Geoinformatica (2008) 12:455–475
DOI 10.1007/s10707-007-0038-8

On Detecting Spatial Outliers

Dechang Chen · Chang-Tien Lu · Yufeng Kou · Feng Chen

Received: 29 March 2006 / Revised: 7 August 2007 /
Accepted: 26 September 2007 / Published online: 23 October 2007
© Springer Science + Business Media, LLC 2007

Abstract The ever-increasing volume of spatial data has greatly challenged our ability to extract useful but implicit knowledge from them. As an important branch of spatial data mining, spatial outlier detection aims to discover the objects whose non-spatial attribute values are significantly different from the values of their spatial neighbors. These objects, called spatial outliers, may reveal important phenomena in a number of applications including traffic control, satellite image analysis, weather forecast, and medical diagnosis. Most of the existing spatial outlier detection algorithms mainly focus on identifying single attribute outliers and could potentially misclassify normal objects as outliers when their neighborhoods contain real spatial outliers with very large or small attribute values. In addition, many spatial applications contain multiple non-spatial attributes which should be processed altogether to identify outliers. To address these two issues, we formulate the spatial outlier detection problem in a general way, design two robust detection algorithms, one for single attribute and the other for multiple attributes, and analyze their computational complexities. Experiments were conducted on a real-world data set, West Nile virus data, to validate the effectiveness of the proposed algorithms.

Keywords algorithm · outlier detection · spatial data mining

D. Chen
Department of Preventive Medicine and Biometrics,
Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
e-mail: dchen@usuhs.mil

C.-T. Lu · Y. Kou · F. Chen (✉)
Department of Computer Science, Virginia Polytechnic Institute and State University,
7054 Haycock Road, Falls Church, VA 22043, USA
e-mail: chenf@vt.edu

C.-T. Lu
e-mail: ctlu@vt.edu

Y. Kou
e-mail: ykou@vt.edu

1 Introduction

Outlier detection is one of the major tasks of data mining that aims to identify abnormal patterns (outliers) from large data sets. In different applications, outliers have different names such as anomalies, deviations, exceptions, faults, and irregularities. Although there is no consensus to describe outliers, Barnet's definition is accepted by many statisticians and computer scientists, which views an outlier as one observation that appears to deviate markedly from other members of the sample in which it occurs [4]. In the past decades, outlier detection has attracted substantial attention and distinguished itself as an important branch of data mining. Traditional outlier detection has many practical applications. For example, it can help identify intrusions in computer networks [41], locate malfunctioned parts in a manufacture streamline [10], pinpoint suspicious usages of credit cards [8], and monitor unusual changes of stock prices [9].

In recent years, the existence of huge amount of spatial data calls for spatial outlier detection methods to identify anomalies in the spatial context. A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood [33]. It is usually viewed as a local anomaly whose nonspatial attribute values are extreme to its neighbors [7]. In contrast to traditional outliers, spatial outliers do not necessarily deviate from the remainder of the whole data set. Informally, traditional outliers can be called "global outliers" since they are based on global comparisons, while spatial outliers can be called "local outliers" since they are derived from local comparisons. Spatial outlier detection plays an important role in many applications. It can help locate extreme meteorological events such as tornadoes and hurricanes [22], [43], identify disease outbreaks [38] and tumor cells [28], and discover abnormal highway traffic patterns [35]. In addition, spatial outlier detection can be potentially applied to pinpoint significant military targets in satellite images, determine the locations of potential gas/oil wells, and detect water pollution incidents.

Traditional outlier detection approaches may not be directly applied to extract abnormal spatial patterns due to the special properties of spatial data. First, classical outlier detection is designed to process numerical and categorical data, whereas spatial data have more complex structures that contain extended objects such as points, lines, and polygons. Second, traditional outlier detection does not separate spatial relationships among input variables, while spatial patterns often exhibit spatial continuity and autocorrelation with nearby samples. As indicated by the geographical rule of thumb, "Everything is related to everything else, but nearby things are more related than distant things [37]." In the identification of spatial outliers, the attribute space is generally divided into two parts, non-spatial attributes and spatial attributes. Spatial attributes record the spatial information such as locations, boundaries, and directions, which determine the spatial relationships among neighbors. Based on the spatial neighborhood relationship, non-spatial attributes are used to identify abnormal observations.

Several methods have been proposed for spatial outlier detection. However, most of the existing spatial outlier detection algorithms mainly focus on identifying single attribute outliers and could potentially misclassify normal objects as outliers when their neighborhoods contain real spatial outliers with very large or small attribute values. In addition, many spatial applications contain multiple non-spatial

attributes which should be processed altogether to identify outliers. To address these issues, we introduce two effective algorithms for the single-attribute and multiple-attribute spatial outlier detection. These two algorithms are robust in the sense that their performance is not subject to the number of outliers that exist in the data. Experiments conducted on West Nile virus data demonstrate the effectiveness of the proposed algorithms.

The contributions of this paper are: (1) formally define the problem of spatial outlier detection; (2) design and implement one robust algorithm for the single-attribute outlier detection; (3) develop one Mahalanobis-distance-based algorithm to detect spatial outliers with multiple attributes; (4) evaluate the effectiveness of the proposed algorithms by experimenting on the West Nile virus data.

The paper is organized as follows. Section 2 reviews related work. Section 3 provides a general framework on the single attribute outlier detection, discusses the deficiency of the existing approaches, and proposes one algorithm to identify single attribute spatial outliers. In Section 4, the algorithm for detecting spatial outliers with multiple attributes is presented. Section 5 presents and analyzes the experimental results. We conclude in Section 6 with directions for future work.

2 Related work

The existing traditional outlier detection algorithms can be classified into the following categories: clustering-based, distribution-based, depth-based, density-based, and distance-based. A few clustering-based algorithms have been designed to identify outliers as exceptional data points that do not belong to any cluster [12], [26], [42]. Since these algorithms are not specifically designed for outlier detection, their efficiency and effectiveness are not optimized. Distribution-based methods use a standard distribution to fit the data set so that data points deviating from this distribution are defined as outliers [40]. The primary limitation of these methods is that in many applications, the exact distribution of a data set is unknown beforehand. Depth-based methods organize the data in different layers of k-d convex hulls where data in the outer layers tend to be outliers [29], [32]. These methods are not widely used due to their high computation costs for multi-attribute data. Density-based algorithms define outliers in terms of their local reachability densities [6], [17]. Local outlier factor (LOF) is a typical example of density based algorithms which evaluate the outlierness of an object by comparing its density with those of its neighbors. Distance-based methods may be the most widely used techniques which define an outlier as a data point having an exceptionally far distance to the other data points [18], [30].

The above methods for detecting outliers focus on low dimensional data. For detecting outliers with numerous attributes, traditional outlier detection approaches are ineffective due to the "curse of high dimensionality," i.e., the sparsity of the data objects in a high dimensional space [3]. It has been shown that the distance between any pair of data points in a high dimensional space is so similar that either every data point or none of the data points can be viewed as an outlier if the concept of proximity is used to define outliers [1]. As a result, traditional Euclidean distance cannot be used to effectively detect outliers in high dimensional data sets. Two categories of research work have been conducted to address this issue. One is to

project high dimensional data to low dimensional data [2], [3], [5], [16], and the other is to re-design distance functions to accurately define the proximity relationship between data points [1].

