Detecting Spatial Outliers with Multiple Attributes

Chang-Tien Lu
Dept. of Computer Science
Virginia Polytechnic Institute
and State University
7054 Haycock Road
Falls Church, VA 22043
ctlu@vt.edu

Dechang Chen
Preventive Medicine and
Biometrics
Uniformed Services University
of the Health Sciences
Bethesda, MD 20814
dchen@usuhs.mil

Yufeng Kou
Dept. of Computer Science
Virginia Polytechnic Institute
and State University
7054 Haycock Road
Falls Church, VA 22043
ykou@vt.edu

Abstract

A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood. Identification of spatial outliers can lead to the discovery of unexpected, interesting, and useful spatial patterns for further analysis. Previous work in spatial outlier detection focuses on detecting spatial outliers with a single attribute. In the paper, we propose two approaches to discover spatial outliers with multiple attributes. We formulate the multi-attribute spatial outlier detection problem in a general way, provide two effective detection algorithms, and analyze their computation complexity. In addition, using a real-world census data, we demonstrate that our approaches can effectively identify local abnormality in large spatial data sets.

1 Introduction

Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [4], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [9]. The identification of outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud detection, athlete performance analysis, voting irregularity analysis, and severe weather prediction.

Spatial data set could be modelled as a collection of spatially referenced objects, such as roads, buildings and cities. Attributes of spatial objects fall into two categories: spatial attributes and non-spatial attributes. The spatial attributes include location, shape and other geometric or topological properties. Non-spatial attributes include length, height, owner, building age and name. A spatial neighborhood [28] of a spatially referenced object is a subset of the spatial data based on the spatial dimension using spatial relationships, e.g., distance and adjacency. Comparisons between spatially referenced objects are based on non-spatial attributes.

In a spatial context, local anomalies are of paramount importance. Spatial outliers are spatially referenced objects whose non-spatial attribute values are significantly different from those of other spatially referenced objects in their spatial neighborhoods. Informally, a spatial outlier is a local instability, or an extreme observation with respect to its neighboring values, even though it may not be significantly different from the entire population. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases [21, 22, 25]. These application domains include transportation, ecology, public safety, public health, climatology, and location based services. In these applications, there may be more than one non-spatial attributes associated with each spatial location. For example, in census data set, each census track contains several non-spatial attributes, including population, population density, income, poverty, housing, education, and race [27]. Detecting outliers from these spatial data with multiple attributes will help demographist and social worker to identify local anomalies for further analysis.

This paper focuses on detecting spatial outlier with multiple attributes. We formulate spatial outlier detection problems in a general way, propose two effective algorithms, analyze their computational costs, and demonstrate the effectiveness of our proposed approaches using a real-world census data set. The paper is organized as follows. Section 2 reviews related work in outlier detection. In Section 3, we formulate the problem, propose two spatial outlier detection algorithms, and analyze their computational complexity. The experimental results and analysis are provided in Section 4. Finally, we conclude in Section 5 with directions for future work.

Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)
1082-3409/03 $17.00 © 2003 **IEEE**

2 Related Work

Numerous outlier detection tests, known as discordancy tests, have been studied in the field of statistics. These tests are developed for different circumstances, depending on the data distribution, the number of expected outliers, and the types of expected outliers [4, 12]. The main idea is to fit the data set to a known standard distribution, and develop a test based on distribution properties. In computational geometry, each data object is represented as a point in a k-dimensional space with an assigned depth. Depth-based approaches [18, 20, 26] organize data objects in convex hull layers in the data space according to peeling depth, and outliers are expected to be found from data objects with shallow depth values. In the context of KDD, many outlier detection algorithms have been recently proposed. They provide outlier tests based on different concepts, such as distance, density, and local property. Knorr and Ng presented the notion of distance-based outliers [13, 14]. For a k dimensional data set T with N objects, an object O in T is a $DB(p, D)$-outlier if at least a fraction p of the objects in T lies greater than distance D from O. Ramaswamy et al. proposed a formulation for distance-based outliers by calculating the distance of a point from its k^{th} nearest neighbor [19]. After ranking points by the distance to its k^{th} nearest neighbor, the top n points are declared as outliers. Breunig et al. introduced the notion of a "local" outlier in which the outlier-degree of an object is determined by taking into account the clustering structure in a bounded k nearest neighborhood of the object [6, 11]. The major limitation of applying the above algorithms for spatial outlier detection is that they do not distinguish between spatial and non-spatial attributes and are not suitable for detecting spatial outliers.

