An Entropy-based Method for Assessing the Number of Spatial Outliers

Xutong Liu, Chang-Tien Lu, Feng Chen
Department of Computer Science
Virginia Polytechnic Institute and State University
{xutongl, ctlu, chenf}@vt.edu

Abstract

A spatial outlier is a spatial object whose non-spatial attributes are significantly different from those of its spatial neighbors. A major limitation associated with the existing outlier detection algorithms is that they generally require a pre-specified number of spatial outliers. Estimating an appropriate number of outliers for a spatial data set is one of the critical issues for outlier analysis. This paper proposes an entropy-based method to address this problem. We define the function of spatial local contrast entropy. Based on the local contrast and local contrast probability that derived from non-spatial and spatial attributes, the spatial local contrast entropy can be computed. By incrementally removing outliers, the entropy value will keep decreasing until it becomes stable at a certain point, where an optimal number of outliers can be estimated. We considered both the single attribute and the multiple attributes of spatial objects. Experiments conducted on the US Housing data validated the effectiveness of our proposed approach.

Index Terms—Spatial Outlier, Local Contrast, Spatial Local Contrast Entropy

1. Introduction

Detecting spatial outliers is an important topic in the field of spatial data mining. The goal of identifying such anomalies is to discover hidden but potentially useful knowledge. As a result, outlier detection in spatial data has attracted significant attention nowadays from geographers and data mining experts. *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 neighbors* [1]. In contrast to traditional outliers, spatial outliers are local anomalies that are extremely distinct from their neighbors, but do not necessarily deviate from the remainder of the whole data set [2]. In this sense, spatial outliers can be called "local outliers," because they focus on local instability. On the other hand, traditional outliers can be referred as "global outliers," because they are based on a global comparison.

Recently, several spatial outlier detection methods have been proposed [3-13]. A major limitation associated with them is that they all assume a pre-specified number of outliers, typically 5% of the entire data set. However, there is a well-known problem of masking and swamping effects for outlier detection. Masking occurs when true outliers are not accurately identified; swamping happens when some normal objects are erroneously flagged as outliers [14]. These detrimental effects can be alleviated by determining an appropriate number of outliers. In this regard, estimating an optimal number of outliers for a spatial data set has become one of the most essential issues in outlier analysis.

In this paper, we present an entropy-based method to tackle this problem. We define the function of Spatial Local Contrast Entropy (SLCE). Based on the relationship between outliers and the overall entropy, that is, the data set with more outliers has a higher entropy value than that with less outliers, we expect that, by incrementally removing outliers, the entropy value will decrease sharply, and reach a stable state when all the outliers have been removed.

The rest of this paper is organized as follows. Section 2 surveys the related work. Section 3 presents fundamental concepts, including the entropy and the characteristics of spatial outliers. Section 4 proposes our approach to assess the optimal number of spatial outliers. Section 5 presents experimental evaluation on a real data set by applying our method to the Point Object Method (POD [15]), and discusses the empirical results. Finally, Section 6 provides some concluding remarks.

2. Related Work

There are two basic categories for spatial outlier detection, namely, graphic and quantitative approaches. Graphic approaches utilize visualization techniques to highlight outlying objects. The representative algorithms include variogram clouds [6], pocket plots [4], scatter plot [3], and Moran scatter plot [5]. Quantitative approaches conduct statistical tests for measuring the local

978-1-4244-2660-7/08/$25.00 ©2008 IEEE

inconsistencies. Examples include Z-value [7] and iterative-Z [8] approaches. Other works explored the special property of spatial data. Shekhar *et al.* introduced a method for detecting spatial outliers in graph data sets [9]. Zhao *et al.* proposed a wavelet-based approach to detect region outliers [10]. Cheng and Li presented a multi-scale approach to detect spatial-temporal outliers [11]. Adam *et al.* proposed an algorithm which considers both the spatial and semantic relationship among neighbors [12]. Lu *et al.* proposed algorithms to detect spatial outliers with multiple non-spatial attributes by using Mahalanobis distance [13].