Traditional outlier detection algorithms can be applied to spatial data. However, their performance is not assured since they treat spatial attributes and non-spatial attributes equally. For spatial outlier detection, spatial and non-spatial dimensions should be considered separately. The spatial dimension is used to define the neighborhood relationship, while the non-spatial dimension is often used to define the discrepancy quantity. A number of algorithms have been specifically designed to deal with spatial data. These methods can be generally grouped into two categories, namely, graphic approaches and quantitative tests. Graphic approaches are based on visualization of spatial data which highlights spatial outliers. Examples include variogram clouds and pocket plots [15], [27]. Quantitative methods provide tests to distinguish spatial outliers from the remainders of the data set. Scatterplot [13] and Moran scatterplot [23] are two representative approaches. A Scatterplot shows the attribute value on the X-axis and the average of the attribute values over the neighborhood on the Y-axis. Nodes far away from the least square regression line are flagged as potential spatial outliers. A Moran scatterplot is a plot of normalized attribute value against the neighborhood average of normalized attribute values. It contains four quadrants where spatial outliers can be identified from the upper left and lower right quadrants. Other work involving quantitative approaches includes the graph-based outlier detection [35] and the locally adaptive statistical analysis [19]. Recently, some neighborhood-based approaches have been proposed to detect spatial outliers with single or multiple attributes [20], [21]. These approaches mainly work for data that do not contain a large number of outliers.

3 Detection of outliers with a single attribute

In this section, we define the problem of detecting spatial outliers with a single attribute, discuss deficiencies of some existing detection methods, then introduce a new detection algorithm. The computational complexity of the proposed algorithm is also examined.

3.1 Problem formulation

Suppose there exists a set of spatial points $X = \{x_1, x_2, \ldots, x_n\}$ in a space with dimension $p \geq 1$. An attribute function f is defined as a mapping from X to R (the set of all real numbers) such that $f(x_i)$ represents the attribute value of spatial point x_i. For a given point x_i, let $NN_k(x_i)$ denote the k nearest neighbors of point x_i. A neighborhood function g is defined as a map from X to R such that for each x_i, $g(x_i)$ returns a summary statistic of attribute values of all the spatial points inside $NN_k(x_i)$. For example, $g(x_i)$ can be the average attribute value of the k nearest neighbors of x_i. To detect spatial outliers, the attribute value of each point x_i is compared with those attribute values of its neighbors in $NN_k(x_i)$. Such comparison is done through a comparison function h, which is a function of f and g. There are many choices for the form of h. For example, h can be the difference $(f - g)$ or the ratio (f/g). The selection of h function depends on the properties of the practical applications. Let

$h_i = h(x_i)$ for $i = 1, 2, \ldots, n$. Given the attribute function f, neighborhood function g, and comparison function h, a point x_i is a spatial outlier or simply S-outlier if h_i is an extreme value of the set $\{h_1, h_2, \ldots, h_n\}$. We note that the definition depends on the choices of functions k, g, and h.

The definition given above is quite general. As a matter of fact, outliers involved in various existing spatial outlier detection techniques are special cases of S-outliers [34]. These include outliers detected by scatterplot [13], Moran scatterplot [23], and pocket plots [15], [27].

A straightforward method to detect S-outliers can be stated as follows. Assume all $k(x_i)$ are equal to a fixed number, denoted as k. The neighborhood function g evaluated at a spatial point x is taken to be the average attribute value of all the k nearest neighbors of x. The comparison function $h(x)$ is chosen to be the difference $f(x) - g(x)$. Applying such an h to the n spatial points leads to the sequence $\{h_1, h_2, \ldots, h_n\}$. A spatial point x_i is treated as a candidate of S-outlier if its corresponding value h_i is extreme among the data set $\{h_1, h_2, \ldots, h_n\}$. To identify the extreme values of this data set, we begin with standardizing the data set $\{h_1, h_2, \ldots, h_n\}$. Let μ and σ denote the mean and standard deviation of $\{h_1, h_2, \ldots, h_n\}$. The standardized value for each h_i is $z_i = \frac{h_i - \mu}{\sigma}$. Clearly, h_i is extreme in the data set $\{h_1, h_2, \ldots, h_n\}$ iff z_i is extreme in the standardized data set. Correspondingly, x_i is a potential S-outlier if $y_i = |z_i|$ is large. This algorithm is described in [35]. We call it Z algorithm, since it is based on the Z-score $\frac{h_i - \mu}{\sigma}$.

3.2 Deficiency of existing approaches

One drawback of the above Z algorithm is that regular spatial points could be falsely detected as spatial outliers due to the presence of neighboring points with very high/low attribute values. Thus the true spatial outliers could be ignored due to falsely detected spatial outliers if the expected number of spatial outliers is limited. We show these two problems using an illustrative example. In Fig. 1, each object

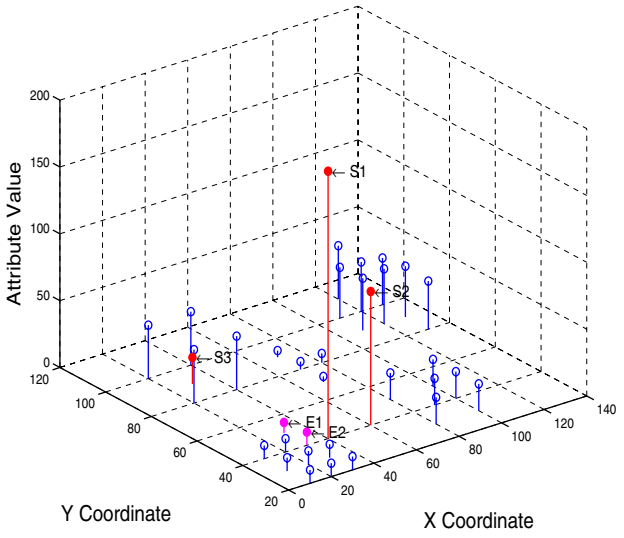


Fig. 1 A spatial data set. Objects are located in the X–Y plane. The height of each vertical line segment represents the attribute value of the corresponding object

Table 1 The X, Y, and Z coordinates of $S1$, $S2$, $S3$, $E1$ and $E2$ in Fig. 1

Point	X coordinate	Y coordinate	Z coordinate
$S1$	40	40	200
$S2$	60	40	100
$S3$	30	90	20
$E1$	30	50	10
$E2$	30	40	10

is located in the X–Y plane with its associate attribute value in the Z-coordinate. In particular, the X, Y, and Z coordinate values of points $S1$, $S2$, $S3$, $E1$, and $E2$ are listed in Table 1. Assuming the expected number of spatial outliers is 3 and k is chosen to be 3, then we can easily observe that points $S1$, $S2$ and $S3$ are spatial outliers, since their attribute values are significantly different from those of their neighbors. However, the obtained result from running the above Z algorithm indicates that $S1$, $E1$ and $E2$ are spatial outliers, as shown in Fig. 2. This detection error is mainly due to the large attribute value difference between point $S1$ and its neighboring points. For example, since $S1$ is inside the neighborhood of $E1$, the neighborhood function at $E1$ obtains a value much larger than the attribute value of $E1$, so that $E1$ is erroneously marked as an outlier. In general, the Z algorithm will lead to some true spatial outliers being ignored and some false spatial outliers being wrongly identified. This disadvantage is also shared by other existing detection approaches. For example, $S1$, $E1$, and $E2$ are detected as the top three spatial outliers by the Moran scatterplot method, since these three points are located in the upper-left and lower-right quadrants and are far away from the origin (0,0), as shown in Fig. 3. $E1$, $E2$, and $S2$ are identified as the top three spatial outliers by the scatterplot approach, since the distances of the three points to the regression line are larger than the distances of other points to the regression line, as shown in Fig. 4. To remedy the above mentioned defect of the Z algorithm, an appropriate neighborhood function

