Spatial outlier detection method based on spatial clustering

DENG Min, LIU Qiliang, LI Guangqiang

Department of Surveying and Geo-informatics, *Central South University*, *Hunan Changsha* 410083, *China*

Abstract: Spatial outlier detection has been a hot issue in the field of spatial data mining and knowledge discovery. Spatial outliers may be utilized to discover and predict the potential change laws or development tendency of geographical phenomenon in the real world. Among the existing spatial outlier detection methods, there are mainly two aspects of issues. On the one hand, these methods primarily consider that all the entities for outlier detection are correlated. Actually, spatial correlation decreases with the increase of distance. Entities will become independent with each other at a distance of rang. Thus, current methods can only discover the obviously outliers in the whole, some local outliers may not be detected. On the other hand, the spatial outlier measures are not enough robust, which are seriously influenced by the construction process of spatial neighborhoods of spatial entities and the possible outliers in spatial neighborhoods. To overcome these two limitations, spatial clustering as a means is firstly employed to extract the local autocorrelation patterns, called clusters. Then, a robust spatial outlier measure is proposed to determine spatial outliers in each cluster. This method is able to detect spatial outliers more accurately. Finally, a practical example is utilized to demonstrate the validity of the spatial outlier detection method proposed in this paper. The comparative experiment is also provided to further demonstrate the method in this paper to be superior to classic SOM method.

Key words: spatial outlier detection, spatial clustering, spatial outlier measure, spatial data mining

CLC number: TP751.1 **Document code:** A

Citation format: Deng M, Liu Q L, Li G Q. 2010. Spatial outlier detection method based on spatial clustering. *Journal of Remote Sensing.* **14**(5): 944—958

1 INTRODUCTION

Spatial outlier detection is one of hot issues in the fields of spatial data mining and knowledge discovery (Li *et al.*, 2002; Pei *et al.*, 2001). It aims to discover a small part of spatial entities which deviate from the normal patterns in spatial database. Spatial outliers may imply some unexpected geographical events, processes or development tendency of geographical phenomenon. In recent years, spatial outlier detection has played an important role in many applications, such as geological disaster monitoring, extreme meteorological events monitoring, mineralization forecasting, and remotely sensed image processing.

Spatial outliers can be defined as the spatial entities whose non-spatial attribute values are significantly different from the values of their spatial neighbors (Shekhar *et al*., 2003). In the process of spatial outlier detection, it needs to define a neighborhood for each point entity, and then spatial outliers are identified by means of the non-spatial attribute deviation. The non-spatial attribute is referred as the thematic attribute of a spatial entity, for example the heavy metal concentration of a soil sampling point. Existing methods of spatial outlier detection can be roughly classified into five types: (1) statistical methods (Hawkins, 1980); (2) graphics-based methods, such as variable plot or clouds (Haslett *et al*., 1991); (3) distance-based method (Liu *et al*., 2001; Shekhar *et al*., 2001, 2003; Ma & He, 2006; Chen *et al*., 2008; Zheng *et al*., 2008; Li *et al.*, 2009); (4) density-based methods (Breunig *et al*., 2000; Chawla & Sun, 2006; Huang *et al*., 2006), and (5) clustering-based methods (Li *et al*., 2008, 2009). The statistical methods use some statistical distribution to fit the dataset, so that they do not work well in the case that the dataset does not satisfy the assumed statistical distribution. Both distance-based and density-based methods only consider the non-spatial deviation in a neighborhood, the property of local correlation among spatial entities is neglected. Moreover, a robust method to measure the deviation degree of a spatial entity should be developed. A few spatial clustering methods have the ability to discover spatial outliers as the entities which do not belong to any cluster. However, clustering-based methods usually cannot find high quality outliers. To overcome the above-mentioned limitations, we integrate the clustering-based method with the distance-based method to develop a spatial clustering based spatial outlier detection method. A robust spatial outlier measure is also proposed.

Received: 2009-08-25; **Accepted**: 2010-03-19

Foundation: National 863 High Technology Research and Development Program of China (No. 2009AA12Z206), Key Laboratory of Geo-Informatics of State Bureau of Surveying and Mapping (No. 200805), Scientific Research Foundation of Jiangsu Key Laboratory of Resource and Environmental Information Engineering at China University of Mining and Technology (No.20080101) and Innovation Research Foundation of Central South University (No. 713360010).

First author biography: DENG Min (1974—), male, professor. He received his doctoral degree from Wuhan University in June 2003 and Asian Institute of Technology in December 2004. Current research interests include spatio-temporal data mining, reasoning and analyzing. He has more than 90 papers published in journals. E-mail: dengmin028@tom.com

2 RELATED WORKS, EXISTED PROBLEMS AND OUR STRATEGY ON SPATIAL OUTLIER DETECTION

2.1 Related works and existed problems

Shekhar *et al*. (2001, 2003), Chen *et al.* (2008) define the spatial outlier measure employed by the difference between the non-spatial attribute value of target entity and the average (or median) value of the entities in its neighborhood. Then, statistical test is utilized to identify spatial outliers. This method is useful for discovering global outliers but may not be able to discover local outliers (Chawla & Sun, 2006). Indeed, the local spatial outlier measures are defined based on the concept of density in some literatures (e.g. Breunig *et al.,* 2000; Chawla & Sun, 2006; and Huang *et al*., 2006). However, the detection results are seriously affected by the definition of neighborhood (Zheng *et al*., 2008) and the outliers in the neighborhood. The methods proposed by Liu *et al*. (2001), Zheng *et al*. (2008) and Li *et al*. (2009) consider the distances among a set of neighbors. These methods only take the spatial relationships among the entities in a neighborhood into count, the local correlation among the entities is seldom considered. Furthermore, the method to define spatial proximate relation and the outliers in a neighborhood also should be paid more attention to. In a word, current methods of spatial outlier detection mainly have two aspects of disadvantages as follows.

On the one hand, existing methods only consider that each entity is related to all the other ones in its neighborhood, the heterogeneity in global is usually ignored. Actually, spatial entities are usually locally correlative; the heterogeneous character is obvious in the global view (De Smith *et al.*, 2007). For example, in a large region, there are usually some sub regions in which the non-spatial attribute values of the entities are quite different. In this condition, the entities in a neighborhood may not correlate to each other, and the spatial outlier measure method cannot work well. Moreover, if there are many outliers in a region, then some local outliers may be merged. Taking the simulated dataset used by Chawla and Sun (2006) for instance, spatial entities and their non-spatial attribute values are shown in Fig.1(a). The spatial local outlier measure (SLOM for short) core for each spatial entity is performed in Fig.1(b). There are five spatial outliers detected by the SLOM method which are marked by slash in Fig.1(b). In region II and III, the SLOM values of the two shadow locations (0.14 and 0.12) are significantly large, and they should be two local outliers. But, there are too many outliers in region I, so other local outliers cannot be discovered. We can hold the hypothesis that a pollution source locates in region I, thus many spatial outliers occurred in this region corresponding. Outliers in other region (II and III) may be caused by some potential factors, but current methods cannot used to find the reasons. Based on the first law of geography (Tobler, 1970), the non-spatial values of the monitoring points are more related in region I, II, III respectively than

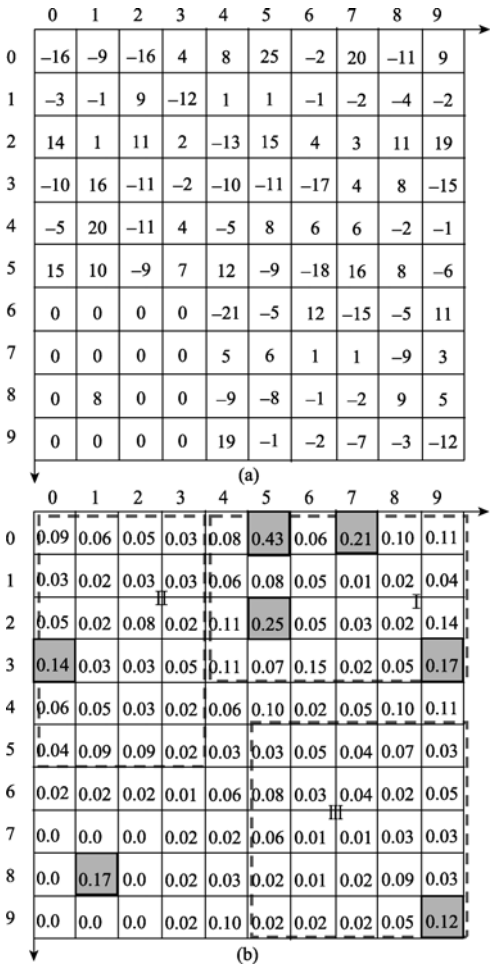

Fig. 1 Simulated spatial data set(Chawala & Sun, 2006)
(a) Original data matrix; (b) SLOM matrix

those in different regions. So the local character of the spatial dataset should be fully considered.