Recently, several works have been focused on assessing an appropriate number of clusters and outliers. Celeux *et al.* proposed an entropy-based criterion, normalized entropy criterion (NEC), to estimate the number of clusters associated to a mixture model [16]. This entropy criterion is derived from the relation between the mixture model and cluster analysis. Lu *et al.* introduced a new evolutionary algorithm for identifying the optimal number of clusters [17]. They defined an entropy-based fitness function to measure how well the mixture model fits the data samples. Barbara *et al.* studied the connection between clusters and the entropy, and concluded that the clusters with similar points have lower entropy values than those with dissimilar ones [18]. They designed a clustering algorithm, named as COOLCAT, which groups points within clusters by trying to minimize the expected entropy value of clusters. This algorithm can also be utilized to identify the optimal number of clusters. Nikhil *et al.* proposed a new concept of probabilistic entropy based on the exponential behavior of information gain [19]. Beghdadi *et al.* presented a nonlinear-noise filtering method based on the entropy definition [20].

3. Theoretical Preliminary

This section introduces the theoretical concepts of the proposed approach, including entropy, spatial local contrast, spatial local contrast probability, and spatial local contrast entropy.

Entropy: Entropy is the measure of information and uncertainty of a random variable [21, 22]. If *X* is a discrete random variable, *S(X)* is the set of possible distinct values that *X* can take, and *p(x)* is the probability function of *X*, the entropy *E(x)* can be defined as follows:

$$\mathrm{E(x)} = -\sum_{\mathrm{x}\in \mathrm{S(x)}} \mathrm{p(x)} \log\big(\mathrm{p(x)}\big) \qquad (1)$$

Spatial Local Contrast: The spatial local contrast can be viewed as the difference between an object and its surrounding neighbors. Generally, it is directly related to its spatial outlierness. Given a point object *i,* the center of an area A_i (A_i includes the point object and its surrounding *k* neighbors), its spatial outlierness value can be represented by O_i. A spatial local contrast D_i is the function of the outlierness O_i:

$$\mathrm{D_i} = \mathrm{f(O_i)} \qquad (2)$$

Spatial Local Contrast Probability: Based on the concept of "Spatial Local Contrast," we define "Spatial Local Contrast Probability" as follows:

$$\mathrm{P_i} = \frac{\mathrm{D_i}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{D_i}} = \frac{\mathrm{f(O_i)}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f(O_i)}} \qquad (3)$$

In Equation (3), *n* is the number of objects in the spatial data set. A zero spatial local contrast area, i.e., a homogeneous region, corresponds to a zero probability. Based on the definition of spatial local contrast probability, spatial local contrast entropy can be computed.

Spatial Local Contrast Entropy (SLCE): Spatial local contrast entropy can be formalized as:

$$\mathrm{H} = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\frac{\mathrm{f(O_i)}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f(O_i)}} \log\left(\frac{\mathrm{f(O_i)}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f(O_i)}}\right)\right) \qquad (4)$$

Motivated by the fact that the spatial local contrast probabilities of outliers are higher than those of other data points, we infer that an outlier point significantly contributes to the spatial local contrast entropy because its spatial local contrast probability is high. The fundamental concept of the proposed technique is that when spatial outliers are incrementally removed from the data set, the spatial local contrast entropy value will be continuously decreased until it reaches a stable state when all the outliers have been removed.

4. Proposed Approach

Previous analysis shows that the spatial local contrast entropy will be zero when a data set is homogeneous. In contrast, the spatial local contrast entropy will be very high if there are a number of outliers in a data set. The outlying objects substantially contribute to the spatial local contrast entropy because the spatial local contrast probabilities of outliers are high. Therefore, if we plot a figure in which the *x*-coordinate denotes the number of removed outliers and the *y*-coordinate denotes the spatial local contrast entropy value, we expect a curve which decreases quickly, and then reaches a point where the spatial local contrast entropy value is stabilized. This point is an appropriate estimate for the optimal number of spatial outliers.

4.1 The Sliding Window

Removing a point from the data sets will affect the spatial local contrast entropy. This is because the size of the spatial data set has been changed. For example, if the original size is n, it will become $(n - 1)$ after removing one point. In this regard, the comparison of entropy values between the data sets of different size is not legitimate.

To address this issue, we introduce the concept of a sliding window, which sets a fixed value for a data set as m. That is, after removing one outlier, we compute the spatial entropy of the m point objects. Figure 1 shows the concept of applying the sliding window concept to our approach.

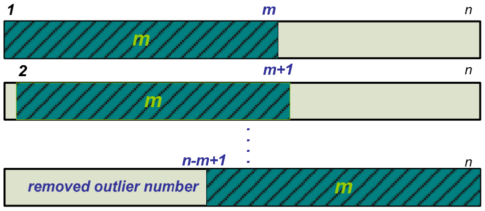

Figure 1. A sliding window

First, sort the data based on their outlierness values by a descending order. Suppose there are n data objects in the entire data set. We set a sliding window with m point objects, which specifies the computation of the spatial local contrast entropy.