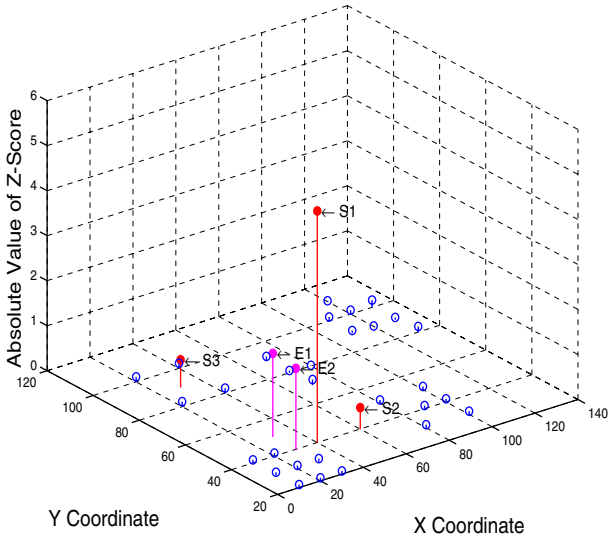


Fig. 2 Graphical illustration of the Z algorithm. Data are shown in Fig. 1 and Table 1. The height of each vertical line segment indicates the absolute value of the Z-score. $S1$, $E1$, and $E2$ would be claimed to be the potential spatial outliers since their absolute values of the Z-scores are the largest three

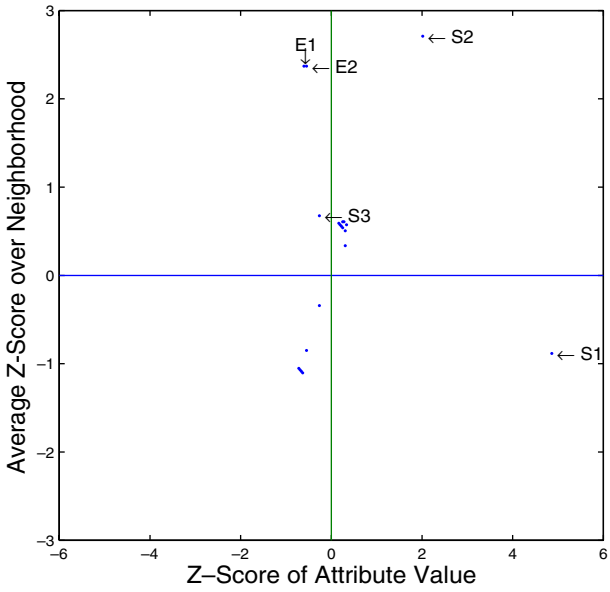


Fig. 3 Moran scatterplot used to detect the spatial outliers of the data in Fig. 1. Here Z score represents the normalized attribute value. Three best candidates of outliers are $S1$, $E1$, and $E2$, located in the upper left and lower right quadrants

needs to be selected. In the algorithm presented below, the median is used as the neighborhood function. The use of median reduces the impact caused by the extreme neighboring points.

Another drawback of the above Z algorithm is that the data set $\{h_1, h_2, \ldots, h_n\}$ is standardized using the sample mean μ and sample standard deviation σ. When multiple outliers exist in the data, these quantities are usually the biased estimates

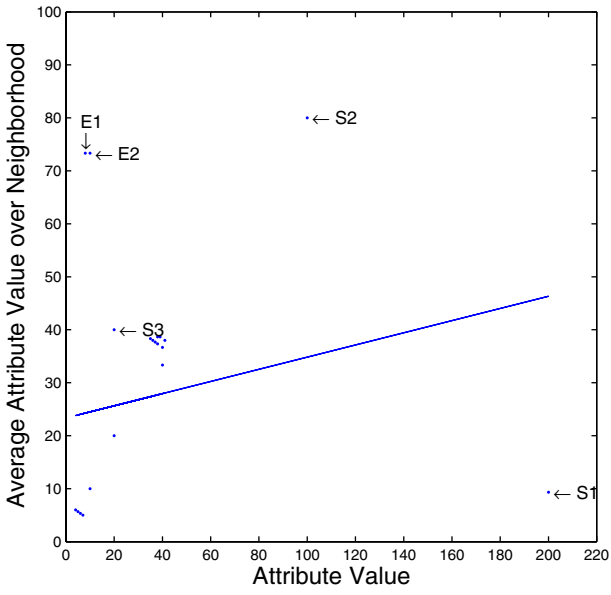


Fig. 4 Scatterplot used to analyze the data in Fig. 1. $E1$, $E2$, and $S2$ are the three best candidates of spatial outliers since their distances to the regression line are the largest

of the true population mean and standard deviation. As a result, some true spatial outliers can escape the detection and regular objects can erroneously become outliers. This is known as the problem of masking and swamping [14]. To resolve this, robust estimates of the mean and standard deviation need to be used.

3.3 Detection algorithm

In this section, we present our detection algorithm, which can be viewed as an improved version of the Z algorithm. Assume that all $k(x_i)$ are equal to a fixed number k. [The algorithm can be easily generalized by replacing the fixed k by a dynamic $k(x_i)$.] Under the framework of Section 3.1, outlier detection algorithms depend on the choices of the neighborhood function g and comparison function h. Selection of g and h determines the performance of the algorithm. In Algorithm 1 below, $g(x)$ is taken to be the median of the attribute values of the points in $NN_k(x_i)$ and $h(x)$ is chosen to be the difference between $f(x)$ and $g(x)$. Applying such an h to the n spatial points leads to the sequence $\{h_1, h_2, \ldots, h_n\}$. A spatial point x_i is treated as a candidate of S-outlier if its corresponding value h_i is extreme among the data set $\{h_1, h_2, \ldots, h_n\}$. To quantify this extremeness, the following observation can be used. If $\{h_1, h_2, \ldots, h_n\}$ is distributed as $N(\mu, \sigma)$, the data objects whose h value is far away from the mean μ can be considered as outliers [11]. Our detection algorithm is stated as follows.

Algorithm 1 (Median algorithm) Given a spatial data set $X = \{x_1, x_2, \ldots, x_n\}$, an attribute function f, one positive integer number k, and $\alpha \in (0, 1)$,

1. Compute, for each spatial point x_i, the k nearest neighbor set $NN_k(x_i)$, the neighborhood function $g(x_i) =$ median of the data set $\{f(x) : x \in NN_k(x_i)\}$, and the comparison function $h_i = h(x_i) = f(x_i) - g(x_i)$.
2. Let μ^* and σ^* denote the robust mean and standard deviation estimates of the data set $\{h_1, h_2, \ldots, h_n\}$. Standardize the data set and compute the absolute values $y_i = |\frac{h_i - \mu^*}{\sigma^*}|$ for $i = 1, 2, \ldots, n$.
3. x_i is a candidate of S-outlier if $y_i \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

In Algorithm 1, μ^* can be the median and σ^* can be the well-known median absolute deviation (MAD), defined as

$$\text{median}\{|h_1 - \text{median}(H)|, |h_2 - \text{median}(H)|, \cdots, |h_n - \text{median}(H)|\}$$

with median(H) denoting the median of the set $\{h_1, h_2, \ldots, h_n\}$. In practice, α can be chosen to be 0.001, 0.01, 0.05, or 0.10.

