between spatially adjacent spatial units in a small neighborhood. An example spatial framework and its four-neighborhood contiguity matrix is shown in Figure 2. Three supervised learning techniques for classification and regression that model spatial dependency are: (1) Markov random field (MRF) based classifiers; (2) logistic spatial autoregression (SAR) model; and (3) geographically weighted regression (GWR).

Markov random field-based Bayesian classifiers: Maximum likelihood classification (MLC) is one of the most widely used parametric and supervised classification technique in the field of remote sensing.[77,78] However, MLC is a per-pixel based classifier and assumes that samples are *i.i.d.* Ignoring spatial autocorrelation results in *salt and pepper* kind of noise in the classified images. One solution is to use MRF-based Bayesian classifiers[79] to model spatial context via the *a priori* term in Bayes' rule. This uses a set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix). A more detailed theoretical and experimental comparison of these two methods can be found in Ref 80.

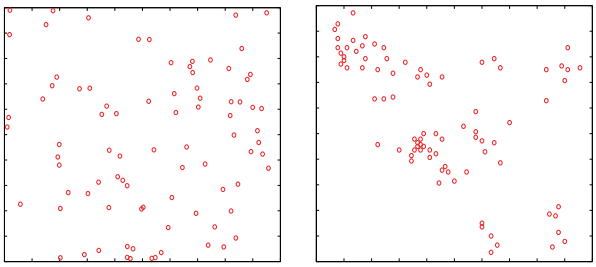
Logistic spatial autoregressive model (SAR): Logistic SAR decomposes a classifier $\hat{f}_C$ into two steps, namely, spatial autoregression and logistic transformation. Spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or the dependent variable, are directly modeled in the regression equation.[81] If the dependent values y_i are related to each other, then the regression equation can be modified as:

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

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

One limitation of the SAR model is that, it does not account for the underlying spatial heteorgeneity that is natural in geographic spaces. Thus, in Eq. (1), the model parameter estimates β and the model errors

 © 2011 John Wiley & Sons, Inc.



(a) CSR pattern (b) Clustered pattern

FIGURE 9 | Complete spatial random (CSR) and spatially clustered patterns.

ϵ are assumed to be uniform throughout the entire geographic space. One proposed method to account for spatial variation in model parameters and errors is *Geographically Weighted Regression* (GWR).[82,83] The regression equation shown for GWR, shown by Eq. (2), has the same structure as standard linear regression, with the exception that the parameters are spatially varying.

$$y = X\beta(s) + \epsilon(s), \tag{2}$$

where $\beta(s)$ and $\epsilon(s)$ represent the spatially varying parameters and the errors, respectively.

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.

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)[56] is synonymous with a homogeneous Poisson process, the patterns of which are independently and uniformly distributed over space, that is, the patterns are equally likely to occur anywhere and do not interact with each other. In contrast, a clustered pattern is distributed dependently and attractively in space.

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

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

Notice in Figure 9 (a) that the CSR pattern seems to exhibit some clustering. This is not an unrepresentative realization but illustrates a well-known property of homogeneous Poisson processes: event-to-nearest-event distances are proportional to χ_2^2 random variables, whose densities have a substantial amount of probability near zero.[56] True clustering, by contrast, is shown in Figure 9(b).

Common methods: Data mining and Machine learning literature have explored a large number of clustering algorithms which compute the statistical significance of spatial clusters to ensure that they are not random. The multitude of clustering algorithms can be classified into several groups as follows:

(1) *Hierarchical* clustering methods start with all patterns as a single cluster and successively perform splitting or merging until a stopping criterion is met. This results in a tree of clusters, called *dendograms*. The dendogram can be cut at different levels to yield desired clusters. Well-known 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). More discussion of these methods can be found in Refs 84–86.

(2) *Partitional* clustering algorithms start with each pattern as a single cluster and iteratively reallocate data points to each cluster until a stopping criterion is met. These methods tend to find clusters of spherical shape. *K-Means* and *K-Medoids* are commonly used partitional algorithms. Squared error is the most frequently used criterion function in partitional clustering. The recent algorithms in this category include partitioning around medoids (PAM), clustering large applications (CLARA), clustering large applications based on randomized search (CLARANS), and expectation-maximization (EM). Related papers include Refs 87 and 88.

(3) *Density-based* clustering algorithms try to find clusters based on the density of data points in a region. These algorithms treat clusters as dense regions of objects in the data space. The density-based clustering algorithms include density-based spatial

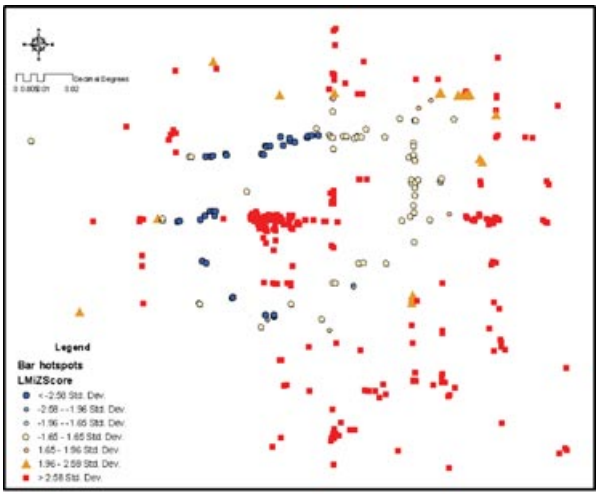
 © 2011 John Wiley & Sons, Inc.

clustering of applications with noise (DBSCAN), ordering points to identify clustering structure (OPTICS), and density-based clustering (DECODE). Related research is discussed in Refs 90–94.

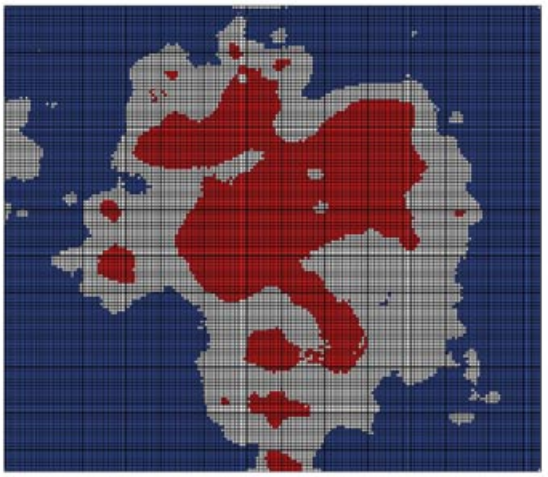
Spatial Hotspot Analysis

Hotspots are a special kind of clustered pattern. As in clustered patterns, objects in hotspot regions have high similarity in comparison to one another and are quite dissimilar to all the objects outside the hotspot. One important feature that distinguishes a hotspot from a general cluster is that the objects in the hotspot area are more active compared with all others (density, appearance, etc.). Spatial correlation of the attribute values within a hotspot could be high and possibly drops dramatically at the boundary, whereas in traditional clustering, the attribute values within a cluster could be *i.i.d*. Hotspot discovery/detection in SDM is a process of identifying spatial regions where more events are likely to happen, or more objects are likely to appear, in comparison to other areas.