Second, compute the spatial local contrast entropy from the 1^{st} point to the m^{th} point. The Spatial Local Contrast Entropy (SLCE) is computed as follows:

$$H_0 = -\sum_{i=1}^{m} P_i \log P_i \tag{5}$$

After removing the first outlier whose outlierness value is the highest, we reorganize the structure of the remaining data set to reflect the neighborhood change due to the removal of the outlier. We then re-compute the spatial local contrasts and spatial local contrast probabilities. The sliding window is shifted by one object, that is, the current window is from the 2^{nd} point to the $(m + 1)^{th}$ point. The corresponding SLCE is calculated as follows:

$$H_1 = -\sum_{i=2}^{m+1} P_i \log P_i \tag{6}$$

After removing the second outlier, the window is shifted again and the spatial entropy is computed from the 3^{rd} point to the $(m + 2)^{th}$ point:

$$H_2 = -\sum_{i=3}^{m+2} P_i \log P_i \tag{7}$$

Continue this procedure till the final spatial local contrast entropy is aggregated from the $(n - m + 1)^{th}$ point to the n^{th} point:

$$H_{n-m} = -\sum_{i=n-m+1}^{n} P_i \log P_i \tag{8}$$

4.2 Algorithm Descriptions

This section introduces the proposed algorithm, named as SLCE, to compute the Spatial Local Contrast Entropy, and discusses the major characteristics of our approach.

```
Algorithm: Spatial Local Contrast Entropy (SLCE)

Input:
    k : Number of neighbors;
    m : Size of sliding windows;
    KNN : The set of the neighborhood relationship;
    O(x_i) : An outlierness function;
Output:
    H : Sum of spatial entropy

n = size(points);
/*Step 1: Sort spatial points in a data set*/
Y = sort(points)
for( i = 1; i ≤ (n-m); i++) {
    /* Retrieve the top outlier based on outlierness values*/
    outlierID = outliers(i);
    /* Remove  the outlier from KNN */
    ruleoutOutlier(outlierID);
    /* Step2: Recompute the k nearest neighbor set */
    KNN = adjustNeighbors(KNN);
    /* Step3: Calculate the spatial local contrast */
    for( j = 1; j ≤ n - i; j++)
    { $D(x_j) = f(O(x_j))$; }
    /* Calculate the spatial local contrast probability */
    for ( j = 1; j ≤ m; j++)
    { $P(x_j) = D(x_j) / sum(D)$; }
    /* Step 4: Calculate the SLCE of the sliding window*/
    /* Get the j^th outlier from the dataset*/
    while( getNode(outlier, j, m) )
    { $H(j) = H(j) + (-1) \cdot P(x_j) \cdot \log_2(P(x_j))$; }}
Output(H);
```

The outlierness function $O(x_i)$ is computed based on different approaches to identify spatial outliers. Given a spatial data set $X = \{x_1, x_2, \ldots, x_n\}$, an outlierness function $O(x_i)$, two positive integer number k (the number of neighbors) and m (the size of sliding window), the major steps in this algorithm are described as follows:

Step 1: Sort spatial points based on outlierness values

Compute a data set $Y = \{y_1, y_2, \ldots, y_n\}$ by decendingly sorting spatial points based on their corresponding outlierness values.

Step 2: Recompute the nearest neighbor set

Remove the top outlier from the current list, and for each spatial point x_i, recompute the k nearest neighbor set $NN_k(x_i)$.

Step 3: **Compute local contrast probabilily**

Let $D(x_i)$ and $P(x_i)$ denote the spatial local contrast and spatial local contrast probability of a point $x_{i,}$ respectively. They are both the functions of $O(x_i)$. *That is*, $D(x_i) = f(O(x_i))$, and $P(x_i)= f(O(x_i))/\Sigma f(O(x_i))$ for $i = 1, 2, \ldots, m$.

Step 4: Compute the spatial entropy

Using the *getNode* function, retrieve all the outliers in the sliding window. Based on the $D(x_i)$ and $P(x_i)$, calculate the spatial local contrast entropy: $SLCE = - \Sigma (f(O(x_i))/ \Sigma f(O(x_i))) \log_2(f(O(x_i)) / \Sigma f(O(x_i)))$.