Recent work by Shekhar et al. introduced a method for detecting spatial outliers in graph data set [23, 24]. The method is based on the distribution property of the difference between an attribute value and the average attribute value of its neighbors. Several spatial outlier detection methods are also available in the literature of spatial statistics. 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. Example methods include variogram clouds and pocket plots [8, 17]. Quantitative methods provide tests to distinguish spatial outliers from the remainder of data. Scatterplot and Moran scatterplot are two representative approaches. A Scatterplot [7, 15] 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; a scatter sloping upward to the left indicates a negative spatial autocorrelation. Nodes far away from the regression line are flagged as possible spatial outliers. A Moran scatterplot [16] is a plot of normalized attribute value against the neighborhood average of normalized attribute values. A Moran scatterplot contains four quadrants. The upper left and lower right quadrants indicate a spatial association of dissimilar values: low values surrounded by high value neighbors and high values surrounded by low value neighbors. Spatial outliers can be identified from these two quadrants. The above methods for detecting spatial outliers focus on the case of single attribute.

For detecting outlier with multiple attributes, traditional outlier detection approaches could not be used properly due to the sparsity of the data objects in high dimensional data space [3]. It has been shown that the distance between any pair of data points in high dimensional space is so similar that either each data point or none data point can be viewed as an outlier if the concepts of proximity is used to define outliers [1]. As a result, using traditional Euclidean distance function cannot effectively get outliers in high dimensional data set due to the averaging behavior of the noisy and irrelevant dimensions. To address this problem, two categories of research work have been conducted. The first is to project the high dimensional data to low dimensional data that has abnormally low local density [2, 3, 5, 10]. The second approach is to re-design distance functions to accurately define the proximity relationship between data points [1].

All these multi-attribute outlier detection approaches deal with non-spatial attributes. For spatial outlier detection, there are two dimensions: spatial dimension and non-spatial dimension. In detecting spatial outliers, 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 used to define the distance function.

3 Algorithms

In this section, we define the multi-attribute spatial outlier detection problem and propose our algorithms. The first algorithm is based on computing the average of attribute values of neighbors, while the second algorithm is based on computing the median of attribute values of neighbors.

3.1 Problem Formulation

Suppose q measurements (attribute values) $y_1, y_2, \cdots, y_q$ ($q \geq 1$) are made on the spatial object $\mathbf{x}$. We use $\mathbf{y}$ to denote the vector $(y_1, y_2, \cdots, y_q)^T$, where T represents the transpose operation. That is, $\mathbf{y} = (y_1, y_2, \cdots, y_q)^T$. Given a set of spatial points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ 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 $\mathbf{x}$, $f(\mathbf{x})$ equals the attribute vector $\mathbf{y}$. For convenience, we write

$$\begin{aligned} \mathbf{y}_i &= f(\mathbf{x}_i) \\ &= (f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \ldots, f_q(\mathbf{x}_i))^T \\ &= (y_{i1}, y_{i2}, \cdots, y_{iq})^T \end{aligned}$$

for $i = 1, 2, \ldots, n$. Denote the set $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ by A.

For a given integer k, let $NN_k(\mathbf{x}_i)$ denote the k nearest neighbors of point $\mathbf{x}_i$ 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 $g(\mathbf{x})$, denoted $g_j(\mathbf{x})$, returns a summary statistic of attribute values y_j of all the spatial points inside $NN_k(\mathbf{x})$.

For the purpose of detecting spatial outliers, we compare all of the components of $\mathbf{y}$ at $\mathbf{x}$ with the corresponding quantities from the neighbors of $\mathbf{x}$. 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 $h = f - g$, which represents 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 $h(\mathbf{x}_i)$ by h_i.

