Algorithms for Spatial Outlier Detection

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. One drawback of existing methods is that normal objects tend to be falsely detected as spatial outliers when their neighborhood contains true spatial outliers. In this paper, we propose a suite of spatial outlier detection algorithms to overcome this disadvantage. We formulate the spatial outlier detection problem in a general way and design algorithms which can accurately detect spatial outliers. In addition, using a real-world census data set, we demonstrate that our approaches can not only avoid detecting false spatial outliers but also find true spatial outliers ignored by existing methods.

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 [1, 5], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [4]. The identification of outliers can lead to the discovery of useful 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 [6, 11, 16]. 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, including transportation, ecology, public safety, public health, climatology, and location based services [12].

Recent work by Shekhar et al. introduced a method for detecting spatial outliers in graph data set [13, 14]. 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 [3, 10]. Quantitative methods provide tests to distinguish spatial outliers from the remainder of data. Scatterplot [2, 8] and Moran scatterplot [9] are two representative approaches.

One major drawback of the existing detection approaches is that their application will lead to some true spatial outliers being ignored and some false spatial outliers being identified. To minimize such defect, we propose two iterative algorithms that detect spatial outliers by multi-iterations. Each iteration identifies only one outlier and modifies the attribute value of this outlier so that this outlier will not impact the subsequent iterations negatively. We also propose a non-iterative algorithm which uses the median as the neighborhood function, thus reducing the negative impact caused by the presence of neighboring points with very high/low attribute values. Using a real-world census data, we show that our algorithms can avoid detecting false spatial outliers and can find true spatial outliers ignored by existing methods when the expected number of spatial outliers is limited.

Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)
0-7695-1978-4/03 $ 17.00 © 2003 IEEE

2 Problem Formulation

Given a set of spatial points $X = \{x_1, x_2, \dots, x_n\}$ in a space with dimension $p \geq 1$, an attribute function f is defined as a mapping from X to R (the set of real number). Attribute function $f(x_i)$ 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, where $k = k(x_i)$ depends on the value of x_i for $i = 1, 2, \dots, n$. 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, we compare the attribute value of each point x_i with those attribute values of its neighbors $NN_k(x_i)$. 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. Let $y_i = h(x_i)$ for $i = 1, 2, \dots, n$. Given the attribute function f, function k, neighborhood function g, and comparison function h, a point x_i is a spatial outlier or simply S-outlier if y_i is an extreme value of the set $\{y_1, y_2, \dots, y_n\}$. 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 [15]. These include outliers detected by z algorithm [13], Scatterplot [2,8], Moran scatterplot [9] and pocket plots [3,10].

3 Proposed Algorithms

We state our algorithms to detect S-outliers. For simplicity, the description assumes all $k(x_i)$ are equal to a fixed number k. The algorithms can be easily generalized by replacing fixed k with dynamic $k(x_i)$. As seen above, 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 each algorithm. In Algorithm 1 below, 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. Comparison function $h(x)$ is taken to be the ratio of $f(x)$ to $g(x)$. Very large or very small value $h(x)$ (detected by the threshold θ) is an indication that x might be an S-outlier. Algorithm 1 is also termed as an iterative r(ratio) algorithm, since iterations are coupled with the ratios.

Algorithm 1 (Iterative r Algorithm)

1. Given a spatial data set $X = \{x_1, x_2, \dots, x_n\}$, an attribute function f, a number k of nearest neighbors, and an expected number m of spatial outliers. For each spatial point x_i, compute the k nearest neighbor set $NN_k(x_i)$, the neighborhood function $g(x_i)$ $= \frac{1}{k} \sum_{x \in NN_k(x_i)} f(x)$ and the comparison function $h_i = h(x_i) = \frac{f(x_i)}{g(x_i)}$.

2. Let h_q or h_q^{-1} denote the maximum of h_1, h_2,…, h_n, h_1^{-1}, h_2^{-1}, …, h_n^{-1}. For a given threshold θ, if h_q or $h_q^{-1} \geq \theta$, treat x_q as an S-outlier.

3. Update $f(x_q)$ to be $g(x_q)$. For each spatial point x_i whose $NN_k(x_i)$ contains x_q, update $g(x_i)$ and h_i.

4. Repeat steps 2 and 3 until either the threshold condition is not met or the total number of S-outliers equals m.

In Algorithm 2 below, the neighborhood function g is the same as in Algorithm 1. But 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 $\{h_1, h_2, \dots, h_n\}$. 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 $\{h_1, h_2, \dots, h_n\}$. Let μ and σ denote the sample mean and sample standard deviation of $\{h_1, h_2, \dots, h_n\}$. The standardized value for each h_i is $y_i = \frac{h_i - \mu}{\sigma}$, and so the standardized data set becomes $\{y_1, y_2, \dots, y_n\}$. Now it is clear that x_i is extreme in the original data set iff h_i is extreme in the standardized data set. Correspondingly, x_i is a possible S-outlier if $|y_i|$ is large enough (again detected by θ).