Hotspot detection is mainly used in the analysis of crime and disease data. Crime data analysis[95] aims at finding areas that have greater than average numbers of criminal or disorderly events, or areas where people have a higher than average risk of victimization. Figure 10 shows two types of hotspots, namely, point hotspots and area hotspots. The design of hotspot maps is primarily oriented toward aiding law enforcement to make appropriate placement of their resources for crime investigation. For example, Figure 10(b) shows locations of bars with seven different colors obtained by using LISA,[67] the red squares in the center, and peripheries of the map show the high crime activity bars. Maps such as the ones shown in Figure 10(a) show specific bars or hotspots where an increased attention for crime mitigation is necessary. On the other hand, if an analyst was interested in the geographic distribution of a particular crime type (e.g., Vandalism) based on an underlying baseline variable, one can make use of techniques such as kernel density estimation that is a part of tools such as CrimeStat.[96] For example, Figure 10(b) shows the hotspots of vandalism incidents from the same city; the red cells indicate areas where there is a significantly high clustering of vandalism reports and the blue cells indicate cells where there is a significantly low concentration of vandalism, and grey indicates the area where there is no significant concentration. This map leads one to understand that, there is a significant clustering of vandalism incidents in the center of the city around the downtown areas.



(a) Point hotspot of bar locations in Lincoln, NE, USA



(b) Hotspot areas for vandalism

FIGURE 10 | Spatial crime hotspots from the city of Lincoln, NE[89] (Best viewed in color).

Hotspot analysis finds applications in cancer/disease data analysis, hotspots of locations where disease are reported intensively are detected, which may indicate a potential breakout of this disease, or suggest an underlying cause of the disease. Other domains of application include transportation (to identify unusual rates of accidents along highways) and ecology (to conduct geoinformatic surveillance for geospatial hot-spot detection[97]).

Common methods: Many of the standard clustering algorithms have been adapted for spatial hotspot analysis. These include *K*-Means, hierarchical clustering, etc. Many other methods such as STAC (spatio-temporal analysis of crime)[96] and LISA (local indicators of spatial association)[67] have been developed to aid law enforcement agencies for crime

 © 2011 John Wiley & Sons, Inc.

TABLE 2 | Algorithmic Strategies for Classical versus Spatial Data Mining

Classical data mining	Algorithmic strategies for spatial data mining
Divide-and-conquer	Space partitioning
Filter-and-refine	Minimum-bounding rectangle (MBR)
	Predicate approximation
Ordering	Plane sweeping, space filling curve
Hierarchical structures	Spatial index, tree matching
Parameter estimation	Parameter estimation with spatial autocorrelation

mitigation. Spatial hotspot analysis methods of particular utility in public health applications such as syndromic surveillance and outbreak detection have been proposed. These methods include various frequentist and Bayesian statistical measures such as the spatial scan statistic[98,99] and space-time scan statistic.[100,101]

COMPUTATIONAL ISSUES

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

To illustrate these computational challenges, we use the case study of parameter estimation for the SAR model. The massive sizes of geospatial data sets in many application domains make it important to develop scalable parameter estimation algorithms of the SAR model solutions for location prediction and classification. As noted previously, many classical data mining algorithms, such as linear regression, assume that the learning samples are *i.i.d.* This assumption is violated in the case of spatial data due to spatial autocorrelation;[81] in such cases, classical linear regression yields a weak model with not only low prediction accuracy[102,103] but also residual error exhibiting spatial dependence. Modeling spatial dependencies improves overall classification and regression accuracies significantly.

However, estimation of SAR model parameters is computationally very expensive because of the need to compute the determinant of a large matrix in the likelihood function.[104–108] The maximum likelihood (ML) function for SAR parameter estimation contains two terms: a determinant term and an *SSE* term (Eq. 3). The former involves computation of the determinant of a very large matrix, which is a well-known hard problem in numerical analysis. Estimating the parameters of a ML-based SAR model solution, the log-likelihood function can be constructed, as shown in Eq. (3). The estimation procedure involves computation of the logarithm of the determinant (log-det) of a large matrix, that is, $(\mathbf{I} - \rho\mathbf{W})$.

$$\ell(\rho|\mathbf{y}) = \frac{-2}{n}\underbrace{\ln|\mathbf{I} - \rho\mathbf{W}|}_{log-det} + \underbrace{\begin{array}{l}\ln((\mathbf{I} - \rho\mathbf{W})\mathbf{y})^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\,\mathbf{x}^T)^T \\ \times(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\,\mathbf{x}^T)((\mathbf{I} - \rho\mathbf{W})\mathbf{y})\end{array}}_{SSE} \quad (3)$$

As a result, the exact SAR model parameter estimation for a very small 10,000-point spatial problem can take tens of thousands of minutes on common desktop computers. Computation costs make it difficult to use SAR for important spatial problems that involve millions of points, despite its promise to improve prediction and classification accuracy. In the equation, $\mathbf{y}$ is the n-by-1 vector of observations on the dependent variable, where n is the number of observation points; ρ is the spatial autoregression parameter; $\mathbf{W}$ is the n-by-n neighborhood matrix that accounts for the spatial relationships (dependencies) among the spatial data; $\mathbf{x}$ is the n-by-k matrix of observations on the explanatory variable, where k is the number of features; and β is a k-by-1 vector of regression coefficients. Spatial autocorrelation term $\rho\mathbf{W}\mathbf{y}$ is added to the linear regression model in order to model the strength of the spatial dependencies among the elements of the dependent variable, $\mathbf{y}$. The computational bottleneck in accounting for spatial autocorrelation is to evaluate the log-det for large problem sizes. Research in SDM has explored both approximate and exact solutions to the SAR model.[109]

SPATIAL ANALYSIS TOOLS

This section surveys currently existing spatial analysis tools and presents a brief critique of the SDM functionalities they support. Spatial analysis methods including many of the SDM techniques, such as co-location mining, spatial hotspot analysis, and

 © 2011 John Wiley & Sons, Inc.