Given the attribute function f, neighborhood function g, and comparison function h, a point $\mathbf{x}_i$ is an S-outlier (spatial outlier) if h_i is an extreme point of the set $\{h_1, h_2, \ldots, h_n\}$. We note that the definition is very general and depends on the choices of functions g and h. The following problem characterizes the task of designing algorithms for detecting spatial outliers:

Spatial Outlier Detection Problem

Given:

- A set of spatial points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- Neighborhoods $NN_k(\mathbf{x}_1), NN_k(\mathbf{x}_2), \ldots, NN_k(\mathbf{x}_n)$
- An attribute function $f : X \to R^q$
- A neighborhood function $g : X \to R^q$
- A comparison function $h : X \to R^r$

Design:

- Algorithms to detect spatial outliers

3.2 Spatial Outlier Detection Algorithms

We introduce two multi-attribute spatial outlier detection algorithms. Different choices of g and h may lead to 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.

Detecting unusual attribute vector by the difference between f and g, i.e., $h = f - g$, is available. We do this through checking the Mahalanobis distance between $h(\mathbf{x})$ and the average h value from the neighbors of $\mathbf{x}$. The Mahalanobis approach considers both the average value and its variance and covariance of the attributes measured. It accounts for ranges of variance between attributes and compensates for interactions (covariance) between attributes. To describe this method, let us first note the following: a) Under certain conditions, we may show that $h(\mathbf{x})$ follows a multivariate normal distribution. b) If $h(\mathbf{x})$ is distributed as $N_q(\boldsymbol{\mu}, \Sigma)$, i.e., q-dimensional vector $h(\mathbf{x})$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ, then $(h(\mathbf{x}) - \boldsymbol{\mu})^T \Sigma^{-1} (h(\mathbf{x}) - \boldsymbol{\mu})$ is distributed as χ_q^2, where χ_q^2 is the chi-square distribution with q degrees of freedom. Therefore the probability that $h(\mathbf{x})$ satisfies $(h(\mathbf{x}) - \boldsymbol{\mu})^T \Sigma^{-1} (h(\mathbf{x}) - \boldsymbol{\mu}) > \chi_p^2(\alpha)$ is α. Here $\chi_q^2(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with q degrees of freedom. For example, $\chi_{10}^2(0.05) = 18.31$.

Now suppose there are n spatial referenced objects $\mathbf{x}_1$, $\cdots$, $\mathbf{x}_n$. For the sample $h(\mathbf{x}_1)$, $\cdots$, $h(\mathbf{x}_n)$, calculate the sample mean

$$\boldsymbol{\mu}_s = \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{x}_i)$$

and sample variance-covariance matrix

$$\Sigma_s = \frac{1}{n-1} \sum_{i=1}^{n} [h(\mathbf{x}_i) - \boldsymbol{\mu}_s][h(\mathbf{x}_i) - \boldsymbol{\mu}_s]^T.$$

Then we should expect that the probability of $h(\mathbf{x})$ satisfying $(h(\mathbf{x}) - \boldsymbol{\mu}_s)^T \Sigma_s^{-1} (h(\mathbf{x}) - \boldsymbol{\mu}_s) > \chi_q^2(\alpha)$ is roughly α.

Set $d^2(\mathbf{x}) = (h(\mathbf{x}) - \boldsymbol{\mu}_s)^T \Sigma_s^{-1} (h(\mathbf{x}) - \boldsymbol{\mu}_s)$. For any $\mathbf{x}$, if $d^2(\mathbf{x})$ is unusually large, $\mathbf{x}$ will be teated as a spatial outlier. In other words, if $d^2(\mathbf{x}) > \theta$, $\mathbf{x}$ is a spatial outlier, where θ is a predetermined number depending on a specified confidence level. It follows from the above discussion that many algorithms for detecting S-outliers are available. Choosing g to be the average attribute vectors from the neighborhood yields the following algorithm.