Step 5: Repeat step 2 to step 4 for *(n - m)* iterations

From the original data set, we compute the spatial local contrast entropy H_0, from the 1^{st} point to the m^{th} point. After removing the first outlier, we reorganize the data set, and recompute the $NN_k(x_i)$ and $O(x_i)$ for those points with neighborhood updates. We can then calculate the next spatial entropy H_1 from the 2^{st} point to the $(m+1)^{th}$ point. We continue this procedure until the last spatial entropy H_{n-m} is computed from $(n - m + 1)^{th}$ point to the n^{th} point. Based on the set of SLCE values, we plot a curve in which the number of removed outliers is denoted by the *x*-axis and the corresponding entropy value is denoted by the *y*-axis. From this curve, we assess the cut-off point to estimate an optimal number of spatial outliers.

5. Experiment Results and Analysis

In this section, we present experimental results on a real data set, Fair Market Rents data, provided by the PDR-DHUD (Policy Development and Research, U.S. Department of Housing and Urban Development). The data set includes the rental prices for efficiencies, one-bedroom apartments, two-bedroom apartments, three-bedroom apartments, and four-bedroom apartments in 3000+ counties of the US. In the experiment, the outlierness function is defined by the POD method [15], which is a graph-based method to identify the spatial outliers. Note that our method can also be applied to other spatial outlier detection methods. The POD method first constructs a graph based on the *k* nearest neighbor relationship in the spatial domain, assigns the non-spatial attributed differences as edge weighs, and then continuously cuts high weight edges to identify isolated points or regions that are dissimilar to their neighboring objects as spatial outliers. The experiment is composed of two components, including single attribute (the rental prices for one-bedroom apartments) and multiple attributes (the rental prices for one-bedroom apartments, the rental prices for two-bedroom apartments, the rental prices for three-bedroom apartments and the rental prices for four-bedroom apartments).

In the first set of experiments, we set *m,* the size of the sliding window, as 50, 100 and 150 respectively, and *k,* the number of the nearest neighbors, as 8 and 10. Figure 2, 3 and 4 were generated by our algorithm for the single attribute. As can be seen, the values of spatial entropy decreases in the beginning and reaches a stable state at a certain point, which shows an estimate of the optimal number of outliers is around 120. When considering the multiple attributes of the spatial objects, the value of *m* is set as 50, 100 and 150, respectively, and *k* is equal to 10. Figure 5, 6 and 7 were generated by the SLCE algorithm for multiple attributes. We can observe a similar precipitating trend in which an estimated outlier number is around 90.

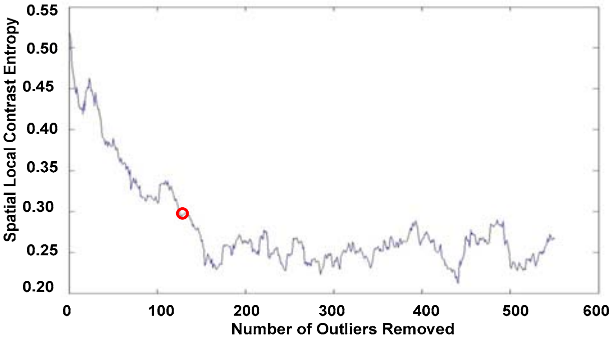

Figure 2. Single attribute, SLCE curve (k=8, m=50)

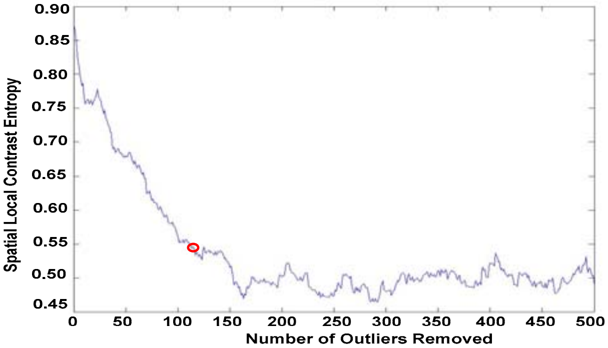

Figure 3. Single attribute, SLCE curve (k=10, m=100)

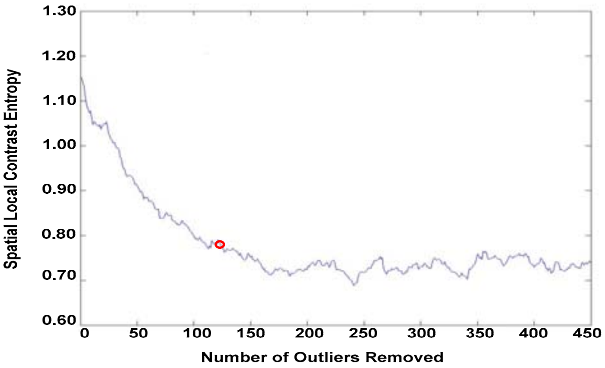

Figure 4. Single attribute, SLCE curve (k=10, m=150)