On the other hand, robust spatial outlier measure should be developed. For spatial point entity, more parameters are often needed to construct the spatial neighborhood. Different parameter may lead to different detection results. Furthermore, there may be some outliers in the neighbors of a spatial entity, and they also made negative impact on the spatial outlier measure. It can be concluded that current spatial outlier measure is usually seriously influenced by the dynamic construction of spatial neighborhoods of spatial entities and the possible outliers in spatial neighborhoods.

2.2 Strategy on detecting spatial outliers

To solve the limitations above, our strategy is to integrate spatial clustering with the distance-based spatial outlier measure to develop a spatial clustering based spatial outlier detection method. Spatial clustering can be seen as a procedure that partitions the spatial entities according to spatial correlation. Therefore, in our strategy, spatial clustering is utilized to group the spatial entities in a spatial database into some meaningful subgroups, also called clusters. Those entities in a same cluster are similar to each other, and the entities in different clusters have a larger difference. That is, spatial entities in a same cluster are

more related to each other. The differences among entities in different clusters can be comprehended as the heterogeneity. Further, a more robust spatial outlier measure method is also developed in this paper. In this method, the Delaunay triangulated irregular network (D-TIN for short) is firstly utilized to construct spatial proximity among each cluster, and then to recover the possible outliers in a neighborhood according to the non-spatial attribute gradient. Finally, the difference between the observation value of an entity and the inverse distance weighted interpolation value of that entity is taken for a spatial outlier measure that is similar to the distance based methods (Zheng *et al.*, 2008; Li *et al.*, 2009).

To sum up, our strategy to detect spatial outliers can be implemented in three steps: (1) spatial clustering; (2) detection of spatial outliers in each cluster, and (3) spatial outlier interpretation and evaluation.

3 SPATIAL CLUSTERING BASED SPATIAL OUTLIER DETECTION

3.1 Spatial cliustering

Existing spatial clustering methods can be mainly classified as: partition methods, hierarchical methods, density-based methods and grid-based methods (Han & Kamber, 2005). In practical applications, spatial entities usually distribute unevenly, so spatial clustering method should adapt to the change of local densities. Thus, we take a spatial clustering method adaptive to the change of local densities-ADBSC (Li *et al.*, 2009) to cluster spatial entities. ADBSC algorithm uses maximum distance in k-spatial nearest neighborhood to measure the spatial local density, and spatial entities with similar densities are grouped into a same cluster. Two parameters should be set for the ADBSC algorithm, the number of the spatial nearest neighborhood (k) and the threshold of the density variation proportion (ε). A heuristic strategy is employed to determine the parameters similar to Ester *et al*. (1996). Before detecting spatial outliers, spatial entities are first clustered into some clusters, and each cluster can be seen as a local strongly correlated region.

3.2 Robust spatial outlier measure

First, the spatial proximity relation among entities is constructed by D-TIN in each spatial cluster. The D-TIN can be utilized to identify the natural neighboring entities (Mccullagh & Ross, 1980). But this method may be inaccurate at the borders and the interspaces of the dataset (Zheng *et al*., 2008; Li *et al*., 2009). In this paper, the results of spatial clustering can be used as a prior distance constraint, so the error in creating spatial neighborhood can be significantly reduced. Take a simulated dataset to illustrate the advantage of our method, the spatial proximate relationship among the simulated dataset constructed by D-TIN is shown in Fig.2(a); spatial proximate relationship established after the spatial clustering process (k=4, ε=22%) is present in Fig.2(b). It can be found that, in Fig.2(a), the neighborhoods of the entities on the border are obviously imprecise. Indeed, the results obtained by the clustering procedure are more reasonable.

Next, we will recover the non-spatial attribute values of some possible outliers in the neighborhood. Generally, the number of spatial entities in a neighborhood is small, spatial outliers cannot be identified by the statistical test. The entities whose non-spatial attribute values are maximum and minimum in a neighborhood can be removed by a trim procedure (Chawla & Sun, 2006). Additionally, the spatial outlier measure may be unreliable cause by the directional bias when many entities gathered in certain direction after the trim procedure. In this paper, we employ the non-spatial attribute gradient to recover the non-spatial value of the likely outliers in a neighborhood. The recovery operation does not change the inherent non-spatial attribute of a spatial entity, it is only a temporal process when calculate the spatial outlier measure. The details can be described as follows.

Definition 1: Non-spatial attribute gradient: Given a spatial entity P, the spatial neighbors of which is expressed as $\boldsymbol{N}(P)=\{X_1, X_2 \ldots X_n\}$, and $f(X_i)$ is referred as the non-spatial attribute value. The non-spatial attribute gradient is the ratio of the absolute value of the non-spatial attribute value difference between entity P and entity X_i in its neighborhood to the distance between them, it can be expressed as:

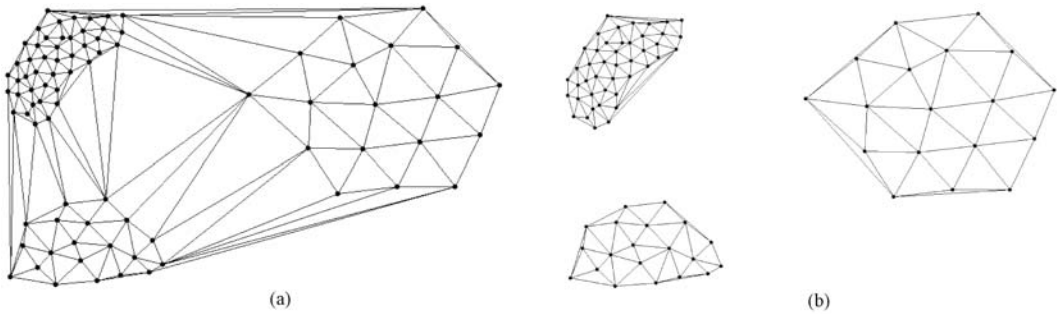

Fig. 2 Construction of spatial neighborhood
(a) D-TIN for all the dataset; (b) D-TIN for each cluster

$$G(P,X_i)=\frac{|f(P)-f(X_i)|}{D(P,X_i)} \quad \forall X_i \in \boldsymbol{N}(P) \tag{1}$$

where, $D(P, X_i)$ is the Euclidian distance between entity P and X_i.

For a spatial entity P, the non-spatial attribute value recovery operation for the possible outliers in its neighborhood can be portrayed as follows:

(1) Let $f(P)=0$, calculate the non-spatial attribute gradient between P and each entity in its neighbor, denoted as $G(P, X_i)$. Then, rank all the $G(P, X_i)$ values from small to large, and obtain the median of the sequence, denoted as $M(P)$.

(2) For each entity X_i in the neighborhood of P, compute the deviation of the non-spatial attribute gradient, $GD(X_i)$, there having: $GD(X_i)=|G(P,X_i)-M(P)|$, and then rank the deviation values from large to small, defined as sequence $\boldsymbol{GD}(P)$.

(3) The neighbors of P can be classified into three ranks on the basis of the deviation of the non-spatial attribute gradient, and the top $[(n+1)/3]$ entities form the candidate recover set $\boldsymbol{R}(P)$. Then, recover the non-spatial attribute value of the entities in $\boldsymbol{R}(P)$. $f_R(X_i)$ is the fixed value, defined as follows:

$$f_R(X_i)=M(P)D(P,X_i) \quad \forall X_i \in \boldsymbol{R}(P) \tag{2}$$

From Eq. (2), it can be found that, the recover process also can consider the spatial correlation among entities (change with distance). Finally, the robust spatial outlier measure and the method to identify spatial outliers can be described as follows:

Definition 2: Robust spatial outlier measure: Given a spatial entity P, the robust spatial outlier measure (RSOM for short) is defined as the difference between the non-spatial attribute value of P and its inverse distance weighted interpolation value, expressed by

$$RSOM(P)=\left|f(P)-\frac{\sum_{i=1}^{n}\frac{f(X_i)}{D(P,X_i)}}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}}\right|=\frac{\sum_{i=1}^{n}\frac{|f(p)-f(X_i)|}{D(P,X_i)}}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}}=\frac{\sum_{i=1}^{n}G(P,X_i)}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}}$$
$$\forall X_i \in N(P) \tag{3}$$

Definition 3: Spatial outlier: Given a spatial database $\boldsymbol{SDB}=\{X_1, X_2, X_3, \ldots, X_n\}$, the robust spatial outlier measures of which form the set $\boldsymbol{S}_{RSOM}=\{\text{RSOM}(X_1), \text{RSOM}(X_2), \ldots \text{RSOM}(X_n)\}$. The average RSOM value in $\boldsymbol{S}_{RSOM}$ is denoted as μ, and the standard deviation is denoted as σ. Let $\boldsymbol{S}_{\text{outlier}}$ express the spatial outlier set, and it can be defined as follows:

$$\boldsymbol{S}_{\text{outlier}}=\{X_i \mid RSOM(X_i)-\mu>2\sigma, X_i \in \boldsymbol{SDB}\} \tag{4}$$

For small samples (the number is smaller than 30), the spatial outlier judgment law in Eq. (4) may be not robust, so we also given the robust valuation of the parameter μ and σ, μ= *Median* ($\boldsymbol{S}_{RSOM}$), *Median* ($\boldsymbol{S}_{RSOM}$) is the median of $\boldsymbol{S}_{RSOM}$; σ= *MAD* ($\boldsymbol{S}_{RSOM}$), *MAD* ($\boldsymbol{S}_{RSOM}$) is the well-known median absolute deviation (MAD), defined as:

$$MAD(S_{RSOM})=Median\{|RAOM(X_1)-Median(S_{RSOM})|, |RAOM(X_2)-Median(S_{RSOM})|, \ldots, |RAOM(X_n)-Median(S_{RSOM})|\} \tag{5}$$

Next, a hypothetical dataset is utilized to illustrate the advantage of the *RSOM* compared to the classic *SOM* method. The *SOM* method employs the difference between non-spatial value of an entity and its inverse distance weighted interpolation value to measure the deviation of a spatial entity. In Fig. 3, each node of the D-TIN represents a spatial entity. The numbers represent non-spatial attribute values of the entities. It can be easily found that the entity B is a spatial outlier, since its non-spatial attribute value is significantly different from those of its neighbors. From Table 1, it can be found that the *SOM* value of the entity A is obviously large because outlier B is located nearby. So, the entity A may be wrongly identified as a spatial outlier. However, the *RSOM* method proposed in this paper can identify A as a normal entity with the help of the non-spatial attribute value recovery operation in a neighborhood.