Spatial Outlier Detection Algorithm 1: Mean Algorithm

1. Given the spatial data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, predefined threshold θ, attribute function f, and the number k of nearest neighbors

2. For each fixed j $(1 \leq j \leq q)$, standardize the attribute function f_j, i.e., $f_j(\mathbf{x}_i) \leftarrow \frac{f_j(\mathbf{x}_i) - \mu_{f_j}}{\sigma_{f_j}}$ for $i = 1, 2, \ldots, n$.

3. For each spatial point $\mathbf{x}_i$, compute the k nearest neighbor set $NN_k(\mathbf{x}_i)$

4. For each spatial point $\mathbf{x}_i$, compute the neighborhood function g such that $g_j(\mathbf{x}_i)$ = average of the data set $\{f_j(\mathbf{x}) : \mathbf{x} \in NN_k(\mathbf{x}_i)\}$, and the comparison function $h(\mathbf{x}_i) = f(\mathbf{x}_i) - g(\mathbf{x}_i)$.

5. Compute $d^2(\mathbf{x}_i) = (h(\mathbf{x}_i) - \boldsymbol{\mu}_s)^T \Sigma_s^{-1}(h(\mathbf{x}_i) - \boldsymbol{\mu}_s)$. If $d^2(\mathbf{x}_i) > \theta$, $\mathbf{x}_i$ is a spatial outlier w.r.t. A.

We call the above algorithm $Mean$ algorithm, since the algorithm is based on computing the average attribute value of neighborhoods.

Replacing g in the above algorithm by the "median" of the attribute vectors from the neighborhood, we have the following detection algorithm. The motivation of using median is the fact that median is a robust estimator of the "center" of a sample. Since Algorithm 2 focuses on computation based on the difference between the attribute value of each point and the median value of its k nearest neighbors, we call it Median Algorithm.

Spatial Outlier Detection Algorithm 2: Median Algorithm

1. Given the spatial data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, predefined threshold θ, attribute function f, and the number k of nearest neighbors

2. For each fixed j ($1 \leq j \leq q$), standardize the attribute function f_j, i.e., $f_j(\mathbf{x}_i) \leftarrow \frac{f_j(\mathbf{x}_i) - \mu_{f_j}}{\sigma_{f_j}}$ for $i = 1, 2, \ldots, n$.

3. For each spatial point $\mathbf{x}_i$, compute the k nearest neighbor set $NN_k(\mathbf{x}_i)$ based on its spatial location.

4. For each spatial point $\mathbf{x}_i$, compute the neighborhood function g such that $g_j(\mathbf{x}_i)$ = median of the data set $\{f_j(\mathbf{x}) : \mathbf{x} \in NN_k(\mathbf{x}_i)\}$, and the comparison function $h(\mathbf{x}_i) = f(\mathbf{x}_i) - g(\mathbf{x}_i)$.

5. Compute $d^2(\mathbf{x}_i) = (h(\mathbf{x}_i) - \boldsymbol{\mu}_s)^T \Sigma_s^{-1}(h(\mathbf{x}_i) - \boldsymbol{\mu}_s)$. If $d^2(\mathbf{x}_i) > \theta$, $\mathbf{x}_i$ is a spatial outlier w.r.t. A.

We note that in the above two algorithms, if the expected number m of S-outliers is given, instead of θ, then those m outliers could be picked up according to the m largest values of d^2.

3.3 Computational Complexity

For the Mean Algorithm, Step 2 is to standardize the attribute function, which costs $O(qn)$. In Step 3, the neighborhood is computed for each spatial point, in which a k nearest neighbor (KNN) query is issued. The time complexity is then based on that of KNN query. For the KNN query, there are two choices. We can use a grid-based approach, which processes KNN query in constant time if the grid directory resides in memory, leading to a complexity of $O(n)$. If an index structure (e.g. R-tree) exists for the spatial data set, spatial index can be used to process KNN query, which has complexity of $O(logn)$, leading to a complexity of $O(nlogn)$. For Step 4, the computation of neighborhood function g and comparison function h takes $O(qkn)$. In Step 5, the computation of Mahalanobis distance costs $O(q^2{*}n)$. In summary, the total computational cost for the Mean Algorithm is $O(qn) + O(n) + O(qkn) + O(q^2 * n)$ for grid-based structure, or $O(qn) + O(nlogn) + O(qkn) + O(q^2 * n)$ for index-based structure. If $n \gg k$ and $n \gg d$, the total time complexity is $O(n)$ for grid-based structure, or $O(nlogn)$ for index-bases structure. The time complexity is then primarily determined by the KNN query. The Median Algorithm has the same time complexity as the Mean Algorithm. The only difference between the two algorithms lies in the computation of neighborhood function g. Nevertheless, the time complexity for computing average and median for k neighbors is the same, i.e., $O(k)$.