Algorithm 2 (Iterative z Algorithm)

1. For each spatial point x_i, compute the k nearest neighbor set $NN_k(x_i)$, the neighborhood function $g(x_i)$ $= \frac{1}{k} \sum_{x \in NN_k(x_i)} f(x)$, and the comparison function $h_i = h(x_i) = f(x_i) - g(x_i)$.

2. Let μ and σ denote the sample mean and sample standard deviation of the data set $\{h_1, h_2, \dots, h_n\}$. Standardize the data set and compute the absolute value $y_i = |\frac{h_i - \mu}{\sigma}|$ for $i = 1, 2, \dots, n$. Let y_q denote the maximum of $y_1, y_2, \dots, y_n$. For a given threshold θ, if $y_q \geq \theta$, treat x_q as an S-outlier.

3. Update $f(x_q)$ to be $g(x_q)$. For each spatial point x_i whose $NN_k(x_i)$ contains x_q, update $g(x_i)$ and h_i.

4. Recalculate μ and σ of the data set $\{h_1, h_2, \dots, h_n\}$. For $i = 1, 2, \dots, n$, update $y_i = |\frac{h_i - \mu}{\sigma}|$.

5. Repeat steps 2, 3, and 4 until either the threshold condition is not met or the total number of S-outliers equals m.

The simplest choice of θ in Algorithm 1 is $\theta = 1$. It can be larger than 1 depending on different scenarios. A common value of θ in Algorithm 2 may be taken to be 2 or 3.

This is based on the result that in Algorithm 2 the comparison function h is normally distributed if the attribute function f is normally distributed [15]. There is no clear guideline on which of the two algorithms (iterative r and z algorithms) should be selected in applications. If the attribute function f can take negative values, then it is obvious that the iterative r algorithm should not be used due to the current description of the algorithm. If the attribute function f is non-negative, the selection depends on the properties of the practical applications. In general, we recommend that both algorithms be used.

In both Algorithms 1 and 2, once an S-outlier is detected, some corrections are made immediately. These include replacing the attribute value of the outlier by the average attribute value of its neighbors and some subsequent updating computation. The effect of these corrections is to avoid normal points close to the true outliers to be claimed as possible outliers. There is a direct method to reduce the risk of overstating the number of outliers without replacing the attribute value of the detected outlier, as describe in Algorithm 3. The method in Algorithm 3 defines the neighborhood function differently. Instead of the average attribute value, $g(x_i)$ is chosen to be the median of the attribute values of the points in $NN_k(x_i)$. The motivation of using median is the fact that median is a robust estimator of the "center" of a sample.

Algorithm 3 (Median Algorithm)

1. For each spatial point x_i, compute 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 sample mean and sample standard deviation 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. For a given positive integer m, let $i_1, i_2, \ldots, i_m$ be the m indices such that their y values in $\{y_1, y_2, \ldots, y_n\}$ represent the m largest. Then the m S-outliers are $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$.

A quick illustration of Algorithms 1, 2, and 3 is to apply them to the data in Figure 1. Table 1 shows the results using the three algorithms with parameters $k = m = 3$, compared with the existing approaches. As can be seen, all the three proposed algorithms accurately detect $S1$, $S2$, and $S3$ as spatial outliers, but z algorithm, Scatterplot, and Moran Scatterplot, falsely identify $E1$ and $E2$ as spatial outliers. In this table, the rank of the outliers is defined in an obvious way. For example, in iterative r and z algorithms, the rank is the order of iterations, while in both z and Median algorithms, the rank is determined by the y value.

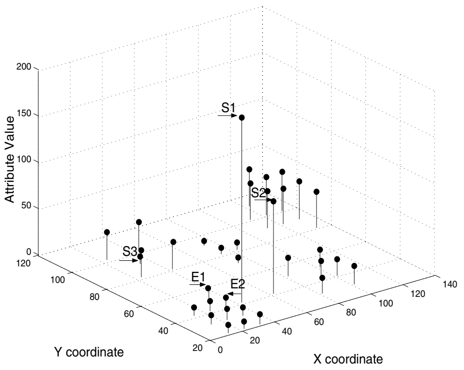

Figure 1. A spatial data set. Objects are located in the $X - Y$ plane. The height of each vertical line segment represents the attribute value of each object.