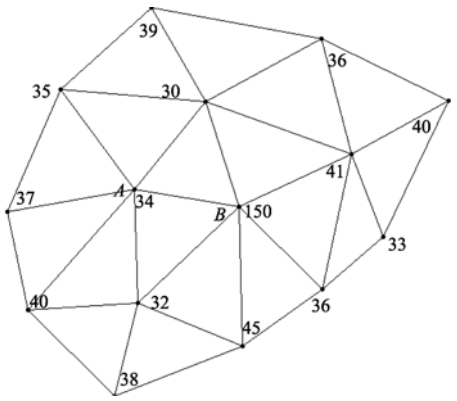

Fig. 3 An example

Table 1 Results of spatial outlier measure

Data point	Observation value	Interpolated value		Spatial outlier measure	
		SOM	*RSOM*	*SOM*	*RSOM*
A	34	56.59	34.40	22.59	0.4
B	150	36.14	38.76	113.86	111.24

In addition, spatial entity may contain multiple non-spatial attributes in practical applications. The spatial outlier measure method developed in this paper also can adapt to detect multiple non-spatial attribute outliers. It can be implemented in three steps. The first is to normalize all the non-spatial attribute dimensions, and then to calculate the *RSOM* value for each non-spatial attribute dimension. The last is to obtain the overall deviation of the entire non-spatial attributes for each spatial entity.

Given a spatial database ***SDB*** which contains n spatial entities, the dimensions of the non-spatial attribute is d. In the first step, the complexity of the ADBSC algorithm is about $O(nlogn)$. It takes $O(nlogn)$ to construct D-TIN and $O(6n)$ to form the spatial proximate relationship (Li *et al*., 2009). The cost of standardize is $O(dn)$. To calculate the *RSOM*, it takes about $O(2nlogn)$ to recover the non-spatial attribute in each neighborhood, $O(6dn)$ to compute the *RSOM*, and $O(nlogn)$ to sort the *RSOM* value. So, the total complexity for our method is

O ($nlogn$)+ O ($nlogn$)+ O($6n$)+ O(dn)+ O ($2nlogn$)+ O($6dn$)+ O ($nlogn$)≈ O ($nlogn$).

4 EXPERIMENTS

In this paper, the soil heavy metal monitoring dataset in a South China city is utilized to illustrate the validity of our method. The Cr concentration is selected as the non-spatial value for each monitoring point. There are 104 monitoring points in the dataset. The spatial distribution of the dataset is shown in Fig. 4. Moreover, the experiment result is compared to the *SOM* method. The *SOM* method employed the D-TIN to construct the spatial neighborhoods for the entire dataset, and the ranked m spatial entities in the *SOM* set are identified as spatial outliers.

First of all, spatial clustering is performed. The clustering result is presented in Fig. 5(k=5, ε=23%). Ten spatial clusters are obtained (shown in Fig. 5(b)). Spatial proximate relationship among each cluster is shown in Fig. 5(c). The spatial outliers identified by our method are shown in Table 2, and the spatial distribution of spatial outliers is presented in Fig. 6.

The spatial outliers detected by *SOM* method are shown in Table 3. We select five spatial outliers from Table 3, and form the spatial outliers set $S_{outliers}$= {97, 71, 22, 70, and 40}. Comparing the result obtained by *SOM* to that obtained by *RSOM*, one can find that the *SOM* method only is able to discover the global spatial outliers, some local outliers are ignored. For example, point 90 in cluster 9 and point 86 in cluster 10, though the deviation of them is not obvious in global, the non-spatial attribute values of them are quite different from their neighbors locally (shown in Table 4), and they should be local outliers. At the same time, it can prove that the spatial outlier identification method developed in this paper is also robust, and there are only five entities in cluster 10.

Fig. 4 Spatial distribution of the sample points

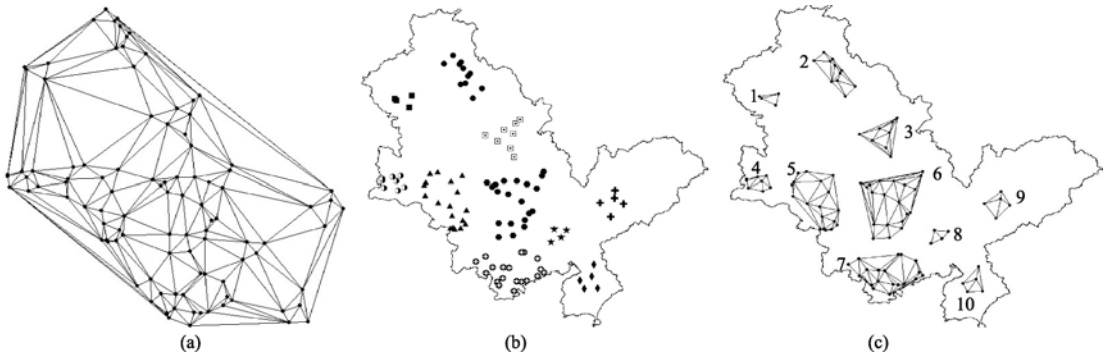

Fig. 5 Spatial clustering results and the construction of spatial neighborhoods
(a) D-TIN for entire dataset; (b) spatial clustering result ; (c) spatial proximate relationship for each cluster

Table 2 Results of spatial outlier detection via clustering based method

Cluster number	Point number	*RSOM*	Cluster number	Point number	*RSOM*	Cluster number	Point number	*RSOM*
2	1	43.04	4	51	25.57	6	16	39.05
	40	36.98		55	25.80		98	35.18
3	42	31.10	5	22	52.40	7	69	46.79
	48	30.75		97	61.65		70	47.68
9	90	31.30	10	86	20.78		71	59.62

The land use type, elevation, and pollution source are utilized to further analyze the cause of the spatial outliers. The distribution of pollution sources is present in Fig. 6. In Table 5, land use types and the evaluators of spatial outliers and their neighbors are enumerated.

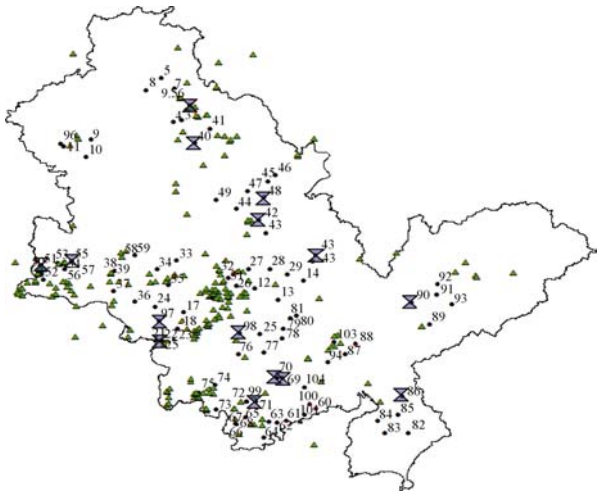

Fig. 6 Locations of spatial outliers
(▲-Pollution source⧗-Spatial outlier)

Table 3 Results of spatial outlier detection via *SOM* method

Point number	*SOM*
97	60.67
71	59.43
22	52.66
70	47.16
40	45.12
34	43.21
1	41.63
69	39.67
35	32.88
…	…

Table 4 Non-spatial attributes of entities in cluster 9 and 10

Cluster number	Point number	Cr concentration
9	89	19.26
	90	2.88
	91	29.34
	92	22.88
	93	9.47
10	82	20.63
	83	12.05
	84	28.35
	85	24.58
	86	3.9

According to above analysis, the main causes of the spatial outliers can be concluded in Table 6. Moreover, we can obtain the following rules: (1) Elevation difference between the spatial outliers and their neighbors is the most important cause of the spatial outlier, and the spatial distribution of the soil heavy metal is seriously influenced by the elevation. (2) The spatial outliers 71, 1, 55, 51, 42 are closely related to the pollution source, they may be impacted by some factories nearby. (3) The land use types of the spatial outliers are mainly paddy field and vegetable field, pesticides, fertilizers abuse may be an important factor for these spatial outliers. (4) Many spatial outliers occurred when the vegetable field is close to the paddy field. This may be a potential spatial association rule.