TABLE 3 | Spatial Analysis Techniques in Popular Software

Technique	Software Tool
Co-location mining	Oracle 10g[110]
Spatial clustering	ArcGIS 9.3 Spatial Statistics tool,[111] Oracle 10g,[110] CrimeStat,[96] Terra Seer
Spatial hotspots	ArcGIS 9.3 Spatial Statistics tool,[111] CrimeStat,[96] GeoDa,[112]
Spatial outliers	ArcGIS 9.3 Spatial Statistics tool,[111] GeoDA[112]
Spatial network hotspots	CrimeStat,[96] SANET[113]
Kriging	ArGIS 9.3 Geostatistical Analyst,[111] S+ Spatial Stats,[114] fields package and geoR in R[115]
Spatial autoregression	S+ Spatial stats,[114] GeoDa[112]
Conditional autoregression	CrimeStat[96]
Geographically weighted regression	ArcGIS 9.3[111]

geographically weighted regression, have found their way into commercial products such as Oracle,[110] ArcGIS,[111] TerraSeer, etc. Beyond these commercial products, there are many public domain and open source tools such as GeoDA,[112] CrimeStat,[96] and many libraries in R[115] that provide useful functionalities for performing spatial autoregression, kriging, and techniques for measuring spatial autocorrelation. Table 3 lists different spatial analysis methods and various tools supporting them.

Many of the above listed tools offer functionalities to perform rigorous significance testing of SDM tools except commercial databases such as Oracle 10g. Added to this, many of the above tools support exploratory analysis with visualization using an interactive display. Despite their usefulness in many applications, tools such as CrimeStat and Oracle 10g are limited in their capabilities to provide interactive map based visualization of results.

FUTURE DIRECTIONS AND RESEARCH NEEDS

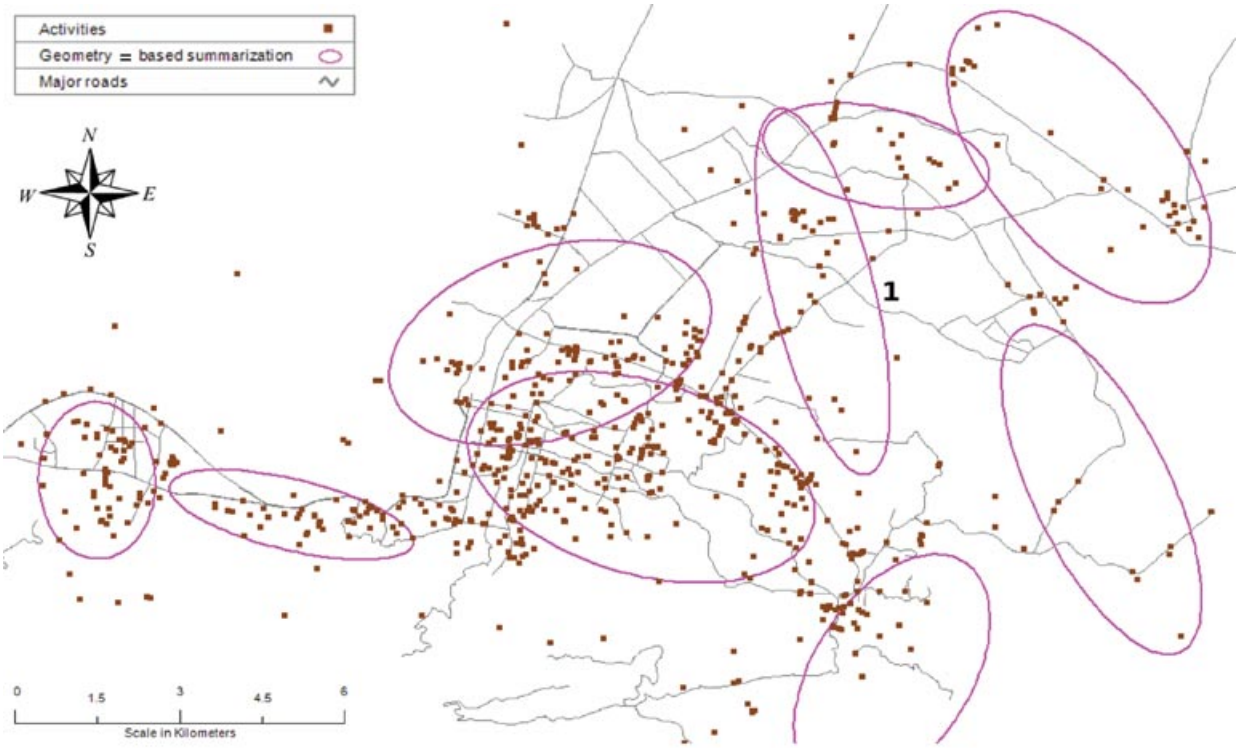
This section presents future directions and research needs in SDM. There are several new areas of research, but the two we will focus on are network-based SDM and spatio-temporal data mining.

Network Patterns

Many spatial phenomena such as distribution of crimes and distribution of accidents in large cities may be constrained by the transportation network structure. One of the main challenges in SDM is to account for the network structure in the data set. For example, in hotspot detection, spatial techniques do not consider the spatial network structure of the data set, that is, they may not be able to model graph properties such as one-ways, connectivities, left-turns, etc. In this section, we present *Spatial network activity hotspots*, an interesting SDM problem that has a spatial network as a part of its input.

Spatial network activity hotspots: The problem of identifying *Spatial network hotspots*(SNAH) is to discover those connected subsets of a spatial network whose attribute values are significantly higher than expected (Figure 11(b)). Finding SNAH is particularly important for crime analysis (high-crime-density street discovery) and law enforcement (planning effective and efficient patrolling strategies). In urban areas, many human activities are centered about spatial infrastructure networks, such as roads and highways, oil/gas pipelines, and utilities (e.g., water, electricity, telephone). Thus, activity reports such as crime logs may often use network-based location references (e.g., street addresses). In addition, spatial interaction among activities at nearby locations may be constrained by network connectivity and network distances (e.g., shortest paths along roads or train networks) rather than the geometric distances used in traditional spatial analysis. Traditional methods that employ a geometric summarization scheme to identify concentrations of crime may not account for large crime concentrations that are normally accounted for by the network-based methods. For example, Figure 11(a) and (b) show a comparison between an ellipse-based geometric hotspot method and a network-based hotspot method for a data set from the recent Haiti earthquake. Crime prevention may focus on identifying subsets of ST networks with high activity levels, understanding underlying causes in terms of network properties, and designing network control policies. Identifying and quantifying *SNAH* is a challenging task due to the need to choose the correct statistical model. In addition, the discovery process in large spatial networks is computationally very expensive due to the difficulty of characterizing and enumerating the population of streets to define

 © 2011 John Wiley & Sons, Inc.