A quick illustration of Algorithms 1 is to apply it to the data in Fig. 1. Table 2 shows the results using the algorithm with the top three outliers, compared with the existing approaches. As can be seen, Algorithm 1 accurately detects $S1$, $S2$, and $S3$ as spatial outliers, but the Z algorithm, scatterplot, and Moran scatterplot, incorrectly identify $E1$ and $E2$ as spatial outliers. In this table, the rank of the outliers is defined in a straightforward manner. For both z and median algorithms, the rank is determined by the size of y value, i.e., objects with larger y values receive higher ranks. The rank of outliers from scatterplot is based on the magnitude of the distance of the spatial point to the least square regression line. For Moran scatterplot, the rank

Table 2 Top three spatial outliers detected by Z, median, scatterplot, and Moran scatterplot algorithms

Rank	Methods			
	Scatterplot	Moran scatterplot	Z algorithm	Median algorithm
1	E1	S1	S1	S1
2	E2	E1	E1	S2
3	S2	E2	E2	S3

is determined by the magnitude of the difference between the normalized attribute value and the corresponding neighborhood average.

3.4 Computational complexity

In the first step of the median algorithm, one needs to compute, for each object, the k nearest neighbors (KNN query) and the median attribute value. There are two choices to conduct a KNN query. We can use a grid-based approach, which processes a KNN query in constant time if the grid directory resides in memory, leading to the complexity of $O(1)$. If an index structure (e.g., R-tree) exists for the spatial data set, the spatial index can be used to process a KNN query, leading to a cost of $O(\log n)$. It takes $O(k)$ to compute the median attribute of k neighbors [24]. So the complexity of the first step is $O(n(1+k))$ for the grid-based structure or $O(n(\log n + k))$ for the spatial index structure. For the second step, it takes $O(n)$ to compute the median, $O(n)$ to compute MAD, and $O(n)$ to standardize. Thus the total cost is $O(n)$. In the third step, a cost of $O(n)$ is required. The total complexity for the algorithm is then $O(n(1+k)) + O(n) + O(n) = O(n)$ for the grid-based structure (if $n \gg k$), or $O(n(\log n + k)) + O(n) + O(n) = O(n \log n)$ for the spatial index-based structure (if $n \gg k$). It is seen that the computation cost of this algorithm is primarily determined by the KNN query.

4 Detection of spatial outliers with multiple attributes

In many applications, there may be more than one non-spatial attribute associated with each spatial location. For example, in the Census data, each census track contains several nonspatial attributes, including population, population density, income, poverty, housing, education, and race. Detecting outliers from such spatial data with multiple attributes will help demographists and social workers to identify local anomalies for further analysis. In this section, we define the multi-attribute spatial outlier detection problem, propose our detection procedure, and discuss its corresponding computational complexity.

4.1 Problem formulation

Suppose $q(\geq 1)$ measurements (attribute values) are made on the spatial object x. We use a to denote the vector of these q values at x. Given a set of spatial points

$X = \{x_1, x_2, \ldots, x_n\}$ in a space with dimension $p \geq 1$, an attribute function f is defined as a map from X to R^q (the q dimensional Euclidean space) such that for each spatial point x, $f(x)$ equals the attribute vector a.

Let $NN_k(x_i)$ denote the k nearest neighbors of point x_i with $k=k(x_i)$ for $i=1, 2, \ldots, n$. A neighborhood function g is defined as a map from X to R^q such that the jth component of $g(x)$, denoted $g_j(x)$, returns a summary statistic of the jth attribute values from all the spatial points inside $NN_k(x)$. For the purpose of detecting spatial outliers, all of the components of a at x should be compared with the corresponding quantities from the neighbors of x. A comparison function h is a function of f and g, whose domain is X and range is in R^r with $r \leq q$. Examples of h include $h = f - g$, a map from X to R^q with $r = q$, and $h = f_1/g_1$, a map from X to R with $r = 1$. Denote $h(x_i)$ by h_i. A point x_i is an S-outlier if h_i is an extreme point of the set $\{h_1, h_2, \ldots, h_n\}$. The task of designing algorithms for detecting spatial outliers with multiple attributes is formulated as follows. Given a set of spatial points $X = \{x_1, x_2, \ldots, x_n\}$, a sequence of neighborhoods $NN_k(x_1), NN_k(x_2), \ldots, NN_k(x_n)$, an attribute function $f : X \rightarrow R^q$, a neighborhood function $g : X \rightarrow R^q$, and a comparison function $h : X \rightarrow R^r$, design algorithms to detect spatial outliers with multiple attributes.

4.2 Detection algorithm

Different choices of g and h may lead to different algorithms and thus potentially different outliers. The criterion on the selection of g and h is that most of the resulting outliers should possess practical meanings. For example, examining outliers should often lead to causation investigations. When multiple attributes are present, spatial outliers should be detected by using all the attribute values simultaneously.

In the algorithm described below, we choose g to be a vector of size q with each component denoting a median. We then compute the difference between f and g, e.g., $h = f - g$ and then check the Mahalanobis distance from each point $h(x)$ to the center of the data set $\{h_1, h_2, \ldots, h_n\}$. The points that have distances larger than a predetermined threshold will be returned as outliers. The Mahalanobis distance provides a suitable way to identify points which are far from all of the others in a multidimensional space. It has been widely used in discriminant analysis, clustering, and principle analysis [25], [36], [39]. It has many advantages over Euclidian distance when dealing with multivariate data. For example, the Euclidian distance treats each variable as equally important in calculating the distance, while Mahalanobis distance automatically accounts for the scaling of the coordinate axes.

As in the case of single attribute, the problem of masking and swamping may exist for data with multiple outliers. This problem can be significantly alleviated by using robust estimates of location and shape involved in the Mahalanobis distance. In this paper, we will employ the MCD estimates of location and shape [14].

For the sample $h(x_1), \cdots, h(x_n)$ associated with n spatially referenced objects $x_1, \cdots, x_n$, the MCD is defined to be the mean and covariance matrix based on the sample of size s ($s \leqslant n$) that minimizes the determinant of the covariance matrix. That is,

$$MCD = \left(\mu_J^*, \Sigma_J^*\right)$$

where

$$J = \left\{\text{set of s points} : |\Sigma_J^*| \leqslant |\Sigma_M^*|, \ \ \forall \text{ set } M \text{ s.t. } |M| = s\right\}$$

$$\mu_J^* = \frac{1}{s}\sum_{i\in J} h(x_i)$$

$$\Sigma_J^* = \frac{1}{s}\sum_{i\in J}\left[h(x_i) - \mu_J^*\right]\left[h(x_i) - \mu_J^*\right]^T.$$

A large Mahalanobis distance indicates a possible outlier. To check whether or not such a distance is large enough, we need a cut-off point, which is based on the following observation. Suppose $h(x)$ is distributed as $N_q(\mu, \Sigma)$, i.e., q-dimensional vector $h(x)$ follows a multivariate normal distribution with mean vector μ and variance-covariance matrix Σ. Suppose the true parameters μ and Σ are approximated by the MCD μ^* and Σ^*, respectively. Let $d^2(x)$ denote $(h(x) - \mu^*)^T \Sigma^{*^{-1}} (h(x) - \mu^*)$, then $\frac{c(m-q+1)}{qm} d^2(x)$ is approximately distributed as $F_{q,m-q+1}$, where $F_{q,m-q+1}$ is the F distribution with q and $(m - q + 1)$ degrees of freedom, and the parameters m and c can be calculated from the asymptotic formulas or simulations. Thus, the probability that $h(x)$ satisfies $\frac{c(m-q+1)}{qm} d^2(x) > F_{q,m-q+1}(\alpha)$ is about α, where $F_{q,m-q+1}(\alpha)$ is the upper (100α)th percentile of a F distribution with q and $m - q + 1$ degrees of freedom. Then intuitively, if a point x satisfies the condition $\frac{c(m-q+1)}{qm} d^2(x) > F_{q,m-q+1}(\alpha)$, then x should be treated as an S-outlier candidate.