4 Experiments

We empirically evaluated our detection algorithms by mining a real-life census data set. The experiment results indicate that our algorithms can effectively identify spatial outliers with multiple attributes.

The census data is the most detailed tabulation of American demographic data compiled by U.S. Census Bureau [27]. It contains detailed data on population, race and ethnicity, age and sex, education, employment, income, poverty, housing, and many other attributes for each of the following different levels of geography: 1) the United State and major regions of the country; 2) each state and metropolitan area; 3) all 3000+ counties in the United States; 4) municipalities, census tracts, and block groups. More than 3000 counties were processed in our experiment. The location of each county is determined by one or more polygons consisting of hundreds of longitude and latitude pairs. The neighborhoods were chosen to be dynamic, i.e., the neighborhood of a county was chosen to be the set of adjacent counties.

In the experiment, we used the following 11 attributes: population in 2001, population percent change from April 1 2000 to July 1 2001, population percent change from 1990 to 2000, percentage of persons under 5 years old in 2000, percentage of persons under 18 years old in 2000, percentage of persons over 65 years old in 2000, percentage of persons under 5 years old in 2000, percentage of persons under 18 years old in 2000, percentage of White persons, percentage of Black persons, percentage of Asian persons, and percentage of American Indian persons. The experiment was

Rank	County	Mahalanobis Distance	Pop 2001	2000-2001 change	1990-2000 change	≤ 5 %	≤ 18 %	≥ 65 %	Female %	White %	Black %	Indian %	Asian %
1	Los Angeles, CA	1142	32.15	0.37	0.23	1.28	0.78	1.24	0.06	2.24	0.06	0.12	7.08
2	Cook, IL	402	17.71	0.41	0.36	0.82	0.15	0.75	0.58	1.77	1.18	0.20	2.56
3	San Francisco, CA	395	2.28	0.56	0.23	2.06	3.44	0.27	0.664	2.18	0.07	0.18	19.11
4	Santa Clara, CA	266	5.31	0.56	0.08	0.72	0.24	1.29	0.61	1.93	0.41	0.14	15.80
5	Menominee, WI	262	0.29	0.42	0.05	2.96	4.20	1.53	0.11	4.56	0.60	13.41	0.49
6	Shannon, SD	244	0.26	0.98	0.92	4.26	6.21	2.43	0.19	5.00	0.60	14.49	0.49
7	Douglas, CO	190	0.36	6.19	11.2	3.05	1.91	2.58	0.19	0.50	0.53	0.18	1.10
8	Buffalo, SD	189	0.29	0.60	0.27	3.98	4.95	2.02	0.92	4.26	0.60	12.52	0.49
9	Rolette, ND	188	0.26	0.04	0.24	2.31	3.45	1.24	0.11	3.71	0.60	11.17	0.42
10	Sioux, ND	184	0.29	0.04	0.22	3.89	4.64	2.24	0.76	4.39	0.60	12.99	0.49

Table 1. The top 10 spatial outlier candidates detected by Mean algorithm and their associated standardized attribute values