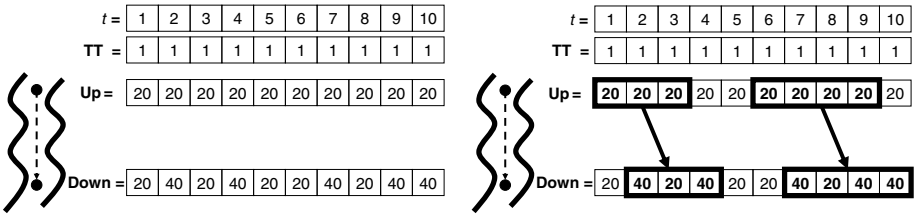


(a) Geometric hotspot detection technique using popular tools[96]

(b) Output using spatial network activity hotspots

FIGURE 11 | Comparison between geometric and network based hotspot for requests during the Haiti earthquake (Best viewed in color).

 © 2011 John Wiley & Sons, Inc.



(a) Example input (b) Example output

FIGURE 12 | Flow anomaly example.

a normal or expected activity level. Preliminary exploration of descriptive and explanatory models for network patterns is available in Ref 116. However, further challenges and research is needed to identify other interesting patterns within network data sets, such as partial segments of roads that are more interesting than other parts.

Spatio-temporal Data Mining

Spatio-temporal data are often modeled using events and processes, both of which generally represent change of some kind. Processes refer to ongoing phenomena that represent activities of one or more types without a specified endpoint.[117–119] Events refer to individual occurrences of a process with a specified beginning and end. Event-types and event-instances are distinguished. For example, a hurricane event-type may occur at many different locations and times, for example, Katrina (New Orleans, 2005) and Rita (Houston, 2005). Each event-instance is associated with a particular occurrence time and location. The ordering may be total if event-instances have disjoint occurrence times. Otherwise, ordering is based on spatio-temporal semantics such as partial order, and spatio-temporal patterns can be modeled as partially ordered subsets. These unique characteristics create new and interesting challenges for discovering spatio-temporal patterns. For example, in contrast to spatial outliers, a spatio-temporal outlier is a spatio-temporal object whose thematic (nonspatial and nontemporal) attributes are significantly different from those of other objects in its spatial and temporal neighborhoods. A spatio-temporal object is defined as a time-evolving spatial object whose evolution or history is represented by a set of instances (EQ), where the space stamp is the location of the object o id at timestamp t. In the remainder of this section, we present research trends in various areas of spatio-temporal data mining.

Flow anomalies: Given a percentage threshold and a set of observations across multiple spatial locations, flow anomaly discovery aims to identify dominant time intervals where the fraction of time instants of significantly mismatched sensor readings exceeds a given percentage threshold. Figure 12 gives a simple example of flow anomalies (FAs). In Figure 12(a), the input to the FA problem consists of two spatial locations [i.e., an upstream (up) and downstream (down) sensor], 10 time instants, and the notion of travel time (TT) or flow between the locations. For simplicity, the TT is set to a constant of 1, but it can be a variable. The output contains two FAs; using the time instants at the upstream sensor, periods 1–3 and 6–9, where the majority of time points show significant differences in between (Figure 12(b)). Discovering FAs is important for water treatment systems, transportation networks, and video surveillance systems. However, mining FAs is computationally expensive due to the large (potentially infinite) number of time instants across a spatial network of locations. Traditional outlier detection methods (e.g. t-test) are suited for detecting transient FAs (i.e., time instants of significant mismatches across consecutive sensors) but cannot detect persistent FAs (i.e., long variable time windows with a high fraction of time instant transient FAs) due to lack of a predetermined window size. Spatial outlier detection techniques do not consider the flow (i.e., TT) between spatial locations and cannot detect any type of FAs. Preliminary work introduced a time-scalable technique called SWEET (Smart Window Enumeration and Evaluation of persistent-Thresholds) that utilizes several algebraic properties in the flow anomaly problem to discover these patterns efficiently.[120–122] However, further research is needed to discover other types of patterns within this environment. In the context of transportation networks, researchers proposed similar ST outlier patterns for identifying traffic accidents known as anomalous window discovery.[123–125]

Teleconnected flow anomalies: An additional pattern that utilizes FAs is teleconnected patterns.[126] A teleconnection represents a strong interaction between paired events that are spatially distant from each other. Identifying teleconnected flow events is computationally hard due to the large number of time instants of measurement, sensors, and locations. For

 © 2011 John Wiley & Sons, Inc.

example, a well-known teleconnected event pair involves the warming of the eastern pacific region (i.e., El Nino) and unusual weather patterns throughout the world.[127] Recently, a RAD (Relationship Analysis of Dynamic-neighborhoods) technique has been proposed that models flow networks to identify teleconnected events.[126] Further research is needed to explore new and interesting patterns that may lie within the RAD model.

Mixed-drove co-occurrence patterns: Another type of dynamic behavior of spatial data sets that might affect co-location patterns is changing the specification of zone of interest and measuring values according to user preferences. Mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) represent subsets of two or more different object-types whose instances are often located in spatial and temporal proximity. Discovering MDCOPs is potentially useful in identifying tactics in battlefields and games, understanding predator–prey interactions, and in transportation (road and network) planning.[128,129] However, mining MDCOPs is computationally very expensive because the interest measures are computationally complex, data sets are larger due to the archival history, and the set of candidate patterns is exponential in the number of object-types. Preliminary work has produced a monotonic composite interest measure for discovering MDCOPs and novel MDCOP mining algorithms.[130]

Cascading spatio-temporal patterns Partially ordered subsets of event-types whose instances are located together and occur in stages are called cascading spatio-temporal patterns (CSTP).[131] Figure 13 shows some interesting partially ordered patterns that were discovered from real spatio-temporal crime data sets from the city of Lincoln, Nebraska.[89] In the domain of public safety, events such as bar closings and football games are considered generators of crime. Preliminary analysis revealed that football games and bar closing events do indeed generate CSTPs. CSTP discovery can play an important role in disaster planning, climate change science[132,133] (e.g., understanding the effects of climate change and global warming), and public health (e.g., tracking the emergence, spread, and re-emergence of multiple infectious diseases[134]). Further research is needed, however, to deal with challenges such as the lack of computationally efficient, statistically meaningful metrics to quantify interestingness, and the large cardinality of candidate pattern sets that are exponential in the number of event types. Existing literature for spatio-temporal data mining focuses on mining totally ordered sequences or unordered subsets.[135–137]

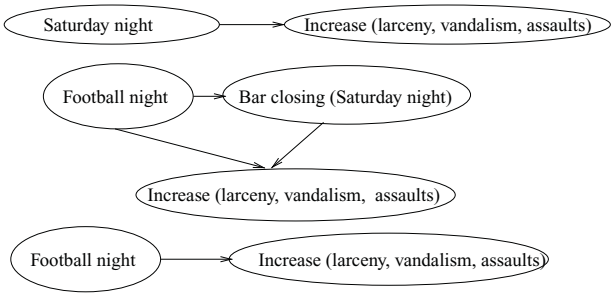


FIGURE 13 | Cascading spatio-temporal patterns from public safety.