We now present the following algorithm, an improved version of the corresponding spatial outlier detection algorithm reported in [20].

Algorithm 2 Given a spatial data set $X = \{x_1, x_2, \ldots, x_n\}$, an attribute function f, one positive integer number k, and $\alpha \in (0, 1)$,

1. For each spatial point x_i, compute the k nearest neighbor set $NN_k(x_i)$.
2. For each spatial point x_i, compute the neighborhood function g such that $g_j(x_i) =$ median of the data set $\{f_j(x) : x \in NN_k(x_i)\}$, and the comparison function $h(x_i) = f(x_i) - g(x_i)$.
3. Compute the MCD-based vector μ^* and matrix Σ^* of the data set $h(x_1)$, $h(x_2), \ldots, h(x_n)$.
4. Compute $d^2(x_i) = (h(x_i) - \mu^*)^T \Sigma^{*^{-1}} (h(x_i) - \mu^*)$.
5. If $d^2(x_i) > \frac{qm}{c(m-q+1)} F_{q,m-q+1}(\alpha)$, x_i is treated as an S-outlier candidate.

4.3 Computational complexity

The complexity of Algorithm 2 is analyzed as follows. Step 1 is to compute the neighborhood for each spatial point, in which a k nearest neighbor (KNN) query is issued. The corresponding complexity is $O(n)$ for the grid-based approach or $O(n \log n)$ for the spatial indexed-based approach. In Step 2, the computation of neighborhood function g and comparison function h takes $O(qkn)$. For Step 3, the time complexity of computing the MCD estimates depends on the specific algorithms. The FAST MCD algorithm used here is a heuristic search algorithm with the complexity of $O(n)$ [31]. In Step 4, it is required to compute the Mahalanobis distance for each spatial point. Each distance computation costs $O(q^3 + q^2)$, where $O(q^3)$ refers to the complexity of matrix inversion by using the Gaussian elimination

algorithm [44], and $O(q^2)$ denotes the complexity of matrix manipulation. So the total complexity of Step 4 is $O(q^3n)$. For step 5, a cost of $O(n)$ is required. In summary, the total computational cost for Algorithm 2 is $O(n) + O(qkn) + O(n) + O(q^3n) + O(n)$ for the grid-based structure, or $O(n \log n) + O(qkn) + O(n) + O(q^3n) + O(n)$ for the index-based structure. If $n \gg k$ and $n \gg q$, the total time complexity is $O(n)$ for the grid-based structure, or $O(n \log n)$ for the index-based structure.

5 Experiment

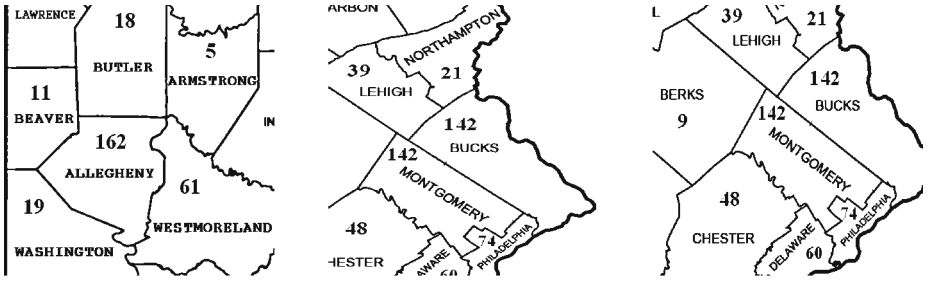
In this section, we present experimental results on the West Nile virus (WNV) data provided by the US Centers for Disease Control and Prevention (CDC), to illustrate the effectiveness of our single and multiple attribute outlier detection algorithms.

Since its first appearance in 1937, WNV has been found in Africa, West Asia, and the Middle East. The virus can infect birds, mosquitoes, horses, humans, and some other mammals. The first outbreak of WNV in USA took place in New York in 1999, and now the WNV cases have been reported in 45 states and the District of Columbia. WNV is mainly maintained in birds. In the USA, American crows and American robins are more likely to be infected by WNV. Thus, many local health departments identify the emergence of WNV by investigating dead birds. WNV can be transmitted to mosquitoes when they feed on infected birds. It may exist in the mosquito's salivary glands for several days before being injected into animals or humans through blood-feeding.

Our WNV data set is based on the 3,109 counties that comprise the contiguous United States. It reports the case numbers of wild birds, mosquitoes, and veterinaries found within each county between January 1, 2001 and December 31, 2003. The location of each county is determined by the "central" longitude and latitude provided by the data. A case number in each year can be treated as an attribute. Since the data contain three types of WNV cases in three consecutive years, there are nine attributes available for each county: Bird-2001, Mosquito-2001, Vet-2001, Bird-2002, Mosquito-2002, Vet-2002, Bird-2003, Mosquito-2003, and Vet-2003. Here Bird-2001 denotes the number of cases of WNV infected birds in the year 2001, Mosquito-2001 denotes the number of cases of WNV infected mosquitoes in the year 2001, and Vet-2001 denotes the number of cases of WNV infected veterinaries identified in the

Table 3 Top seven spatial outlier candidates detected by Z and median algorithms

Rank	Methods	
	Z alg.	Median alg.
1	Allegheny	Allegheny
2	Bucks	Bucks
3	Montgomery	Montgomery
4	Berks	Westmoreland
5	Lancaster	Berks
6	Armstrong	Dauphin
7	Luzerne	Luzerne



a Allegheny Co. **b** Bucks Co. **c** Montgomery Co.

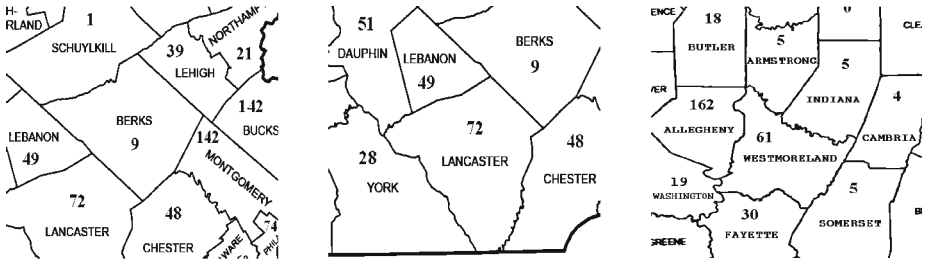
Fig. 5 Distribution of the values of attribute Bird-2002 in the vicinity of Allegheny, Bucks, and Montgomery. The numerical numbers are the values of Bird-2002

year 2001. Other attributes are defined similarly. The purpose of the experiments is to identify which counties are abnormal in terms of the WNV cases. In our analysis, each county was treated as a spatial object, and the number of neighbors for each county was chosen to be dynamic, i.e., the neighborhood of a county was chosen to be the set of adjacent counties.