5 CONCLUSIONS AND FUTURE WORKS

Spatial outlier detection is a powerful tool which can be used to discover and interpret the potential geographic laws or development tendency of geographic phenomenon. Current spatial outlier detection methods seldom consider the local

Table 5 Information of the spatial outliers

Spatial outlier			Neighbors			Spatial outlier			Neighbors		
ID	Land use type	Elevation/m	ID	Land use type	Elevation/m	ID	Land use type	Elevation/m	ID	Land use type	Elevation/m
71	Vegetable field	36.47	69	Vegetable field	89.74	97	Vegetable field	47.94	24	Paddy field	48.36
			70	Paddy field	132.33				18	Vegetable field	264.12
			99	Vegetable field	25.11				21	Paddy field	133.00
			61	Litchi field	28.33				20	Corn field	96.54
			63	Litchi field	36.11				36	Litchi field	15.21
			62	Vegetable field	62.00				17	Vegetable field	61.46
			65	Paddy field	41.85	70	Vegetable field	132.33	99	Vegetable field	25.11
22	Vegetable field	83.52	20	Corn field	96.54				69	Vegetable field	89.74
			75	Vegetable field	100.25				71	Vegetable field	36.47
			21	Paddy field	133.00				77	Vegetable field	20.36
			23	Paddy field	124.41	98	Vegetable field	64.46	76	Vegetable field	144.35
1	Paddy field	79.33	6	Paddy field	81.33				74	Vegetable field	100.51
			95	Vegetable field	124.00				75	Vegetable field	100.25
			2	Vegetable field	95.67				25	Paddy field	13.43
			41	Paddy field	45.01				12	Paddy field	25.34
40	Vegetable field	103.28	3	Vegetable field	153.78				26	Vegetable field	20.56
			4	Paddy field	137.65				31	Vegetable field	48.66
			41	Paddy field	45.01				32	Paddy field	58.64
90	Paddy field	290.47	89	Vegetable field	52.55	69	Vegetable field	89.74	70	Vegetable field	132.33
			91	Paddy field	188.33				71	Vegetable field	36.47
			92	Paddy field	151.67				77	Vegetable field	20.36
			93	Paddy field	257.68				78	Paddy field	277.39
48	Litchi field	56.34	44	Vegetable field	44.33				61	Litchi field	28.33
			43	Paddy field	80.32				104	Vegetable field	77.24
			45	Vegetable field	36.32	86	Vegetable field	147.79	82	Paddy field	147.25
			46	Paddy field	57.00				83	Paddy field	163.88
			47	Vegetable field	17.22				84	Paddy field	95.70
			42	Paddy field	23.33				85	Vegetable field	28.66
42	Paddy field	23.33	44	Vegetable field	44.33	55	Litchi field	45.31	53	Paddy field	34.53
			43	Paddy field	80.32				54	Vegetable field	15.69
			48	Litchi field	56.34				57	Vegetable field	10.10
16	Paddy field	84.05	27	Paddy field	71.48	51	Vegetable field	27.25	50	Vegetable field	13.14
			28	Paddy field	41.04				53	Paddy field	34.53
			15	Vegetable field	74.57						

Table 6 Analysis of the spatial outliers

Spatial outlier	Main causes	Spatial outlier	Main causes	Spatial outlier	Main causes
97	Land use type, elevation	1	elevation, pollution source	90	elevation
71	Land use type, elevation, pollution source	55	Land use type, pollution source	69	elevation
22	Land use type, elevation	51	pollution source	98	elevation
86	Land use type, elevation	40	elevation	70	elevation
42	Land use type, elevation, pollution source	16	elevation	48	Land use type

spatial correlation and the global heterogeneous characters. In this paper, we first employ the spatial clustering technology to discover the local spatial correlation patterns, and then to detect outliers in all clusters, respectively. A robust method of spatial outlier measure is also proposed. The soil heavy metal monitoring dataset has been used to prove the advantage of our method compared to the *SOM* method.

Future works will be focus on the following two directions: (1) To develop multi-scale spatial clustering and spatial outlier detection methods. In this paper, we detect outliers only on a large scale. Detecting multi-scale spatial outliers is meaningful in many practical applications. Multi-scale spatial clustering may be a useful tool to detect multi-scale outliers. (2) To consider multiple non-spatial attribute correlation when the spatial clustering and outlier detection procedure work. Chen *et al*., (2008) employed the Mahalanobis distance to detect multiple non-spatial attribute outliers, but the method may not be reliable for the computation of the covariance and average value, and has a higher time complexity (Fan & Pan, 1980). The spatial outlier detection method involing the local spatial correlation, heterogeneous and multiple non-spatial attribute correlation characters is the work under consideration.

REFERENCES

Breunig M, Kriegel H, Ng R T and Sander J. 2000. LOF: identifying density-based local outliers. Proceedings of the ACM SIGMOD conf. On Management of Data'2000, Dallas, TX

Chawla S and Sun P. 2006. SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems*, **9**(4): 412—429

Chen D C, Lu C T, Kou Y F and Chen F. 2008. On detection of spatial outliers. *Geoinformatica*, **12**(4): 455—475

De Smith M J, Goodchild M F and Longley P A. 2007. Geospatial analysis: a comprehensive guide to principles, techniques and software tools, second edition. UK: The Winchelsea Press

Ester M, Kriegel H P, Sander J and Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd the International Conference on Knowledge Discovery and Data Mining. Portland, OR

Fang K T and Pan E P. 1982. Clustering Analysis. Beijing: Geological Publishing House

Han J and Kamber M. 2005. Data Mining: Concepts and Techniques, Second Edition. San Francisco: Morgan Kaufmann.

Haslett J, Brandley R, Craig P, Unwin A and Wills G. 1991. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, **45**(3): 234—242

Hawkins D. 1980. Identification of Outliers. London: Chapman and Hall

Huang T Q, Qin X L and Wang Q M. 2006. New method of spatial outliers measurement and detection in spatial databases. *Journal of Image and Graphics*, **11**(7): 982—989

Li D R, Wang S L, Li D Y and Wang X Z. 2002. Theories and technologies of spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, **27**(3): 221—233

Li G Q, Deng M, Cheng T and Zhu J J. 2008. A dual distance based spatial clustering method. *Acta Geodaetica et Cartographica Sinica*, **37**(4): 482—487

Li G Q, Deng M, Zhu J J, Cheng T and Liu Q L. 2009a. Spatial outlier detection considering distances among their neighbors. *Journal of Remote Sensing*, **13**(2): 197—202

Li G Q, Deng M, Liu Q L and Cheng T. 2009b. A spatial clustering method adaptive to local density change. *Acta Geodaetica et Cartographica Sinica*, **38**(3): 255—263

Liu H G, Jezek K C and O'Kelly M E. 2001. Detecting outliers in irregularly distribution spatial data sets by locally adaptive and robust statistics analysis in GIS. *International Journal of Geographical Information Science*, **15**(8): 721—741

Ma R H and He Z Y. 2006. Fast mining of spatial outliers from GIS database. *Geomatics and Information Science of Wuhan University*, **31**(8): 679—682

Mccullagh M J and Ross C G. 1980. Delaunay triangulation of a random data set for isarithmic mapping. *The Cartographic Journal*, **17**: 93—99

Pei T, Zhou C H, Luo J C, Han Z J, Wang M, Qin C Z and Cai Q. 2001. Review on the proceedings of spatial data mining research. *Journal of Image and Graphics*, **6**(9): 854—860

Shekhar S, Lu C T and Zhang P S. 2003. A unified method to detecting spatial outliers. *Geoinformatica*, **7**(2): 139—166

Shekhar S, Lu C T and Zhang P S. 2001. Detecting graph-based spatial outliers: algorithms and applications. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California

Tobler W. 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, **46**(2): 234—240

Zheng M Q, Chen C C, Fan M H, Ye D Y and Lin J X. 2008. Spatial outlier detection based on delaunay triangulation. *Microcomputer Applications*, **29**(6): 76—82

采用聚类技术探测空间异常

邓　敏, 刘启亮, 李光强

中南大学 测绘与国土信息工程系, 湖南 长沙 410083

摘　要: 提出了一种基于聚类的空间异常探测方法。该方法通过空间聚类获得局部相关性较强的实体集合, 分别探测空间异常, 给出了一种稳健的空间异常度量指标, 提高了异常探测结果的可靠性。通过实例验证以及与 SOM 方法的比较分析, 证明了该方法的正确性和优越性。

关键词: 空间异常探测, 空间聚类, 空间异常度量, 空间数据挖掘

中图分类号: TP751.1　　文献标志码: A

引用格式: 邓　敏, 刘启亮, 李光强. 2010. 采用聚类技术探测空间异常. 遥感学报, **14**(5): 944—958

Deng M, Liu Q L, Li G Q. 2010. Spatial outlier detection method based on spatial clustering. *Journal of Remote Sensing*. **14**(5): 944—958