Broader Future Directions

In this paper, we have presented the major research achievements and techniques that have emerged from SDM, especially for predicting locations and discovering spatial outliers, co-location rules, and spatial clusters. Current research is mostly concentrated on developing algorithms that model spatial and spatio-temporal autocorrelations and constraints. Spatio-temporal data mining remains, however, still largely an unexplored territory; thus, we conclude by noting other areas of research that require further investigation, such as the mining of movement data involving groups of people, ideas, goods, and streaming data. Any SDM method is influenced by the neighborhood method selected. Hence, new computational algorithms and interest measures that deal with the sensitivity to spatial neighborhood size need to be defined. Most urgently, methods are needed to validate the hypotheses generated by SDM algorithms as well as to ensure that the knowledge generated is actionable in the real world.

ACKNOWLEDGEMENTS

We are particularly grateful to our collaborators Prof. Vipin Kumar, Prof. Paul Schrater, Prof. Sanjay Chawla, Dr. Chang-Tien Lu, Dr. Weili Wu, and Prof. Uygar Ozesmi, Prof. Yan Huang,

 © 2011 John Wiley & Sons, Inc.

and Dr. Pusheng Zhang for their various contributions. We are particularly grateful to Dev Oliver, Xun Zhou, and Zhe Jiang for their feedback and help in preparing this paper. We also thank Prof. Hui Xiong, Prof. Jin Soung Yoo, Dr. Baris Kazar, Prof. Mete Celik, Dr. Betsy George, Dr. R.R Vatsavai and anonymous reviewers for their valuable feedback on early versions of this paper. We would like to thank Kim Koffolt for improving the readability of this paper.

REFERENCES

1. Güting R. An Introduction to Spatial Database Systems. In: Schek HJ, ed. *VLDB Journal*. London, UK: Springer-Verlag; 1994, 3:357–399.
2. Shekhar S, Chawla S. Spatial databases: A tour. *Prentice Hall*. NJ, USA: Upper Saddle River; 2002. ISBN: 0-7484-0064-6.
3. Shekhar S, Chawla S, Ravada S, Fetterer A, Liu X, Lu C-T. Spatial Databases - Accomplishments and Research Needs. *Trans. on Knowledge and Data Engineering* 1999, 11:45–55.
4. Worboys M, Duckham M. *GIS: A Computing Perspective. Second Edition*. Boca Raton, FL, USA: CRC Press; 2004. ISBN: 978-0415283755.
5. Stolorz P, Nakamura H, Mesrobian E, Muntz R, Shek E, Santos J, Yi J, Ng K, Chien S, Mechoso R, Farrara J. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: AAAI Press; 1995, 300–305.
6. Chawla S, Shekhar S, Wu W-L, Ozesmi U. Modeling spatial dependencies for mining geospatial data. In: *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* 2000, 70–77.
7. Albert P, McShane L. A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. *Biometrics* 1995, 51:627–638.
8. Cromley E, McLafferty S. *GIS and public health*. The Guilford Press, 2002.
9. Elliott P, Wakefield J, Best N, Briggs D. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2000. ISBN: 978-0192629418.
10. Eck JE, Chainey S, Cameron JG, Leitner M, Wilson RE. Mapping Crime: Understanding Hot Spots. National Insitute of Justice (NIJ) Special Report NCJ 209393, U.S. Department of Justice, Washington D.C., USA.
11. Lang L. *Transportation GIS*. Redlands, CA, USA: ESRI Press; 1999. ISBN: 978-1879102471.
12. Leipnik MR, Albert DP. *GIS in Law Enforcement: Implementation Issues and Case Studies*. London/New York: CRC Press; 2002. ISBN: 978-0415286107.
13. Shekhar S, Yang T, Hancock P. An Intelligent Vehicle Highway Information Management System. *Intl J Microcomputers in Civil Engineering*. New York: John Wiley & Sons; 1993, 8.
14. Thill J. Geographic information systems for transportation in perspective. *Transportation Research Part C: Emerging Technologies* 2000, 8:3–12.
15. Issaks E, Svivastava M. *Applied Geostatistics*. Oxford: Oxford University Press; 1989.
16. Kanevski M, Parkin R, Pozdnukhov A, Timonin V, Maignan M, Demyanov V, Canu S. Environmental data mining and modeling based on machine learning algorithms and geostatistics. *Environmental Modelling & Software* 2004, 19:845–855.
17. Haining RJ. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press; 1989.
18. Roddick J-F, Spiliopoulou M. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. *SIGKDD Explorations* 1999, 1:34–38.
19. Scally R. *GIS for Environmental Management*. Redlands, CA, USA: ESRI Press; 2006. ISBN: 978-589481429.
20. Krugman P. *Development, Geography, and Economic Theory*. Cambridge, MA: MIT Press; 1995.
21. Ganguly A, Steinhaeuser K. Data mining for climate change and impacts. In: *IEEE International Conference on Data Mining Workshops, 2008. ICDMW'08* 2008, 385–394.
22. Yasui Y, Lele S. A regression method for spatial disease rates: An estimating function approach. *J Am Stat Assoc* 1997, 94:21–32.
23. Calkins H. GIS and public policy. *Geographical Information Systems* 1990, 2:233–245.
24. Greene R. *GIS in public policy: Using geographic information for more effective government*. Radlands, CA, USA: ESRI Press, 2000. ISBN: 978-1879102668.
25. Hohn M, Liebhold LGAE. A Geostatistical model for Forecasting the Spatial Dynamics of Defoliation caused by the Gypsy Moth, Lymantria dispar (Lepidoptera:Lymantriidae). *Environmental Entomology (Publisher: Entomological Society of America)* 1993, 22:1066–1075.
26. Maguire D. Implementing spatial analysis and GIS applications for business and service planning. In: Longley PA, Clarke G, eds. *GIS for Business and Service Planning*. New York: John Wiley & Sons; 1995, 171–191. ISBN: 978-0470235102.