5.1 Single attribute outlier detection

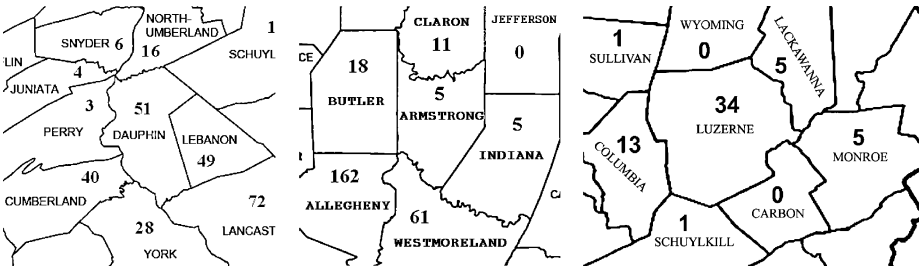
As an illustration, in this section we report the results obtained by examining the State of Pennsylvania, which has 67 counties. In the experiment, we used Bird-2002 as the attribute for each county. The median algorithm (Algorithm 1) was executed to discover which counties had abnormal WNV infected bird cases for the year of 2002. We set $\alpha = 0.05$ and totally seven counties were returned as outlier candidates. To compare our results with those from the Z algorithm, we also ran the Z algorithm.

Table 3 shows the seven spatial outlier candidates detected by both algorithms. As can be seen, the top three candidates, Allegheny County, Bucks County, and Montgomery County, are the same for both algorithms. They should be treated as outlier candidates since they have much higher attribute values (162, 142, 142) than their neighbors (far less than 100). See Fig. 5 for more details.



a Berks Co. **b** Lancaster Co. **c** Westmoreland Co.

Fig. 6 Distribution of the values of attribute Bird-2002 in the vicinity of Berks, Lancaster, and Westmoreland. The numerical numbers are the values of Bird-2002



a Dauphin Co. **b** Armstrong Co. **c** Luzerne Co.

Fig. 7 Distribution of the values of attribute Bird-2002 in the vicinity of Dauphin, Armstrong, and Luzerne. The numerical numbers are the values of Bird-2002

The fourth and fifth spatial outlier candidates are the following three counties: Berks, Westmoreland, and Lancaster. As shown in Fig. 6a, Berks should be treated as an outlier candidate, since its attribute value is quite different from those of its neighbors whose average is 58.6. Lancaster (Fig. 6b) should also be treated as an

Table 4 Top 50 spatial outlier candidates and their associated attribute values. The rank is based on the magnitude of the Mahalanobis distance d^2

Rank	County	Robust distance	Bird-2002	Vet-2002	Bird-2003	Vet-2003
1	Harris,TX	47,330.99	208.00	45.00	248.00	3.00
2	Fulton, GA	22,228.23	248.00	0.00	21.00	2.00
3	Suffolk, NY	21,537.95	180.00	4.00	173.00	3.00
4	Lancaster, PA	21,195.24	72.00	42.00	17.00	168.00
5	Tulsa, OK	20,887.80	154.00	37.00	158.00	7.00
6	Albany, NY	14,695.37	137.00	4.00	165.00	3.00
7	El Paso, CO	12,895.51	18.00	6.00	155.00	44.00
8	Hennepin, MN	11,512.54	86.00	35.00	145.00	3.00
9	New Castle, DE	10,420.65	180.00	4.00	29.00	0.00
10	Hartford, CT	9,833.03	181.00	0.00	105.00	4.00
11	Milwaukee, WI	8,450.96	157.00	0.00	2.00	0.00
12	Allegheny, PA	8,318.51	162.00	0.00	5.00	1.00
13	Middlesex, MA	8,031.91	203.00	0.00	95.00	1.00
14	Chester, PA	7,105.70	48.00	4.00	27.00	98.00
15	Davidson, TN	6,979.80	138.00	2.00	3.00	1.00
16	San Juan, NM	6,456.51	0.00	0.00	0.00	90.00
17	Larimer, CO	6,253.59	6.00	31.00	125.00	32.00
18	Bay, FL	6,077.66	7.00	2.00	108.00	10.00
19	Escambia, FL	5,858.14	124.00	39.00	76.00	3.00
20	Cobb, GA	5,829.41	119.00	1.00	57.00	0.00
21	Bucks, PA	5,282.78	142.00	5.00	6.00	34.00
22	Gwinnett, GA	5,161.84	50.00	0.00	93.00	0.00
23	District of Columbia, DC	5,160.79	175.00	0.00	2.00	0.00
24	Rockland, NY	5,067.30	138.00	0.00	36.00	1.00
25	DeKalb, GA	4,903.97	124.00	0.00	46.00	0.00

Table 4 (continued)

Rank	County	Robust distance	Bird-2002	Vet-2002	Bird-2003	Vet-2003
26	Monmouth, NJ	4,693.65	157.00	4.00	46.00	18.00
27	Putnam, NY	4,123.20	13.00	0.00	109.00	0.00
28	Rockingham, NH	3,726.77	17.00	0.00	101.00	0.00
29	Middlesex, CT	3,536.28	31.00	0.00	55.00	0.00
30	Fairfield, CT	3,339.80	121.00	1.00	94.00	0.00
31	Cook, IL	3,255.73	101.00	20.00	34.00	0.00
32	Montgomery, PA	3,096.85	142.00	2.00	30.00	23.00
33	Oklahoma, OK	2,717.62	51.00	45.00	59.00	7.00
34	Henrico, VA	2,685.48	14.00	0.00	70.00	4.00
35	Lancaster, NE	2,564.99	85.00	20.00	32.00	2.00
36	Los Angeles, CA	2,455.29	0.00	0.00	65.00	0.00
37	Fremont, WY	2,416.72	1.00	4.00	10.00	52.00
38	Hamilton, TN	2,215.35	73.00	1.00	30.00	2.00
39	Shelby, TN	2,215.33	83.00	33.00	24.00	1.00
40	Mobile, AL	2,146.92	65.00	12.00	47.00	11.00
41	Goshen, WY	2,122.99	11.00	40.00	64.00	4.00
42	Teller, CO	2,121.21	0.00	0.00	1.00	0.00
43	Ramsey, MN	2,023.93	40.00	0.00	88.00	6.00
44	Holmes, OH	1,900.27	0.00	134.00	1.00	2.00
45	Morris, NJ	1,827.87	79.00	0.00	27.00	1.00
46	Weld, CO	1,806.45	18.00	99.00	39.00	52.00
47	Jefferson, KY	1,778.40	70.00	10.00	6.00	3.00
48	Jefferson, AL	1,773.78	49.00	1.00	51.00	0.00
49	Okaloosa, FL	1,731.70	4.00	2.00	58.00	4.00
50	Marion, FL	1,705.03	40.00	121.00	2.00	20.00

outlier candidate, since it has an attribute value of 72, and all the values from its neighbors are less than 50. Note that Lancaster has other two neighboring counties that do not show up in this figure. They are Harvard County and Cecil County, with attribute values of 45 and 5, respectively. Westmoreland County is identified by the median algorithm, while it does not appear in the result of the Z algorithm. Westmoreland county should be treated as an outlier candidate, since the attribute value of this county is much higher than most of its neighbors (see Fig. 6c). The Z algorithm fails to detect it since its neighbor, Allegheny County, has such a high attribute value (162) that the averaged value of the neighborhood of Westmoreland is close to the attribute value of Westmoreland. The median algorithm does not have this disadvantage since it uses the median value (11.5) to represent the "center" of the values from Westmoreland's neighbors (4, 5, 5, 5, 18, 19, 30, 162).