Rank	County	Mahalanobis Distance	Pop 2001	2000-2001 change	1990-2000 change	≤ 5 %	≤ 18 %	≥ 65 %	Female %	White %	Black %	Indian %	Asian %
1	Los Angeles, CA	1306	32.15	0.37	0.23	1.28	0.78	1.24	0.06	2.24	0.06	0.12	7.08
2	Cook, IL	441	17.71	0.41	0.36	0.82	0.15	0.75	0.58	1.77	1.18	0.20	2.56
3	San Francisco, CA	395	2.28	0.56	0.23	2.06	3.44	0.27	0.664	2.18	0.07	0.18	19.11
4	Santa Clara, CA	367	5.31	0.56	0.08	0.72	0.24	1.29	0.61	1.93	0.41	0.14	15.80
5	Shannon, SD	300	0.26	0.98	0.92	4.26	6.21	2.43	0.19	5.00	0.60	14.49	0.49
6	Menominee, WI	259	0.29	0.42	0.05	2.96	4.20	1.53	0.11	4.56	0.60	13.41	0.49
7	Sioux, ND	238	0.29	0.04	0.22	3.89	4.64	2.24	0.76	4.39	0.60	12.99	0.49
8	Douglas, CO	205	0.36	6.19	11.2	3.05	1.91	2.58	0.19	0.50	0.53	0.18	1.10
9	Buffalo, SD	198	0.29	0.60	0.27	3.98	4.95	2.02	0.92	4.26	0.60	12.52	0.49
10	Harris, TX	191	11.35	0.65	0.59	1.84	1.10	1.80	0.14	1.62	0.66	0.18	2.75

Table 2. The top 10 spatial outlier candidates detected by Median algorithm and their associated standardized attribute values

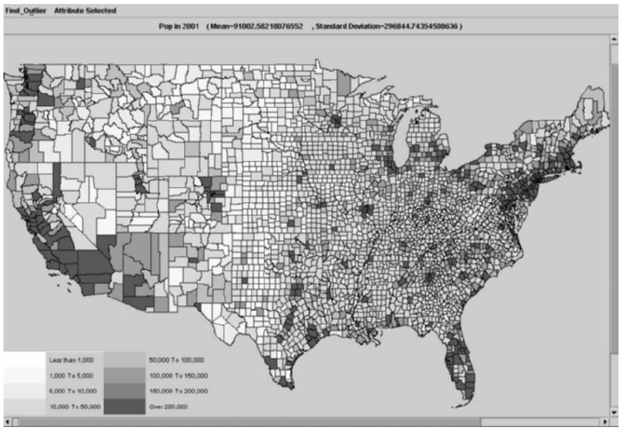

Figure 1. US Population in the Year 2001

conducted on data of all counties in the United States. Figure 1 shows an example attribute population in year 2001 used in our experiment. The high population areas in the east coast, west coast, and around Great Lakes region can be clearly observed.

The multiple attributes may have different magnitudes. For example, population of a county is usually more than 10,000, but the percentage of population change is mostly less than 1. So population of a county may dominate the value of difference function. To avoid this negative impact, we standardized the attribute values for each attribute.

The experiment results are shown in Tables 1 and 2. Note that the attribute values for each county have been standardized. The tables show only top 10 counties which are most likely to be spatial outliers. As can be seen from Tables 1 and 2, Los Angeles is selected as top spatial outlier by both algorithms, because it has the largest Mahalanobis distance, 1142 for the Mean Algorithm and 1306 for the Median Algorithm. Specifically, the largest distance mainly comes from the contribution of the corresponding attribute population (standardized value 32.15), compared with its neighboring counties, e.g., Orange Co. (9.43), Ventura Co.(2.28), San Bernardino Co. (5.64), and Kern Co. (1.97). The second spatial outlier, Cook Co, which encompasses the downtown of Chicago, is also identified due to its high population (standardized value 17.71), compared with its neighboring counties, e.g., Dupage Co. (2.76) Will Co. (1.5), Lake Co.(1.32), Kane Co. (1.12), and McHenry Co.(0.6). The third and fourth spatial outliers, San Francisco, CA and Santa Clara, CA, have high percentage of Asian population (standardized Value 19.11 and 15.80, respectively) compared with the Asian population of their neighboring counties. The remaining seven counties in both tables were detected as spatial outliers because the total contributions to the Mahalanobis distance from various attribute values are significant. From Tables 1 and 2, we can also see that the Mean Algorithm and the Median Algorithm have 9 outliers in common and the rankings of the top 4 outliers are in the same order. This shows that both of the algorithms are effective in detecting spatial outliers.