1 引　言

空间异常探测是空间数据挖掘领域的一项重要研究内容(裴韬等, 2001; 李德仁等, 2002), 旨在海量空间数据中发现小部分偏离普遍模式的空间实体。通常, 这部分异常实体蕴含了意想不到的知识。在地球信息科学中, 空间异常可能代表着地理现象或地理过程的特殊发展规律。近年来, 空间异常探测在地质灾害监测、成矿预测、环境监测与保护、遥感图像数据处理等领域广受关注, 并具有重要应用价值。

空间异常是指非空间属性与其空间邻近域内的其他参考实体的非空间属性显著不同的空间实体(Shekhar 等, 2003)。空间异常探测通常是根据空间属性(即位置)确定空间邻近关系, 通过非空间属性描述异常程度。这里的非空间属性指的是与空间实体对应的属性特征, 例如空间采样点的重金属含量。现有的空间异常探测方法大致分为: (1)基于统计的方法(Hawkins, 1980); (2)基于图形的方法(Haslett 等, 1991); (3)基于距离的方法(Shekhar 等, 2001; Chen 等, 2008); (4)基于局部度量的方法(Breunig 等, 2000; Liu 等, 2001; Ma & He, 2006; Chawla & Sun, 2006; 黄添强等, 2006; 郑旻琦等, 2008; 李光强等, 2009a); (5)基于聚类的方法(李光强等, 2008, 2009b)。基于统计方法要求数据服从一定的统计分布, 适用性不强。基于图形的方法主要是采用可视化方法(如变量云与散点图)寻找异常实体。由于此方法存在诸多缺点, 现已较少使用(黄添强等, 2006)。基于距离的方法和基于局部度量的方法主要是考虑空间实体与其空间邻近实体间非空间属性的偏离程度, 而没有顾及空间实体间的局部相关特征, 且异常度量的结果不够稳健。基于聚类的方法将聚类后未归入任何空间簇的对象视为空间异常, 其主要目的在于发现空间簇, 缺乏对空间异常的准确度量, 是一种粗糙的异常探测方法。为此, 本文将基于聚类的异常探测方法与基于局部度量的方法相融合, 发展了一种更为稳健的空间异常度量, 提出一种基于聚类的空间异常探测方法。

2 相关工作、存在的问题及研究策略

2.1 主要相关工作及存在的问题

空间异常探测过程可以分为 3 个过程: (1)空间

收稿日期: 2009-08-25; 修订日期: 2010-03-19

基金项目: 国家 863 计划项目(编号: 2009AA12Z206); 地理空间信息工程国家测绘局重点实验室开放基金重点项目(编号: 200805); 江苏省资源环境信息工程重点实验室(中国矿业大学)开放基金项目(编号: 20080101)和中南大学研究生学位论文创新资助项目(编号: 713360010)。

第一作者简介: 邓敏(1974—), 男, 江西临川人, 博士, 教授, 博士生导师, 主要研究方向为时空数据挖掘、推理与分析, 发表论文 90 余篇。E-mail: dengmin208@tom.com。

邻近关系确定; (2)空间异常度量; (3)空间异常识别。Shekhar 等(2001)、Chen 等(2008)首先根据实体间的邻接关系(或构建 K-NN 邻近关系)获得空间邻域，根据空间实体与其邻近实体非空间属性平均值(或中值)的差异定义实体的空间异常程度，最后通过统计测试获得异常实体集合。此类方法只能适合发现全局的异常现象，容易忽略局部空间异常(Chawla & Sun, 2006)。Breunig 等(2000)、Chawla & Sun(2006)、黄添强等(2006)借助局部密度的思想定义局部偏离度的概念，局部偏离度较大的若干实体判为空间异常，此类方法受空间邻近域参数选择(郑旻琦等，2008)及空间邻近域内异常实体的影响显著。Liu 等(2001)、郑旻琦等(2008)、李光强等(2009a)采用局部插值非空间属性值与实测非空间属性值的差异定义空间实体局部偏离程度，在整体上探测异常，没有顾及空间实体的局部相关特征，亦容易受邻近域选择及邻近域内空间实体的影响。于是，可以概括当前空间异常探测方法存在的主要问题为两个方面:

(1) 现有方法认为所有空间实体之间均具有同等的相关性，而实际上空间实体间只具有局部的相关性，在整体上更呈现为异质特性(De Smith 等，2007)。例如在一个大区域内，经常出现非空间属性差异比较明显的几个子区域，若不加以区分，则一方面极有可能导致空间邻近域内实体不满足相关性假设，导致异常度量的偏差；另一方面可能由于局部出现过多的异常现象而导致一些在整体上不明显的空间异常难以被发现。下面以 Chawla & Sun(2006)采用的模拟数据集为例，图 1(a)为实体非空间属性及其空间分布情况，图 1(b)为相应的局部异常度量(SLOM)值，根据 SLOM 方法共探测出 5 个空间异常，如斜线位置标识。然而，在区域 II 和 III 中两个阴影位置的 SLOM 值明显偏大，极有可能为空间异常。但是，由于 I 区域内出现较多明显的异常点，导致这两个异常无法被发现。于是，可以直观地假设，I 区域内由于污染扩散导致过多异常现象，而 II、III 区域未受污染影响，依据地理学第一定律中越近越相关的原则(Tobler, 1970), I、II、III 各自区域内实体间的相关性较强，而不同区域内实体间相关性较弱，甚至是独立的，如果不加以区分则极有可能忽略 II、III 区域内的异常现象，严重影响空间异常探测的应用效果。

(2) 缺乏稳健的空间异常度量方法。现有空间邻近域(K-NN 邻域或 ε 邻域)需要过多的人为干预，度量结果受输入参数的影响较大。在计算空间异常程度时通常将空间邻近域内实体等同看待，而没有考虑空间实体间的空间相关性。此外，空间异常度量时邻近域内实体必须满足局部平稳性假设，实体空间邻近域内若包含了空间异常实体，则该实体的空间异常度量会受到明显影响。由于当前空间异常度量方法没有综合考虑这些方面的问题，故导致空间异常探测结果的准确性不高，甚至可能是错误的。

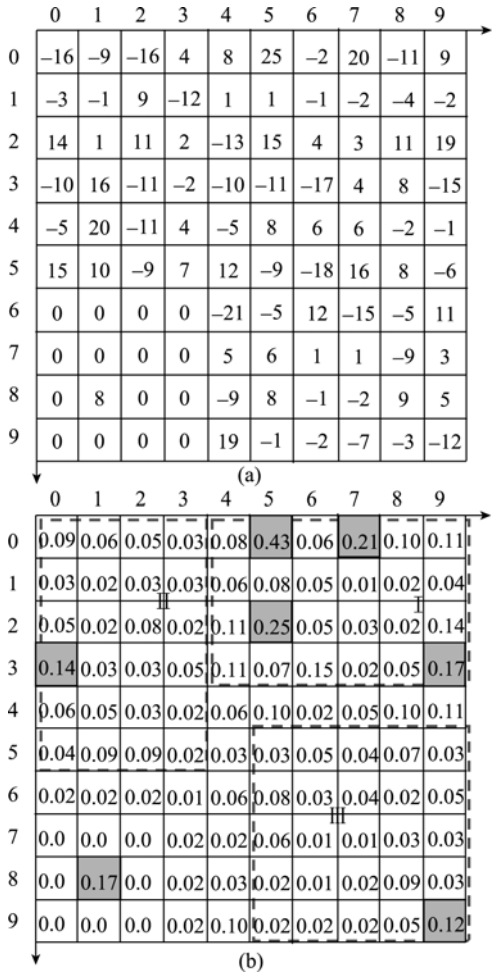

图 1 模拟空间数据集(Chawla & Sun, 2006)
(a) 非空间属性空间分布; (b) 空间异常度量结果

2.2 本文的研究策略

针对上述问题，本文将空间聚类技术与局部异常度量方法相结合，发展了一种基于聚类的空间异常探测方法，具体思路如下:

空间聚类是将空间数据库中的空间实体划分成具有一定意义的若干簇，使得每个簇内实体间具有最大相似度，而簇间实体差别最大，不同空间簇的实体在空间上具有明显的区分。空间聚类可以视为依据空间相关性对空间实体进行划分的过程，同一空间簇的实体其空间相关程度较高，不同簇之间的实体异质性更明显。于是，空间簇可以视为一种局

部相关性显著的空间分布模式。因此，可以从空间聚类的角度更好地描述空间异常，它是指在空间聚类获得的空间簇中非空间属性严重偏离其邻近实体的空间实体。由于空间簇内各实体之间具有较强的相关性，因而空间异常探测的结果更为符合实际。

在空间聚类的基础上，采用 Delaunay 三角网构建实体间的空间邻近关系，并对空间邻近域内可能存在的异常实体依据非空间属性变化梯度进行预先修复，进而采用反距离插值方法计算每个实体非空间属性理论预测值，根据理论预测值和实际值的差异来计算空间异常度。