For the sixth outlier candidate, the median algorithm selected the county of Dauplin. As shown in Fig.7a, Dauplin is selected since it has four neighbors which have very small attribute values (3, 4, 6, 16). The z approach detected Armstrong County. However, this county may not be an outlier candidate. Figure 7b shows that Armstrong has six neighbors, four of which have values similar to that of Armstrong. But the other two neighbors, Allegheny and Westmoreland, have much higher values, which make the Armstrong County falsely detected as an outlier candidate.

The seventh and final outlier candidate, Luzerne County, is the same for both algorithms. As shown in Fig. 7c, it should be treated as an outlier candidate since it has a much higher attribute value (34) than its neighbors (0, 0, 1, 1, 5, 5, 13).

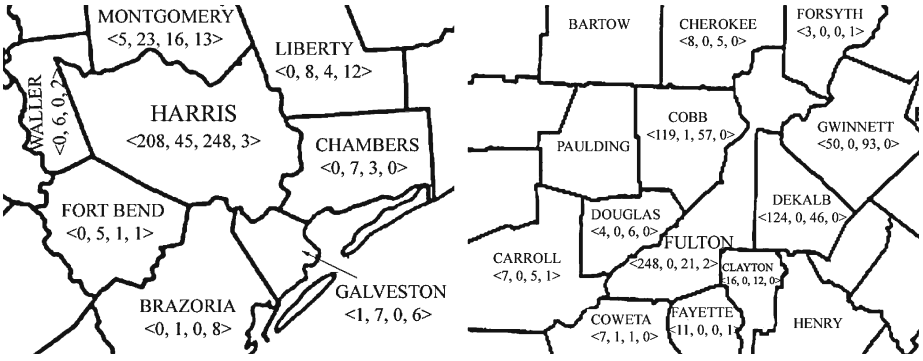
The above example shows that the median algorithm can identify spatial outlier candidates ignored by the Z algorithm and avoid detecting erroneous spatial outliers. This conclusion is valid for all the experiments we have conducted.

5.2 Multiple attribute outlier detection

In conducting multiple attribute outlier detection, we considered all the 3,109 counties from the data. We conducted the experiment using Matlab 6.5. The robust mean and covariance matrix estimates were generated by using the implementation of the FAST MCD algorithm in a third party Matlab toolbox LIBRA [45]. Because the use of all the nine attributes incurred a singular MCD covariance matrix, we present here, as an illustration, the results based on the following four attributes Bird-2002, Vet-2002, Bird-2003, and Vet-2003. The use of these four attributes will lead to a nonsingular Σ^*.

As discussed previously, $(\frac{c(m-q+1)}{qm})d^2(x)$ is distributed approximately as $F_{q,m-q+1}$, where $F_{q,m-q+1}$ is the F distribution with q and $(m-q+1)$ degrees of freedom. Since four attributes were used, $q=4$. The data have more than 3,000 objects, and therefore we calculated the parameters m and c by using the asymptotic formulas [31]: $m=353.422$ and $c=0.472$. For our experiment, we chose $\alpha=0.001$. After running the Algorithm 2, 563 counties were returned as outlier candidates, as partially shown in Table 4.

Below we provide a brief discussion to show why Harris and Fulton (the first two in Table IV) were selected as outlier candidates. As shown in Fig. 8a, the values of 2002-Bird , 2002-Vet, and 2003-Bird for Harris are 208, 45, and 248, respectively. The medians of values of 2002-Bird, 2002-Vet, and 2003-Bird for the neighbors of Harris are 0, 7, and 1 respectively. Harris was selected as an outlier candidate since the difference between the value of each of these three attributes at Harris and the



a Dauphin Co. **b** Armstrong Co.

Fig. 8 Distribution of the values of attributes Bird-2002, Vet-2002, Bird-2003, and Vet-2003 in the vicinity of Harris and Fulton. The four values in each vector are the values of these four attributes, respectively

associated median from the neighbors is large. Similarly, as displayed in Fig. 8b, Fulton was selected as an outlier candidate since for the attributes 2002-Bird and 2003-Bird, the difference between an attribute value at Fulton and the corresponding median from the neighbors is large.

A further examination coupled with specific knowledge of epidemiology of the WNV can be conducted to indicate whether a detected county is a spatial outlier candidate. However, this is beyond the scope of this paper and thus is omitted.

6 Conclusion

In this paper we propose two algorithms to detect spatial outliers. One median based algorithm is developed for single attribute outlier detection, and one Mahalanobis-distance-based algorithm is proposed for multi-attribute outlier detection. Robust estimates are used to approximate the corresponding parameters involved in the algorithms. This will significantly alleviate the well-known masking and swamping effects that may exist in data with multiple outliers or groups of outliers. Illustrative experiments have been performed on the West Nile virus data.

Future research activities can be conducted along several directions. First, in developing our detection algorithms, we assumed the sequence of differences $h_1, h_2, \cdots, h_n$ are independent of each other. In practice, these differences are usually dependent. In our future work, we will explore the effect of such dependence on the performance of the proposed algorithms and consequently will study modified detection algorithms. Our methods in this paper focus on processing "static" data. The proposed algorithms can be extended to discover spatial outliers in continuous geospatial data streams.

References

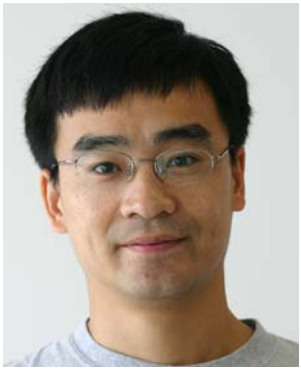
1. C.C. Aggarwal. "Redesigning distance functions and distance-based applications for high dimensional data," *SIGMOD Record*, Vol. 30(1):13–18, March 2001.
2. C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, and J. S. Park. "Fast algorithms for projected clustering," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 61–72, Philadelphia, Pennsylvania, United States, June 1–3, 1999.
3. C.C. Aggarwal and P.S. Yu. "Outlier detection for high dimensional data," in *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 37–46, Santa Barbara, California, United States, May 21–24, 2001.
4. V. Barnett and T. Lewis. Outliers in Statistical Data. Wiley, New York, 1994.
5. S. Berchtold, C. Böhm, and H.-P. Kriegal. "The pyramid-technique: Towards breaking the curse of dimensionality," in *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 142–153, Seattle, Washington, United States, June 2–4, 1998.
6. M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. "Lof: Identifying density-based local outliers." in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, Dallas, Texas, United States, May 14–19, 2000.
7. A. Cerioli and M. Riani. "The ordering of spatial data and the detection of multiple outliers," *Journal of Computational and Graphical Statistics*, Vol. 8(2):239–258, June 1999.
8. P.K. Chan, W. Fan, A.L. Prodromidis, and S.J. Stolfo. "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems*, Vol. 14(6):67–74, 1999.
9. W.S. Chan and W.N. Liu. "Diagnosing shocks in stock markets of Southeast Asia, Australia, and New Zealand," *Mathematics and Computers in Simulation*, Vol. 59(1–3):223–232, 2002.