 © 2011 John Wiley & Sons, Inc.

27. Mennecke B. Understanding the role of geographic information technologies in business: Applications and research directions. *Journal of Geographic Information and Decision Analysis* 1997, 1:44–68.
28. Pick J. *Geographic information systems in business*. Hershey, PA, USA/London, UK: Idea Group Publishing; 2004.
29. Burrough PA, McDonnell RA. Principles of Geographical Information Systems, Spatial Information Systems and Geostatistics. Oxford University Press: Oxford, UK. 1998. ISBN: 978-0198233657.
30. Savic D, Walters G. Hydroinformatics, data mining and maintenance of UK water networks. *Anticorrosion methods and materials* 1999, 46:415–425.
31. Wu H, Lu C. A data mining approach for spatial modeling in small area load forecast. *IEEE Trans Power Syst*. Piscatway, NJ, USA: IEEE Power and Energy Society; 2002, 17(2):516–521.
32. Schiller J. *Location-Based Services*. San Francisco, CA: Morgan Kaufmann, 2004. ISBN: 978-1558609296.
33. Stefanidis A, Nittel S. *GeoSensor Networks*. Boca Raton, FL: CRC; 2004. ISBN: 978-0415324045.
34. Arctur DK, Zeiler M. *Designing Geodatabases*. Redlands, CA: ESRI Press; 2004. ISBN: 158948021X.
35. Beinat E, Godfrind A, Kothuri RV. *Pro Oracle Spatial Apress*. Springer-Verlag; New York/Hiedelberg, USA/Germany; 2004. ISBN: 978-1590593837.
36. SQL Server 2008 R2, Microsoft Corporation, Redmond, WA, 2010. URL: http://www.microsoft.com/sql/prodinfo/futureversion/default.mspx.
37. Google Earth 6, Google Inc, Mountain View, CA, USA, 2010. URL: http://earth.google.com.
38. Bing Maps, Microsoft Corporation, Redmond, WA, USA, 2010, URL: http://www.microsoft.com/maps/.
39. PostGIS 1.5.2, Refractions Research, September 2010, URL: http://postgis.refractions.net/.
40. MySQL Spatial Extensions, Oracle Corporation, Redwood City, CA, 2011. URL: http://dev.mysql.com/doc/refman/5.5/en/.
41. Sloa Digital Sky Survey Data Release 8, Astrophysical Research Consortium (ARC), Seattle, WA URL: http://www.sdss3.org/dr8/.
42. Chamberlin D. Using the New DB2: IBM's Object Relational System. San Francisco, CA: Morgan Kaufmann; 1997. ISBN: 978-1558603738.
43. Stonebraker M, Moore D. Object Relational DBMSs: The Next Great Wave. San Francisco, CA: Morgan Kaufmann, 1997. ISBN: 978-1558603974.
44. OGC Standards and Specifications, Open Geospatial Consortium (OGC), Wayland, MA, 2011. URL: http://www.opengeospatial.org/standards.
45. Frank E, Hall MA, Holmes G, Kirkby R, Pfahringer B. Weka A machine learning workbench for data mining. In: Maimon O, Rokach L, eds. *The Data Mining and Knowledge Discovery Handbook*. London, UK: Springer-Verlag; 2005, 1305–1314.
46. Maimon O, Rokach L. *Data Mining and Knowledge Discovery Handbook*. New York: Springer-Verlag; 2005.
47. Witten I, Frank E. *Data Mining: Practical machine learning tools and techniques*. Sanfrancisco, USA: Morgan Kaufmann Pub, 2005.
48. Bolstad P. *GIS Fundamentals: A First Text on GIS*. Eider Press, 2002.
49. Ganguly AR, Steinhaeuser K. Data mining for climate change and impacts. In: *ICDM Workshops* 2008, 385–394.
50. Agarwal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S. eds. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington, DC: ACM Press; 1993, 207–216.
51. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca JB, Jarke M, Zaniolo C, eds. *Proceedings of 20th International Conference on Very Large Databases*, Kauffman Morgan, Sanfrancisco, CA, USA; 1994, 487–499. ISBN: 1-55860-153-8.
52. Bolstad P. GIS Fundamentals: A first Text on GIS. Eider Press, St. Paul, MN, ISBN: 978-0971764729.
53. Quinlan J. *C4.5: Programs for Machine Learning* Morgan Kaufmann; San Mateo, CA, 1993. ISBN: 978-1558602380.
54. Varnett V, Lewis T. *Outliers in Statistical Data*. John Wiley & Sons; 1994. ISBN: 978-0471930945.
55. Banerjee S, Carlin B, Gelfand A. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall; 2004 ISBN: 978-1584884101.
56. Cressie N. *Statistics for Spatial Data (Revised Edition)*. New York: John Wiley & Sons, 1993.
57. Bailey T, Gatrell A. *Interactive Spatial Data Analysis*. Essex: Longman Harlow; 1995.
58. Fotheringham A, Brunsdon C, Charlton M. *Geographically weighted regression: the analysis of spatially varying relationships*. New York: John Wiley & Sons, 2002.
59. Tobler WR. A computer movie simulating urban growth in the Detroit Region. Economic Geography, Vol. 46, Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods, Clark University, Worcester, MA, 1970. 234–240.
60. Barnett V, Lewis T. *Outliers in Statistical Data*. 3rd ed. New York: John Wiley & Sons, 1994.
61. Hawkins D. *Identification of Outliers*. Chapman and Hall, 1980.
62. Pei Y, Zaïane OR, Gao Y. An Efficient Reference-Based Approach to Outlier Detection in Large

 © 2011 John Wiley & Sons, Inc.

Datasets. In: *Proceedings of the Sixth International Conference on Data Mining*, (ICDM'06). Washington, DC: IEEE Computer Society; 2006, 478–487.