Discussion: As shown in Figures 2-7, for both cases of single attribute and multiple attributes, the dominant pattern as exhibited in the experiment is consistent with the theoretical analysis discussed in Section 4. That is, there does exist an "inflexion" point which corroborates the relationship between the spatial local contrast entropy and the number of outliers removed. Because the object with the highest outlierness value will contribute most to the spatial local contrast entropy, we can greedily reduce the overall entropy value by incrementally removing the outlying object, whose local inconsistence value is the highest. When all outliers have been removed, if we continue to remove normal objects, the spatial local contrast entropy will not change significantly. It is reasonable to assume that the outlierness values of normal objects are randomly distributed. Then the removal of normal objects based on the ranks of their corresponding outlierness values will not drastically disturb the background distribution. Therefore, the background spatial local contrast entropy will keep persistent. Take Figure 2 as an example, we can observe that the estimate of an optimal number of outliers is around 120. When the number of candidate outliers removed is greater than 120, the spatial local contrast entropy indicates a constant mean value around 0.25. The trembling of the curve is due to a normal variance.

Ideally, the SLCE curve will monotonously decrease with a continuously decreasing slope until the spatial local contrast entropy becomes stable when all the outliers have been removed. However, the outlierness value of each object, which is approximated based on the POD method, may not be identical to the true outlierness value. That is, during each incremental process, the object removed is not certainly to be the most "significant" outliers. Therefore, the slope will not monotonously decrease in strict relation to the number of outliers removed. Considering the case of false positives, in which normal objects are misidentified as outliers, potentially the ratio of outliers to normal objects will be erriously increased, and hence will the spatial local contrast entropy. As a result, the curve will not have the distinct characteristic of decreasing monotonicity. It is reasonable that the SLCE curve may have some small jumps before the removal of all outliers. As shown in Figure 4, when the number of outliers removed is around 22, the spatial local contrast entropy stops decreasing and incurs a small jump, but it quickly resumes the decreasing trend when the number is larger than 25. This pattern reveals that, it is necessary to evaluate the curve entirely, in order to estimate a close-to-optimal number of outliers.

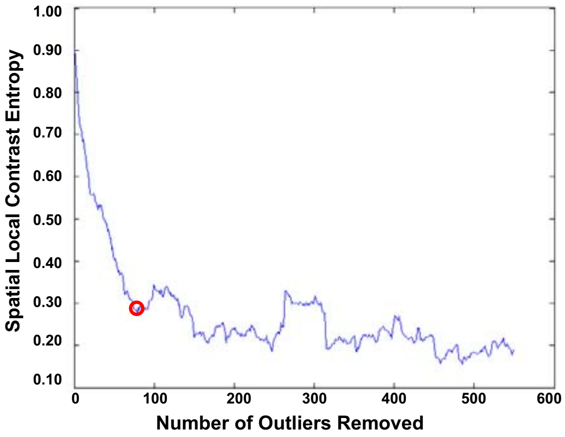

Figure 5. Multiple attributes, SLCE curve (k=10, m=50)

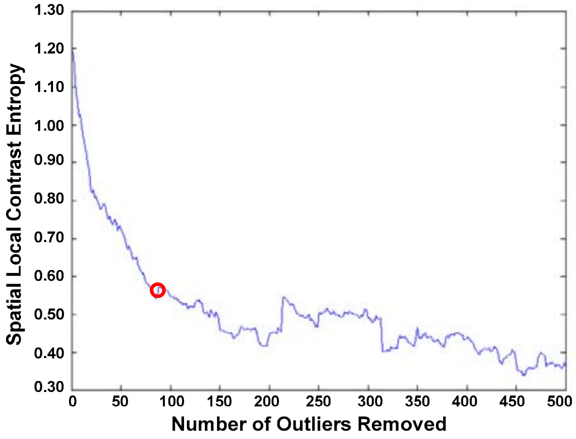

Figure 6. Multiple attributes, SLCE curve (k=10, m=100)