10. A. Conci and C.B. Proença. "A system for real-time fabric inspection and industrial decision," in *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, pp. 707–714, Ischia, Italy, July 15–19, 2002.
11. D. Freedman, R. Pisani, and R. Purves. Statistics. Norton, Vol. 41:212–223, 1998.
12. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise," in *the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Portland, Oregon, United States, August 2–4, 1996.
13. R. Haining. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, 1993.
14. J. Hardin and D.M. Rocke. "The distribution of robust distances," *Journal of Computational and Graphical Statistics*, Vol. 14:1–19, 2005.
15. J. Haslett, R. Brandley, P. Craig, A. Unwin, and G. Wills. "Dynamic graphics for exploring spatial data with application to locating global and local anomalies," *The American Statistician*, Vol. 45:234–242, 1991.
16. A. Hinneburg, C.C. Aggarwal, and D.A. Keim. "What is the nearest neighbor in high dimensional spaces?" in *Proceedings of 26th International Conference on Very Large Data Bases*, pp. 506–515, Cairo, Egypt, September 10–14, 2000.
17. W. Jin, A.K.H. Tung, and J. Han. "Mining top-n local outliers in large databases," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 293–298, San Francisco, California, United States, August 26–29, 2001.
18. E.M. Knorr and R.T. Ng. "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24th International Conference on Very Large Data Bases*, pp. 392–403, New York City, NY, United States, August 24–27, 1998.
19. H. Liu, K.C. Jezek, and M.E. O'Kelly. "Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and gis," *International Journal of Geographical Information Science*, Vol. 15(8):721–741, 2001.
20. C.-T. Lu, D. Chen, and Y. Kou. "Detecting spatial outliers with multiple attributes," in *Proceedings of the 15th International Conference on Tools with Artificial Intelligence*, pp. 122–128, Sacramento, California, United States, November 3–5, 2003.
21. C.-T. Lu, D. Chen, and Y. Kou. "Algorithms for spatial outlier detection," in *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, Florida, pp. 597–600, November 19–22, 2003.
22. C.-T. Lu and L.R. Liang. "Wavelet fuzzy classification for detecting and tracking region outliers in meteorological data," in *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, pp. 258–265, Washington DC, United States, November 12–13, 2004.
23. A. Luc. "Local indicators of spatial association: Lisa." *Geographical Analysis*, Vol. 27(2):93–115, 1995.
24. M. Blum, R.W. Floyd, V. Pratt, R. Rivest, and R. Tarjan. "Time bounds for selection," *Journal of Computer and System Sciences*, Vol. 7:448–461, 1973.
25. A. Mkhadri. "Shrinkage parameter for the modified linear discriminant analysis," *Pattern Recognition Letters*, Vol. 16(3):267–275, 1995.
26. R. T. Ng and J. Han. "Efficient and effective clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 144–155, Santiago de Chile, Chile, September 12–15, 1994.
27. Y. Panatier. VARIOWIN: Software for Spatial Data Analysis in 2D. Springer, New York, 1996.
28. M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. "A brain tumor segmentation framework based on outlier detection," *Medical Image Analysis*, Vol. 9(5):457–466, 2004.
29. F.P. Preparata and M.I. Shamos. Computational Geometry—An Introduction. Springer, 1985.
30. S. Ramaswamy, R. Rastogi, and K. Shim. "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, vol. 29, pp. 427–438, Dallas, Texas, United States, May 16–18, 2000.
31. P.J. Rousseeuw and K.V. Driessen. "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, Vol. 41:212–223, 1999.
32. I. Ruts and P.J. Rousseeuw. "Computing depth contours of bivariate point clouds," *Computational Statistics and Data Analysis*, Vol. 23(1):153–168, 1996.
33. S. Shekhar and S. Chawla. A Tour of Spatial Databases. Prentice Hall, 2002.
34. S. Shekhar, C.-T. Lu, and P. Zhang. "A unified approach to detecting spatial outliers," *GeoInformatica*, Vol. 7(2):139–166, 2003.

35. S. Shekhar, C.-T. Lu, and P. Zhang. "Detecting graph-based spatial outliers: algorithms and applications (a summary of results)," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 371–376, San Francisco, California, United States, August 26–29, 2001.
36. M.E. Tipping and C.M. Bishop. "Mixtures of probabilistic principal component analysers," *Neural Computation*, Vol. 11(2):443–482, 1999.
37. W. Tobler. "Cellular geography," in *Philosophy in Geography*, pp. 379–386, Dordrecht, Holland. Dordrecht Reidel Publishing Company, 1979.
38. W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. "Rule-based anomaly pattern detection for detecting disease outbreaks," in *The Eighteenth National Conference on Artificial Intelligence*, pp. 217–223, Edmonton, Alberta, Canada, July 28–August 1, 2002.
39. L. Xu. "Bayesian ying-yang machine, clustering and number of clusters," *Pattern Recognition Letters*, Vol. 18(11–13):1167–1178, 1997.
40. K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne. "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, Vol. 8(3):275–300, 2004.
41. S. Zanero and S.M. Savaresi. "Unsupervised learning techniques for an intrusion detection system," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 412–419, Nicosia, Cyprus, March 14–17, 2004.
42. T. Zhang, R. Ramakrishnan, and M. Livny. "Birch: an efficient data clustering method for very large databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 103–114, Montreal, Quebec, Canada, June 4–6, 1996.
43. J. Zhao, C.-T. Lu, and Y. Kou. "Detecting region outliers in meteorological data," in *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, pp. 49–55, New Orleans, Louisiana, United States, November 7–8, 2003.
44. G.H. Golub and C.F. Van Loan. Matrix Computations. The Johns Hopkins University Press, 3rd ed., 1996.
45. S. Verboven and M. Hubert. "LIBRA: a Matlab library for robust analysis," *Chemometrics and Intelligent Laboratory Systems*, Vol. 75:127–136, 1996.

Dechang Chen is an associate professor at Division of Epidemiology and Biostatistics, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Maryland, USA. His research interests include applied statistics, bioinformatics, machine learning, ad hoc and sensor networks, and differential equations.

Chang-Tien Lu received the BS degree in Computer Science and Engineering from the Tatung Institute of Technology, Taipei, Taiwan, in 1991, the MS degree in Computer Science from the Georgia Institute of Technology, Atlanta, GA, in 1996, and the Ph.D. degree in Computer Science from the University of Minnesota, Minneapolis, MN, in 2001. He is currently an assistant professor in the Department of Computer Science at Virginia Polytechnic Institute and State University, and is the founding director of the Spatial Data Management Laboratory. His research interests include spatial databases, data mining, data warehousing, geographic information systems, and intelligent transportation systems. Dr. Lu is also affiliated with Virginia Tech Civil and Environmental Engineering Department, Center for Geospatial Information Technology, and Virginia Tech Transportation Institute.



Yufeng Kou received the BS degree in computer science from Northwestern Polytechnic University, XiAn, China, in 1996, the MS degree in computer science from Beijing University of Post and Telecommunications. He is a PhD candidate in computer science department, Virginia Polytechnic Institute and State University. His research interests include spatial data analysis, data mining, data warehousing, and geographic information systems.

Feng Chen received the BS degree in computer science from Hunan University, Changsha, China, in 2001, the MS degree in computer science from Beijing University of Aeronautics & Astronautics. He is a PhD candidate in computer science department, Virginia Polytechnic Institute and State University. His research interests include spatial data analysis, data mining, data warehousing, and geographic information systems.