综上，基于聚类的空间异常探测方法可以分为三个步骤：(1)空间聚类；(2)在空间聚类获得的各个空间簇内分别探测空间异常；(3)空间异常评价与分析。

3　基于聚类的空间异常探测方法

3.1　空间聚类

当前的空间聚类方法主要包括：基于划分的方法、基于层次的方法、基于密度的方法以及基于格网的方法。空间数据分布具有明显的空间分异特性，空间聚类算法必须能够适应局部空间密度的变化。因此，本文采用了一种适应局部密度变化的空间聚类方法—ADBSC(李光强等，2009b)进行空间聚类。ADBSC 算法采用 K-NN 最大距离来反映实体的局部空间密度，进而将局部空间密度相似的实体聚为一类，可以很好地适应空间数据分布不均匀的特性。ADBSC 算法需要两个输入参数，即邻近实体数量 k 和密度变化率阈值 ε。本文在进行空间异常探测之前，首先采用 ADBSC 算法对空间数据进行聚类操作，将所有空间实体划分为若干空间相关性较强的空间簇。在空间聚类过程中，参数选择以及具体空间簇的数量确定仿照 Ester 等(1996)采用了一种试凑法、启发式的策略，通过统计空间实体 k 邻近最大距离的变化特征进行设置。同时，为了便于空间异常探测过程中邻近域的构建，要求尽量将所有空间实体加入空间簇中。

3.2　稳健的空间异常度量

首先通过空间聚类获得各个空间簇，并借助 Delaunay 三角网来构建空间实体间的邻近关系。这种策略一方面充分利用了 Delaunay 三角网能够自然地反映空间实体间的邻近关系的优点(Mccullagh & Ross, 1980)；另一方面，借助了空间聚类对实体间距离的约束作用，可以有效克服传统采用 Delaunay 三角网构建空间邻近域在边界处以及空间分布不均匀区域的误差(郑旻琦等, 2008; 李光强等, 2009a)。图 2(a)为所有空间数据依据 Delaunay 三角网构建的空间邻近关系；图 2(b)为经过 ADBSC 算法聚类后(k=4, ε=22%)构建的空间邻近关系。通过比较发现，由于空间数据分布不均匀，构建空间邻近域时在边界处存在明显误差(图 2(a))，而聚类后针对每个空间簇构建的实体空间邻近关系显然更加合理(图 2(b))。

进而，在实体的空间邻近域内对可能存在的非空间属性异常值进行修复操作。由于空间邻近实体的数量较少，故难以准确探测出异常值。此外，简单的舍弃策略一方面浪费了宝贵的空间数据(Chawla & Sun, 2006)，另一方面极有可能导致空间邻近实体在某个方向聚集，影响异常度量的可靠性(Liu 等, 2001)。为此，本文依据非空间属性变化梯度对空间邻近实体的非空间属性进行修复，具体描述如下:

定义 1　非空间属性变化梯度：给定空间实体 P，其空间邻近实体记为 $\boldsymbol{N}(P)=\{X_1, X_2, \cdots, X_n\}$，$f(X_i)$表示实体 X_i 的非空间属性值，因此非空间属性变化梯

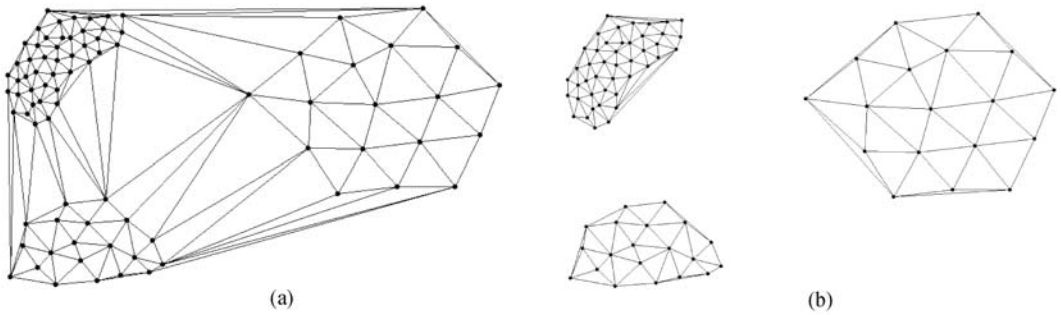

图 2　空间邻近域构建
(a) Delaunay 三角网构建空间邻近关系; (b) 空间聚类后构建空间邻近关系

度定义为空间实体 P 与其某一邻近域实体 X_i 的非空间属性差值的绝对值与二者间距离的比值，记为 $G(P, X_i)$，表达为:

$$G(P,X_i)=\frac{|f(P)-f(X_i)|}{D(P,X_i)} \quad \forall X_i\in \boldsymbol{N}(P) \qquad (1)$$

式中: $D(P, X_i)$ 表示实体 P 与 X_i 间的欧氏距离。

进而，针对任一空间实体 P，其邻域异常值修复过程如下:

① 令 $f(P)=0$，分别计算实体 P 与其空间邻近实体 X_i 的非空间属性变化梯度 $G(P, X_i)$，并从小到大进行排列为 $G_{(1)}, G_{(2)}, \cdots, G_{(n)}$，记为序列 $\boldsymbol{G}(P)$，同时获得 $\boldsymbol{G}(P)$的中位数统计量 $M(P)$。

② 针对任一空间邻近实体 X_i，计算非空间属性变化梯度偏离 $GD(X_i)$, $GD(X_i)=|G(P, X_i)-M(P)|$，从小到大进行排列为 $GD_{(1)}, GD_{(2)}, \cdots, GD_{(n)}$，记为序列 $\boldsymbol{GD}(P)$。

③ 依据非空间属性变化梯度偏离将邻近实体划分为大、中、小 3 个等级，处于最大等级的$[(n+1)/3]$个实体组成待修复集合 $\boldsymbol{R}(P)$。进而，采用非空间属性变化梯度序列中位数值进行修复，$f_R(X_i)$表示修复值:

$$f_R(X_i)=M(P)D(P,X_i) \quad \forall X_i\in \boldsymbol{R}(P) \qquad (2)$$

由式(2)可知，根据非空间属性变化梯度进行异常修复的方法一方面顾及了实体间的相关性(即距离变异)，另一方面采用中位数度量非空间属性的中心趋势，比平均值更加稳健。同时，本文的异常值修复策略旨在消除邻近域内可能存在的异常值对空间异常度量的影响，保证邻近域内实体间非空间属性的局部平稳性假设；但是这种修复只是暂时性的，即仅在空间异常度量的过程中进行，并不改变实体的固有非空间属性，故不会影响修复实体的异常度量。在此基础上，可以进一步计算一个实体的空间异常程度。

对于任一空间实体 P，首先采用其经过修复后的空间邻近实体通过反距离插值获得实体 P 的非空间属性预测值，进一步根据预测值与非空间属性实测值的差异来衡量实体 P 的异常程度，具体定义如下:

定义 2 稳健空间异常度：给定空间实体 P，其稳健异常度是指 P 的非空间属性值与其邻近域内空间实体的非空间属性的反距离插值数值的差异，记为 $RSOM(P)$，表达为:

$$\begin{aligned} RSOM(P)&=\left|f(P)-\frac{\sum_{i=1}^{n}\frac{f(X_i)}{D(P,X_i)}}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}}\right| \\ &=\frac{\sum_{i=1}^{n}\frac{|f(p)-f(X_i)|}{D(P,X_i)}}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}} \\ &=\frac{\sum_{i=1}^{n}G(P,X_i)}{\sum_{i=1}^{n}\frac{1}{D(P,X_i)}} \quad \forall X_i\in \boldsymbol{N}(P) \end{aligned} \qquad (3)$$

定义 3 空间异常：给定空间异常探测空间实体集 $\boldsymbol{SDB}=\{X_1, X_2, X_3, \ldots, X_n\}$，其各自稳健空间异常度组成集合 $\boldsymbol{S}_{RSOM}=\{RSOM(X_1), RSOM(X_2),\cdots, RSOM(X_n)\}$，稳健异常度均值记为 μ，标准差记为 σ，进而空间异常集合记为 $\boldsymbol{S}_{\text{outlier}}$，表达为:

$$\boldsymbol{S}_{\text{outlier}}=\{X_i \mid RSOM(X_i)-\mu>2\sigma, X_i\in \boldsymbol{SDB}\} \qquad (4)$$

由于空间聚类后，异常探测的样本减少，直接采用式(4)可能降低异常探测的稳健性。为了提高异常探测的准确性，针对小样本情况(样本数小于 30)，本文进一步给出了 μ 和 σ 的稳健估值，即 $\mu= Median(\boldsymbol{S}_{RSOM})$, $Median(\boldsymbol{S}_{RSOM})$为 $\boldsymbol{S}_{RSOM}$ 的中位数统计量；$\sigma= MAD(\boldsymbol{S}_{RSOM})$, $MAD(\boldsymbol{S}_{RSOM})$为集合 $\boldsymbol{S}_{RSOM}$ 的中位数绝对偏差，表达为:

$$\begin{aligned} MAD(S_{RSOM})=Median\{&|RAOM(X_1)-Median(S_{RSOM})|, \\ &|RAOM(X_2)-Median(S_{RSOM})|,\ldots, \\ &|RAOM(X_n)-Median(S_{RSOM})|\} \end{aligned} \qquad (5)$$