63. Sun P, Chawla S. On local spatial outliers. In: *ICDM* 2004, 209–216.
64. Wu W, Cheng X, Ding M, Xing K, Liu F, Deng P. Localized outlying and boundary data detection in sensor networks. *IEEE Trans Knowl Data Eng* 2007, 19:1145–1157.
65. Shekhar S, Lu CT, Zhang P. Detecting graph-based spatial outliers. In: Family A, ed. *Intelligent Data Analysis*. Amsterdam, the Netherlands: IOS Press; 2002, 451–468.
66. Haslett J, Bradley R, Craig P, Unwin A, Wills G. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician* 1991, 234–242.
67. Anselin L., Getis A. Local indicators of spatial association-lisa. *Geographical Analysis* 1995, 27:93–155.
68. Anselin L, Getis A. Spatial statistical analaysis and geographic information systems. The Annals of Regional Science, Volume 26, Number 1, 1992, Springer-Verlag: Heidelberg, 1992, 19–33.
69. Lu C-T, Chen D, Kou Y. Detecting spatial outliers with multiple attributes. In: *ICTAI '03: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, page 122, Washington, DC: IEEE Computer Society; 2003.
70. Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer MJ, Herring JR, eds. *Proceedings of the 4th International Symposium on Advances in Spatial Databases, 1995* (SSD'95). London, UK: Springer-Verlag; 1995, 47–66.
71. Morimoto Y. Mining Frequent Neighboring Class Sets in Spatial Databases. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001* (ACM SIGKDD'01). New York: ACM; 2001, 353–358.
72. Shekhar S, Huang Y. Discovering Spatial Colocation Patterns: A Summary of Results. In: Jensen CS, Schneider M, Seeger B, Tsotras VJ, eds. *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases, 2001* (SSTD'01). London, UK: Springer-Verlag; 2001, 236–256.
73. de Almeida CM, Souza IM, Alves CD, Pinho CMD, Pereira MN, Feitosa RQ. Multilevel object-oriented classification of quickbird images for urban population estimates. In: *GIS* 2007, 12.
74. Little B, Schucking M, Gartrell B, Chen B, Ross K, McKellip R. High Granularity Remote Sensing and Crop Production over Space and Time: NDVI over the Growing Season and Prediction of Cotton Yields at the Farm Field Level in Texas. In: *Proceedings of the IEEE International Conference on Data Mining Workshops, 2008* (ICDMW'08). Washington, DC: IEEE Computer Society; 2008, 426–435.
75. Jhung Y, Swain PH. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 1996, 34:67–75.
76. Solberg AH, Taxt T, Jain AK. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing* 1996, 34:100–113.
77. Hixson M, Scholz D, Funs N. Evaluation of several schemes for classification of remotely sensed data. *Photogrammetric Engineering & Remote Sensing* 1980, 46:1547–1553.
78. Strahler A. The use of prior probabilities in maximum likelihood classification of remote sensing data. *Remote Sensing of Environment* 1980, 10:135–163.
79. Li SZ. Markov Random field modeling in image analysis. London, UK: Springer, 2001, ISBN: 978-4431703099.
80. Shekhar S, Schrater PR, Vatsavai RR, Wu W, Chawla S. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia* 2002, 4.
81. Anselin L. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer, Dordrecht; 1988.
82. Brunsdon C, Fotheringham S, Charlton M. Geographically weighted regression-modelling spatial non-stationarity. *The Statistician* 1998, 47:431–443.
83. Fotheringham A, Charlton M, Demšar U. Looking for a relationship? Try GWR. In: Miller HJ, Han J. eds. *Geographic Data Mining and Knowledge Discovery* CRC Press, Boca Raton, FL, USA; 2009, 227–254 ISBN: 978-1-4200-7397-3.
84. Cervone G, Franzese P, Ezber Y, Boybeyi Z. Risk assessment of atmospheric hazard releases using k-means clustering. In: *ICDM Workshops* 2008, 342–348.
85. Karypis G, Han E-H, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer* 1999, 32:68–75.
86. Zhang T, Ramakrishnan R, Livny M. Birch: An efficient data clustering method for very large databases. In: Jagadish, Mumick IS, eds. *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Montreal, Quebec, June 4–6*. ACM Press; 1996, 103–114.
87. Hu T, Xiong H, Gong X, Sung SY. Anemi: An adaptive neighborhood expectation-maximization algorithm with spatial augmented initialization. In: *PAKDD* 2008, 160–171.

 © 2011 John Wiley & Sons, Inc.

88. Ng RT, Han J. Clarans: A method for clustering objects for spatial data mining. *IEEE Trans Knowl Data Eng* 2002, 14:1003–1016.
89. Lincoln city police department, Lincoln city crime records. 2008. URL: http://www.lincoln.ne.gov/city/police/.
90. Wang W, Yang J, Muntz, R. *Sting*: A statistical Information Grid Approach to Spatial Data Mining. In: Jarke M, Carey MJ, Dittrich KR, Lochovsky FH, Loucopoulos P, Jeusfeld MA, eds. *Proceedings of 23rd International Conference on Very Large data bases* (VLDB'97). *August 25–29, Athens, Greece*. Massachusetts: Morgan Kaufmann; Sanfrancisco, USA; 1997, 186–195. ISBN: 1-55860-470-7.
91. Lai C, Nguyen NT. Predicting Density-Based Spatial Clusters Over Time. In: *Proceedings of the Fourth IEEE International Conference on Data Mining* (ICDM'04). Washington, DC: IEEE Computer Society; 2004, 443–446.
92. Ma D, Zhang A. An adaptive density-based clustering algorithm for spatial database with noise. In: *ICDM* 2004, 467–470.
93. Neill D, Moore AW. Rapid detection of significant spatial clusters. In: Kim W, Kohavi R, Gehrke J, DuMouchel W. eds. *Proceedings of the 10th international conference on Knowledge discovery and data mining* (KDD '04). New York: ACM Press; 2004, 256–265. ISBN 1-58113-888-1.
94. Pei T, Jasra A, Hand DJ, Zhu A-X, Zhou C. Decode: a new method for discovering clusters of different densities in spatial data. *Data Min Knowl Discov* 2009, 18:337–369.
95. van Eck NJ, Waltman L. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2007, 15:625–645.
96. Levine N. CrimeStat: A spatial statistics program for the analysis of crime incident locations (v 3.3). *Ned Levine & Associates, Houston, TX, and the National Institute of Justice*. Washington, DC, 2010.
97. Janeja VP, Adam NR. Geo Spatial Data Mining techniques for Homeland Security Applications. In: Shekhar S, Xiong H, eds. *Encyclopedia of Geographic Information Science*. London, UK: Springer-Verlag; 2008, 434–440.
98. Kulldorff M. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 1997, 26:1481–1496.
99. Neill D, Moore A, Cooper G. A Bayesian spatial scan statistic. *Advances in neural information processing systems* 2006, 18:1003.
100. Kulldorff M, Athas W, Feurer E, Miller B, Key C. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health* 1998, 88:1377.
101. Neill D, Moore A, Pereira F, Mitchell T. Detecting significant multidimensional spatial clusters. *Advances in Neural Information Processing Systems* 2005, 17:969–976.
102. Chawla S, Shekhar S, Wu W, Ozesmi U. Modeling spatial dependencies for mining geospatial data. *1st SIAM International Conference on Data Mining*, 2001.
103. Shekhar S, Schrater P, Raju R, Wu W. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia* 2002, 4:174–188.
104. Kazar B, Shekhar S, Lilja D, Boley D. A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets. *SIAM International Conf. on Data Mining Workshop on High Performance and Distributed Mining (HPDM2004)*, April 2004.
105. Li B. Implementing spatial statistics on parallel computers. *Practical Handbook of Spatial Statistics. CRC Press* 1996, 107–148.
106. Pace RK, Zou D. Closed-Fprm Maximum Likelihood Estimates of Nearest Neighbor Spatial Dependence. *Geographical Analysis* 2000, 32:154–172.
107. Pace R, LeSage J. Semiparametric maximum likelihood estimates of spatial dependence. *Geographical Analysis* 2002, 34:76–90.
108. Pace R, LeSage J. Simple bounds for difficult spatial likelihood problems. *http://www.spatial-statistics.com*, 2003.
109. Kazar B, Shekhar S, Lilja D, Vatsavai R, Pace R. Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis. In: Egenhofer MJ, Freska C, Miller HJ, eds. *Proceedings of Third International Conference on Geographic Information Science*, (GIScience 2004), Lecture Notes in Computer Science. London, UK: Springer-Verlag; 2004.
110. Murray Cea. Oracle® Spatial Users Guide and Reference 10g Release 1 (10.1). *Redwood City, Oracle Corporation* 2003, 602.
111. Scott L, Janikas M. Spatial statistics in ArcGIS. In Handbook of Applied Spatial Analysis. Fischer MM, Getis A. eds., London, UK: Springer, 2010, 27–41. ISBN: 978-3642036460.
112. Anselin L. GeoDa 0.9 users guide. *Urbana* 2003, 51:61801.
113. Okabe A, Okunuki K, Shiode S. SANET: A toolbox for spatial analysis on a network. *Geographical Analysis* 2006, 38:57–66.
114. Kaluzny S. S+ SpatialStats: user's manual for Windows and UNIX. London, UK: Springer, 1998, ISBN: 978-0387982267.
115. Ribeiro P Jr, Diggle P. geoR: A package for geostatistical analysis. *R News* 2001, 1:14–18.