Rank	Methods					
	Scatter-plot	Moran Scatterplot	z Alg.	Iterative z Alg.	Iterative r Alg.	Median Alg.
1	E1	S1	S1	S1	S1	S1
2	E2	E1	E1	S2	S2	S2
3	S2	E2	E2	S3	S3	S3

Table 1. The top three spatial outliers detected by Scatterplot, Moran scatterplot, z, iterative z, iterative r, and median algorithms.

4 Experiments

We empirically compared the detection performance of our proposed methods with the z algorithm through mining a real-life census data set. The experiment results indicate that our algorithms can successfully identify spatial outliers ignored by the z algorithm and can avoid detecting false spatial outliers. In this experiment, we tested various attributes from census data compiled by U.S. Census Bureau [17]. The attributes tested include population, population density, percent of white persons, percent of black or African American persons, percent of American Indian persons, percent of Asian persons, and percent of female persons. We first ran the four algorithms (z, iterative r, iterative z, and median algorithms) to detect which counties have abnormal population. There are 3192 counties in the USA. We show the top 10 counties which are most likely to be the spatial outliers.

Table 2 provides the experimental results for all four spatial outlier detection algorithms. For the top 10 spatial outlier detected by z, iterative z, and median algorithms, most of them are the same with slightly different order. In fact, there are eight spatial outliers in common detected by the

Rank	Methods			
	z Alg.	Iterative z Alg.	Median Alg.	Iterative r Alg.
1	Los Angeles,CA,9637494.0	Los Angeles,CA,9637494.0	Los Angeles,CA,9637494.0	Kenedy,TX,413.0
2	Cook,IL,5350269.0	Cook,IL,5350269.0	Cook,IL,5350269.0	Loving,TX,70.0
3	Harris,TX,3460589.0	Harris,TX,3460589.0	Harris,TX,3460589.0	Treasure,MT,802.0
4	Maricopa,AZ,3194798.0	Maricopa,AZ,3194798.0	Maricopa,AZ,3194798.0	Lincoln,NV,4198.0
5	Ventura,CA,770630.0	Dallas,TX,2245398.0	Miami-Dade,FL,2289683.0	Falls Church city,VA,10612.0
6	Dallas,TX,2245398.0	Miami-Dade,FL,2289683.0	Dallas,TX,2245398.0	La Paz,AZ,19759.0
7	Miami-Dade,FL,2289683.0	Wayne,MI,2045473.0	Wayne,MI,2045473.0	Alpine,CA,1192.0
8	Wayne,MI,2045473.0	Bexar,TX,1417501.0	Clark,NV,1464653.0	Hudspeth,TX, 3318.0
9	Bexar,TX,1417501.0	King,WA,1741785.0	Bexar,TX,1417501.0	Fairfax city,VA,21674.0
10	King,WA,1741785.0	San Diego,CA,2862819.0	Tarrant,TX,1486392.0	Gilpin, CO,4823.0

Table 2. The top ten spatial outliers detected by z, iterative z, iterative r, and Median algorithms.

three algorithms. Further examination shows that the top 10 counties selected by iterative z and median algorithms are true outliers. But Ventura Co. in California was falsely detected by the non-iterative z algorithm. This falsely detected outlier was avoided by both the iterative z and median algorithms. The last column of the table shows the top ten candidate outliers from the iterative r algorithm. Although these 10 candidates are true outliers from a practical examination, they are very different from those obtained from the other three methods. This difference is due to the fact that the iterative r algorithm focuses on the ratio between the attribute value and the averaged attribute value of neighbors.

Experimental results from other attributes also show that the iterative algorithms and median method are more accurate than the non-iterative algorithm in terms of falsely detected spatial outliers. For running the algorithms and generating more results, we refer interested readers to [7], where we developed one software package which implements all the existing and proposed algorithms.

5 Conclusion

In this paper we propose three spatial outlier detection algorithms to analyze spatial data: two algorithms based on iteration and one algorithm based on median. The experimental results confirm the effectiveness of our approach in reducing the risk of falsely claiming regular spatial points as outliers, which exists in commonly used detection methodologies. Furthermore, it carries the important bonus of ordering the spatial outliers with respect to their degree of outlierness.

References

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

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

[3] 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.

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

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

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

[7] C.-T. Lu, H. Wang, and Y. Kou. http://europa.nvc.cs.vt.edu/~ctlu/Project/MapView/index.htm.

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

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

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

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

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

[13] S. Shekhar, C.-T. 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.

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

[15] S. Shekhar, C.-T. Lu, and P. Zhang. A Unified Approach to Spatial Outliers Detection. *GeoInformatica, An International Journal on Advances of Computer Science for Geographic Information System*, 7(2), June 2003.

[16] 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.

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