由于本文在进行空间异常度量时充分顾及了空间邻近域的异常实体影响以及实体与邻近域内实体间的相关性，异常度量的结果将更加可靠。下面通过一个算例来说明本文稳健空间异常度量方法的有效性，同时与 *SOM* 方法(李光强等，2009a)的计算结果进行比较，其中 *SOM* 方法直接采用实体实测非空间属性值与反距离插值数值之差作为异常度量的指标。图 3 中 Delaunay 三角网的每个节点表示一个空间实体，数字表示实体的非空间属性，从中可以发现 B 点的非空间属性值明显偏离其他实体，显然是一个空间异常，而 A 点应是一个正常实体。由表 1 中异常度量结果可以发现，由于 A 点空间邻近域内包含了 B 点，结果导致 A 点的异常度(*SOM*)显著增大，极有可能被误判为一个局部的空间异常；而本文提出的稳健异常度量(*RSOM*)方法，在对 A 点进行异常

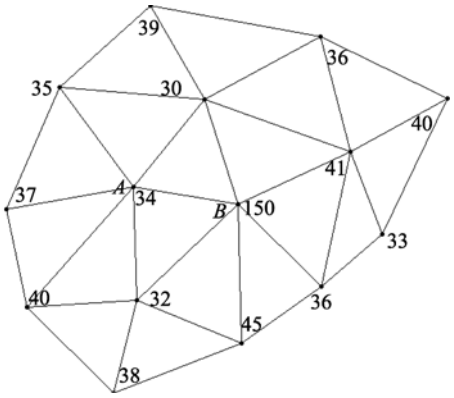

图 3　模拟算例

表 1　空间异常度量值

点号	观测值	插值结果		异常度	
		SOM	*RSOM*	*SOM*	*RSOM*
A	34	56.59	34.40	22.59	0.4
B	150	36.14	38.76	113.86	111.24

度量时，有效地对邻域内的异常值 *B* 进行了修复，故异常度量的结果更加合理，而在此过程中并没有改变异常点 *B* 的固有非空间属性值，因此 *B* 点依然可以被有效识别。

实际中，空间实体的非空间属性可能为多个，在这种情况下首先需要将多个非空间属性分别进行归一化，并计算各属性的偏离度，进而获得非空间属性整体偏离度。对于包含 *n* 个实体的空间数据库 ***SDB***，非空间属性维数为 *d*，ADBSC 算法的复杂度约为 $O(nlogn)$；构建 Delaney 三角网并获得邻近关系的复杂度约为 $O(nlogn)+O(6n)$；非空间属性归一化的复杂度为 $O(dn)$；异常度计算的复杂度约为 $O(2nlogn)+O(6dn)$；排序的复杂度约为 $O(nlogn)$；于是，空间异常探测的复杂度计算为：$O(nlogn)+O(nlogn)+O(6n)+O(dn)+O(2nlogn)+O(6dn)+O(nlogn)$，当 $d \ll n$ 时，算法复杂度近似为 $O(nlogn)$。

4　实例分析

本文采用中国华南某市土壤重金属 Cr 浓度监测数据验证本文算法的可行性和正确性，土壤采样点共 104 个，其空间分布如图 4。将本文提出方法与顾及邻近域内实体间距离的异常探测方法——*SOM* 进行了比较。*SOM* 方法针对整个数据集采用 Delaunay 三角网构建空间邻近关系，采用每个实体的 *SOM* 值度量其异常程度，最终选取 *SOM* 值最大 *m* 个实体构成空间异常集合。

依据基于聚类的空间异常探测的基本步骤，首先进行空间聚类。图 5 为 ADBSC 算法在 k=5 和 ε=23%时的空间聚类结果，此时所有空间实体均加入空间簇中，共获得 10 个空间簇。针对各个空间簇构建的空间邻近关系如图 5(c)。当空间簇内实体数量小于 6 时，本文认为簇内任一实体与其他实体互为邻近实体。进而采用本文提出的稳健空间异常度量方法在各个空间簇中探测空间异常。空间异常探测结果及相应的 *RSOM* 值列于表 2，而空间异常点的空间分布如图 6。

同时给出 *SOM* 方法的探测结果，所有实体的 *SOM* 值从大到小进行排序，如表 3。选取 *SOM* 值最大的 5 个实体组成空间异常集合，即 S_{outliers}={97, 71, 22, 70, 40}。对比本文的探测结果与 *SOM* 方法的探测结果可以发现，*SOM* 方法仅能发现在整体上异常程度较大的异常现象，无法发现局部的异常现象,如簇 9 中 90 点和簇 10 中 86 号点，在整体上异常程度不明显，然而在局部上其非空间属性严重偏离其他实体(表 4)，表现为局部的异常现象。同时发现，本文采用的小样本异常识别策略针对小样本的情况亦十分稳健(簇 9、10 中均仅包含 5 个实体)。

图 4　采样点空间分布

表 2　基于聚类的空间异常探测结果

空间簇	测点编号	*RSOM*	空间簇	测点编号	*RSOM*	空间簇	测点编号	*RSOM*
2	1	43.04	4	51	25.57	6	16	39.05
	40	36.98		55	25.80		98	35.18
3	42	31.10	5	22	52.40	7	69	46.79
	48	30.75		97	61.65		70	47.68
9	90	31.30	10	86	20.78		71	59.62

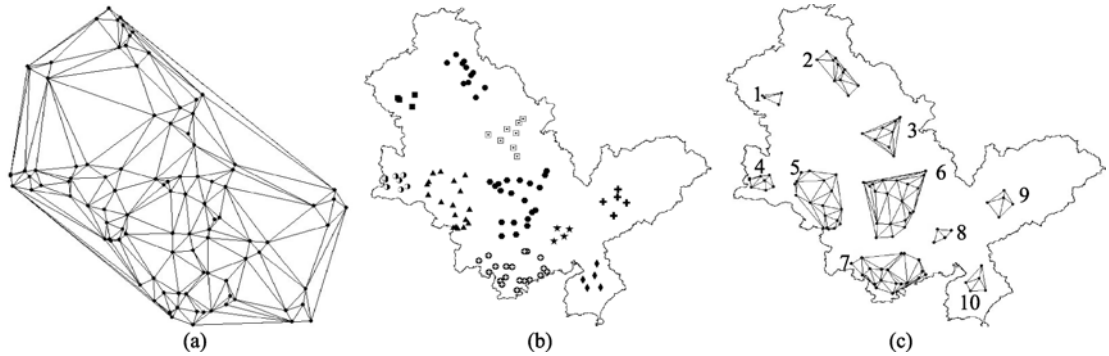

图 5 空间聚类结果及空间邻近域构建
(a) Delaunay 三角网; (b) 空间聚类结果; (c) 空间邻近关系构建

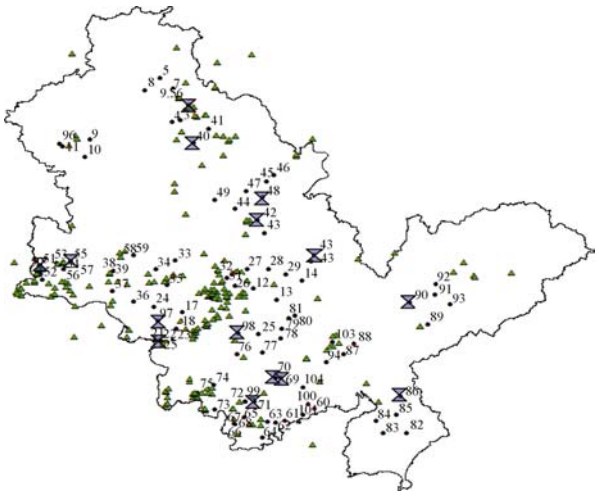

图 6 空间异常点位置
(▲—主要污染源⧗—空间异常)

表 3 SOM 方法的异常探测结果

测点编号	*SOM*
97	60.67
71	59.43
22	52.66
70	47.16
40	45.12
34	43.21
1	41.63
69	39.67
35	32.88
…	…

表 4 空间簇 9 和 10 中实体非空间属性值

空间簇	测点编号	Cr 浓度实测值
9	89	19.26
	90	2.88
	91	29.34
	92	22.88
	93	9.47
10	82	20.63
	83	12.05
	84	28.35
	85	24.58
	86	3.9

从土壤使用类型、高程差异以及污染源等 3 方面分析空间异常产生的原因。这里将重金属污染企业(如电镀厂)视为主要污染源，其空间分布如图 6。空间异常点及其邻近域结点的土壤使用类型及高程信息列于表 5。