5 Conclusion

In this paper we propose two spatial outlier detection algorithms using Mahalanobis distance to analyze spatial data with multiple attributes: one algorithm based on the average of the attribute values from neighborhoods and the other based on median of the attribute values. The experimental results indicate our methods are effective in practical use. Furthermore, it carries the important bonus of ordering the spatial outliers with respect to their degree of outlierness based on the Mahalanobis distance.

Spatial outlier detection is the focus of this paper. However, there are other types of outliers, such as temporal outliers and spatial-temporal outliers, and region outliers where the data contains two neighboring regions with different ranges of attribute values. We are planning to investigate the definitions of these kinds of outliers, as well as to expand our algorithm to identify these local anomalies. Our algorithms assume that the data set can be loaded into memory to process. We are planning to investigate the issue of handling a large, disk-resident spatial data set. The goal will be to minimize the number of page reads or passes over the data set.

References

[1] C. C. Aggarwal. Redesigning Distance Functions and Distance-Based Applications for High Dimensional Data. *SIGMOD Record*, 30(1), March 2001.

[2] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park. Fast algorithms for projected clustering. In *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, pages 61–72. ACM Press, 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*, volume 30. ACM, 2001.

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

[5] S. Berchtold, C. Böhm, and H.-P. Kriegel. The pyramid-technique: Towards breaking the curse of dimensionality. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 142–153. ACM Press, 1998.

[6] M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. OPTICS-OF: Identifying Local Outliers. In *Proc. of PKDD '99, Prague, Czech Republic, Lecture Notes in Computer Science (LNAI 1704), pp. 262-270, Springer Verlag*, 1999.

[7] R. Haining. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, 1993.

[8] 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*, 45:234–242, 1991.

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

[10] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases*, pages 506–515, 2000.

[11] 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*, pages 293–298. ACM Press, 2001.

[12] R. Johnson. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[13] E. Knorr and R. Ng. A Unified Notion of Outliers: Properties and Computation. In *Proc. of the International Conference on Knowledge Discovery and Data Mining*, pages 219–222, 1997.

[14] E. Knorr and R. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proc. 24th VLDB Conference*, 1998.

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

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

[17] Y. Panatier. *Variowin. Software For Spatial Data Analysis in 2D*. New York: Springer-Verlag, 1996.

[18] F. Preparata and M. Shamos. *Computational Geometry: An Introduction*. Springer Verlag, 1988.

[19] 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*, volume 29, pages 427–438. ACM, 2000.

[20] I. Ruts and P. Rousseeuw. Computing Depth Contours of Bivariate Point Clouds. In *Computational Statistics and Data Analysis, 23:153–168*, 1996.

[21] S. Shekhar and S. Chawla. *A Tour of Spatial Databases*. Prentice Hall, 2002.

[22] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. Lu. Spatial Databases: Accomplishments and Research Needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):45–55, 1999.

[23] S. Shekhar, C. Lu, and P. Zhang. Detecting Graph-Based Spatial Outlier: Algorithms and Applications(A Summary of Results). In *Proc. of the Seventh ACM-SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, Aug 2001.

[24] S. Shekhar, C. Lu, and P. Zhang. Detecting Graph-Based Spatial Outlier. *Intelligent Data Analysis: An International Journal*, 6(5):451–468, 2002.

[25] S. Shekhar, C. Lu, P. Zhang, and R. Liu. Data Mining for Selective Visualization of Large Spatial Datasets. In *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence*, 2002.

[26] T. Johnson and I. Kwok and R. Ng. Fast Computation of 2-Dimensional Depth Contours. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 224–228. AAAI Press, 1998.

[27] U.S. Census Burean, United Stated Department of Commence. http://www.census.gov/.

[28] M. F. Worboys. *GIS - A Computing Perspective*. Taylor and Francis, 1995.