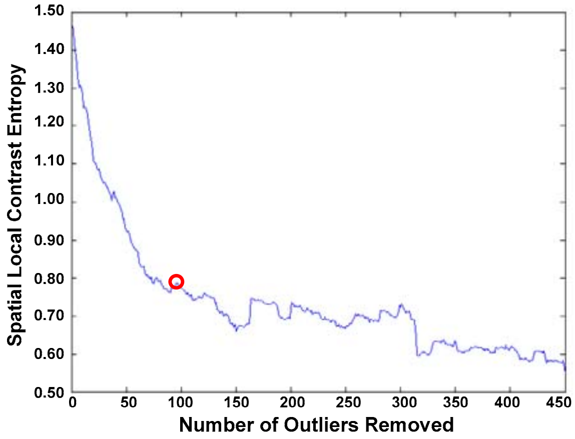

Figure 7. Multiple attributes, SLCE curve (k=10, m=150)

6. Conclusion

Detecting spatial outliers is an important topic in the field of spatial data mining. The goal of identifying such anomalies is to discover hidden but potentially useful knowledge. A major limitation of existing outlier detection algorithms is that they generally require a pre-determined number of spatial outliers, which is not suitable in real applications due to the well-known masking and swamping effects. In this paper, we present an entropy-based approach to assess an optimal number of spatial outliers in a spatial data set. Specifically, we define the formula of spatial local contrast entropy (SLCE), analyze the theoretical characteristics and foundation, and propose an effective algorithm. Experiments were conducted on the cases of single attribute and multiple attributes. The empirical results validate that our proposed approach can appropriately reveal an “inflexion” point for the outlier number assessment.

References

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

[2] A. Cerioli and M. Riani, "The Ordering of Spatial Data and the Detection of Multiple Outliers," *Journal of Computational and Graphical Statistics,* vol. 8, pp. 239-258, 1999.

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

[4] J. Haslett, R. Bradley, 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, pp. 234-242, 1991.

[5] L. Anselin, "Local Indicators of Spatial Association (LISA)," *Geographical Analysis,* vol. 27, pp. 93-115, 1995.

[6] P. Yvan, *Variowin: Software for Spatial Data Analysis in 2D*. New York, 1996.

[7] S. Shekhar, C.-T. Lu, and P. Zhang, "A Unified Approach to Spatial Outliers Detection," *An International Journal on Advances of Computer Science for Geographic Information System,* vol. 7, pp. 139-166, 2003.

[8] C.-T. Lu, D. Chen, and Y. Kou, "Algorithms for Spatial Outlier Detection," in *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 597-600, 2003.

[9] 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 SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 371-376, 2001.

[10] 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, 2003.

[11] T. Cheng and Z. Li, "A Hybrid Approach to Detect Spatial-temporal Outliers," in *Proceedings of the 12th International Conference on Geoinformatics Geospatial Information Research*, pp. 173-178, 2004.

[12] N. R. Adam, V. P. Janeja, and V. Atluri, "Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 576 - 583, 2004

[13] C.-T. Lu, D. Chen, and Y. Kou, "Detecting Spatial Outliers with Multiple Attributes," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 122-128, 2003.

[14] J. Hardin and D. M. Rocke, "The Distribution of Robust Distances," *Journal of Computational and Graphical Statistics,* vol. 14, pp. 928-946, 2005.

[15] Y. Kou, C.-T. Lu, and R. F. DosSantos, "Spatial Outlier Detection: A Graph-Based Approach," in *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pp. 281-288, 2007.

[16] G. Celeux and G. Soromenho, "An Entropy Criterion for Assessing the Number of Clusters In A Mixture Model," *Journal of Classification,* vol. 13, pp. 195-212, 1996.

[17] W. Lu and I. Traore, "Determining the Optimal Number of Clusters Using a New Evolutionary Algorithm," in *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, pp. 712 - 713, 2005.

[18] D. Barbará, Y. Li, and J. Couto, "COOLCAT: An Entropy-based Algorithm for Categorical Clustering," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 582 - 589, 2002.

[19] N. R. Pal and S. K. Pal, "Entropy: a New Definition and Its Applications," *IEEE Transactions On Systems. Man, and Cybertics,* vol. 21, pp. 1260-1270, 1991.

[20] A. Beghdadi and A. Khellaf, "A Noise-Filtering Method Using a Local Information Measure," *IEEE Transactions on Image Processing,* vol. 6, pp. 879-882, 1997.

[21] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal,* vol. 27, pp. 379-423, July 1948.

[22] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," *University of Illinois Press,* 1949.