 © 2011 John Wiley & Sons, Inc.

116. Celik M, Shekhar S, George B, Rogers JP, Shine JA. Discovering and quantifying mean streets: A summary of results. Technical Report 025, University of Minnesota, 2007.
117. Allen JF. Towards a general theory of action and time. *Artif Intell* 1984, 23:123–154.
118. Shekhar S, Xiong H, editors. *Encyclopedia of GIS*. Springer, 2008.
119. Worboys MF. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science* 2005, 19:1–28.
120. Kang JM, Shekhar S, Wennen C, Novak P. Discovering Flow Anomalies: A SWEET Approach. In: *International Conference on Data Mining*, 2008.
121. Franke C, Gertz M. Detection and exploration of outlier regions in sensor data streams. In: *ICDM Workshops* 2008, 375–384.
122. Elfeky MG, Aref WG, Elmagarmid AK. Stagger: Periodicity mining of data streams using expanding sliding windows. *Data Mining, IEEE International Conference* (ICDM'06). Washington, DC: IEEE Computer Society; 2006, 188–199.
123. Dey S, Janeja V, Gangopadhyay A. Temporal Neighborhood Discovery Using Markov Models. In: *2009 Ninth IEEE International Conference on Data Mining* 2009, 110–119. IEEE.
124. Janeja V, Adam N, Atluri V, Vaidya J. Spatial neighborhood based anomaly detection in sensor datasets. *Data Mining and Knowledge Discovery* 2010, 20:221–258.
125. Shi L, Janeja V. Anomalous window discovery through scan statistics for linear intersecting paths (sslip). In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (SIGKDD'09). New York: ACM; 2009, 767–776.
126. Kang JM, Shekhar S, Henjum M, Novak P, Arnold W. Discovering Teleconnected Flow Anomalies: A Relationship Analysis of spatio-temporal Dynamic (RAD) neighborhoods. In: Mamoulis N, Seidl T, Pedersen TB, Torp K, Assent I, eds. *Proceedings of 11th International Symposium on Advances in Spatial and Temporal Databases* (SSTD'09), Lecture Notes in Computer Science 5644. London, UK: Springer-Verlag; 2009.
127. Pastor R. The El Nino Story. Pacific Marine Environmental Laboratory (PMEL), National Oceanic and Atmospheric Administration (NOAA), US Department of Commerce. URL: http://www.pmel.noaa.gov/tao/elnino/el-ninostory.html.
128. Guting R, Schneider M. *Moving Object Databases*. Morgan Kaufmann, 2005.
129. Koubarakis M, Sellis T, Frank A, Grumbach S, Guting R, Jensen C, Lorentzos N, Schek HJ, Scholl M. eds. *Spatio-Temporal Databases: The Chorochronos Approach, LNCS 2520*. London, UK: Springer-Verlag; 2003, 9:352 ISBN: 978-3540405528.
130. Celik M, Shekhar S, Rogers JP, Shine JA, Yoo JS. Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In: *Proceedings of the Sixth International Conference on Data Mining*, ICDM'06. Washington, DC: IEEE Computer Society; 2006, 119–128.
131. Mohan P, Shekhar S, Shine JA, Rogers JP. Cascading Spatio-temporal pattern discovery: A summary of results. In: *Proceedings of 10th SIAM International Conference on Data Mining*, (SDM'10). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM); 2010, 327–338.
132. Committee on Strategic Advice on the US Climate Change Science Program; National Research Council: Restructuring Federal Climate Research to Meet the Challenges of Climate Change. Washington DC: The National Academies Press; 2009.
133. Frelich LE, Reich PB. Will environmental changes reinforce the impact of global warming on the prairie-forest border of central north america? *Frontiers in Ecology and the Environment*, Ithaca, NY, Ecological Society of America Journals, Issue 7, 2010, 371–378.
134. Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. *Nature* 2004, 430:242–249.
135. Celik M, Shekhar S, Rogers JP, Shine JA. Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering* 2008, 20:1322–1335.
136. Huang Y, Zhang L, Zhang P. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering* 2008, 20:433–448.
137. Wang J, Hsu W, Lee ML. A framework for mining topological patterns in spatio-temporal databases. In: *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. New York: ACM. 2005, 429–436.

 © 2011 John Wiley & Sons, Inc.