结合以上信息，对空间异常产生的主要原因进行分析，结果列于表 6。可以发现以下几条规律: (1)空间实体与其邻近域内实体的高程差异是产生空间异常的最主要因素，对土壤重金属含量的分布具有明显的影响; (2)71、1、55、51、42 号点与污染源的联系比较紧密，极有可能是受到附近污染企业的影响; (3)绝大多数空间异常产生在水稻土和菜地，可能与农药、化肥等的过量使用具有密切关系; (4)水稻土和菜地邻接的情况下异常情况较多，可能包含了潜在的空间关联规则，需要进一步实地检测和查证。

5 结论与展望

空间异常探测对于揭示地理现象变化、发展的特殊规律具有重要意义，已成为空间数据挖掘领域的一个研究热点。针对现有空间异常探测方法没有顾及空间实体局部相关、整体分异的特征，本文首先采用空间聚类技术发现空间实体的局部相关模式，即同一空间簇中实体具有最大的相关度，而不同空间簇中实体的独立性更为显著。进而，发展了一种稳健的空间异常度量方法，在每个空间簇中探测空间异常，通过实际算例分析以及与 *SOM* 方法的比较可以发现，本文的方法主要有两个方面的优势: (1) 顾及了实体与空间邻近域内实体间距离及邻近域内空间异常点的影响，与现有的空间异常度量方法相比

表 5　空间异常点信息

异常点			邻近域			异常点			邻近域		
编号	土壤类型	高程/m	编号	土壤类型	高程/m	编号	土壤类型	高程/m	编号	土壤类型	高程/m
71	菜地	36.47	69	菜地	89.74	97	菜地	47.94	24	水稻土	48.36
			70	水稻土	132.33				18	菜地	264.12
			99	菜地	25.11				21	水稻土	133.00
			61	荔枝地	28.33				20	玉米地	96.54
			63	荔枝地	36.11				36	荔枝地	15.21
			62	菜地	62.00				17	菜地	61.46
			65	水稻土	41.85	70	菜地	132.33	99	菜地	25.11
22	菜地	83.52	20	玉米地	96.54				69	菜地	89.74
			75	菜地	100.25				71	菜地	36.47
			21	水稻土	133.00				77	菜地	20.36
			23	水稻土	124.41	98	菜地	64.46	76	菜地	144.35
1	水稻土	79.33	6	水稻土	81.33				74	菜地	100.51
			95	菜地	124.00				75	菜地	100.25
			2	菜地	95.67				25	水稻土	13.43
			41	水稻土	45.01				12	水稻土	25.34
40	菜地	103.28	3	菜地	153.78				26	菜地	20.56
			4	水稻土	137.65				31	菜地	48.66
			41	水稻土	45.01				32	水稻土	58.64
90	水稻土	290.47	89	菜地	52.55	69	菜地	89.74	70	菜地	132.33
			91	水稻土	188.33				71	菜地	36.47
			92	水稻土	151.67				77	菜地	20.36
			93	水稻土	257.68				78	水稻土	277.39
48	荔枝地	56.34	44	菜地	44.33				61	荔枝地	28.33
			43	水稻土	80.32				104	菜地	77.24
			45	菜地	36.32	86	菜地	147.79	82	水稻土	147.25
			46	水稻土	57.00				83	水稻土	163.88
			47	菜地	17.22				84	水稻土	95.70
			42	水稻土	23.33				85	菜地	28.66
42	水稻土	23.33	44	菜地	44.33	55	荔枝地	45.31	53	水稻土	34.53
			43	水稻土	80.32				54	菜地	15.69
			48	荔枝地	56.34				57	菜地	10.10
16	水稻土	84.05	27	水稻土	71.48	51	菜地	27.25	50	菜地	13.14
			28	水稻土	41.04				53	水稻土	34.53
			15	菜地	74.57						

表 6　空间异常点分析结果

异常点	异常主要原因	异常点	异常主要原因	异常点	异常主要原因
97	土壤使用类型差异, 高程差异	1	污染源, 高程差异	90	高程差异
71	土壤使用类型差异, 高程差异, 污染源	55	土壤使用类型差异, 污染源	69	高程差异
22	土壤使用类型差异, 高程差异	51	污染源	98	高程差异
86	土壤使用类型差异, 高程差异	40	高程差异	70	高程差异
42	土壤使用类型差异, 高程差异, 污染源	16	高程差异	48	土壤使用类型差异

更加稳健; (2)顾及了实体的局部相关性, 能够更全面地发现局部的异常现象。

未来的工作主要集中在 2 个方面: (1)空间聚类与空间异常探测中的尺度问题。本文方法虽然可以通过在空间聚类过程中设置不同参数(即 k 值)来获得详细程度不同的空间簇, 进而可以获得重要性程度不同的空间异常, 并在一定程度上能够反映空间聚类及异常探测的尺度特征, 但仍需要进一步研究多尺度空间聚类及多尺度空间异常探测方法, 对于更好地研究地理现象变化发展的规律具有重要意义。(2)在本文的研究基础上, 需要进一步研究顾及非空间属性及其相关性的空间聚类方法和空间异常探测方法。Chen 等(2008)采用马氏距离度量多个非空间属性间相似性的策略虽在一定程度上顾及了非空间属性间的相关性, 但是采用全部数据计算均值与协方差的方法, 一方面计算量巨大, 另一方面忽视了空间实体分布的整体分异特点, 结果未必可靠(方开泰&潘恩沛, 1982)。顾及整体分异与多个非空间属性相关的空间异常探测方法是作者正在开展的工作, 将另文进行讨论。

REFERENCES

Breunig M, Kriegel H, Ng R T and Sander J. 2000. LOF: identifying density-based local outliers. Proceedings of the ACM SIGMOD conf. On Management of Data'2000, Dallas, TX

Chawla S and Sun P. 2006. SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems*, **9**(4): 412—429

Chen D C, Lu C T, Kou Y F and Chen F. 2008. On detection of spatial outliers. *Geoinformatica*, **12**(4): 455—475

De Smith M J, Goodchild M F and Longley P A. 2007. Geospatial analysis: a comprehensive guide to principles, techniques and software tools, second edition. UK: The Winchelsea Press

Ester M, Kriegel H P, Sander J and Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd the International Conference on Knowledge Discovery and Data Mining. Portland, OR

Fang K T and Pan E P. 1982. Clustering Analysis. Beijing: Geological Publishing House

Han J and Kamber M. 2005. Data Mining: Concepts and Techniques, Second Edition. San Francisco: Morgan Kaufmann.

Haslett J, Brandley R, Craig P, Unwin A and Wills G. 1991. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, **45**(3): 234—242

Hawkins D. 1980. Identification of Outliers. London: Chapman and Hall

Huang T Q, Qin X L and Wang Q M. 2006. New method of spatial outliers measurement and detection in spatial databases. *Journal of Image and Graphics*, **11**(7): 982—989

Li D R, Wang S L, Li D Y and Wang X Z. 2002. Theories and technologies of spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, **27**(3): 221—233

Li G Q, Deng M, Cheng T and Zhu J J. 2008. A dual distance based spatial clustering method. *Acta Geodaetica et Cartographica Sinica*, **37**(4): 482—487

Li G Q, Deng M, Zhu J J, Cheng T and Liu Q L. 2009a. Spatial outlier detection considering distances among their neighbors. *Journal of Remote Sensing*, **13**(2): 197—202

Li G Q, Deng M, Liu Q L and Cheng T. 2009b. A spatial clustering method adaptive to local density change. *Acta Geodaetica et Cartographica Sinica*, **38**(3): 255—263

Liu H G, Jezek K C and O'Kelly M E. 2001. Detecting outliers in irregularly distribution spatial data sets by locally adaptive and robust statistics analysis in GIS. *International Journal of Geographical Information Science*, **15**(8): 721—741

Ma R H and He Z Y. 2006. Fast mining of spatial outliers from GIS database. *Geomatics and Information Science of Wuhan University*, **31**(8): 679—682

Mccullagh M J and Ross C G. 1980. Delaunay triangulation of a random data set for isarithmic mapping. *The Cartographic Journal*, **17**: 93—99

Pei T, Zhou C H, Luo J C, Han Z J, Wang M, Qin C Z and Cai Q. 2001. Review on the proceedings of spatial data mining research. *Journal of Image and Graphics*, **6**(9): 854—860

Shekhar S, Lu C T and Zhang P S. 2003. A unified method to detecting spatial outliers. *Geoinformatica*, **7**(2): 139—166

Shekhar S, Lu C T and Zhang P S. 2001. Detecting graph-based spatial outliers: algorithms and applications. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California

Tobler W. 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, **46**(2): 234—240

Zheng M Q, Chen C C, Fan M H, Ye D Y and Lin J X. 2008. Spatial outlier detection based on delaunay triangulation. *Microcomputer Applications*, **29**(6): 76—82

附中文参考文献

方开泰, 潘恩沛. 1982. 聚类分析. 北京: 地质出版社

黄添强, 秦小麟, 王钦敏. 2006. 空间数据库中离群点的度量与查找新方法. 中国图象图形学报, **11**(7): 982—989

李德仁, 王树良, 李德毅, 王新洲. 2002. 论空间数据挖掘和知识发现的理论和方法. 武汉大学学报(信息科学版), **27**(3): 221—233

李光强, 邓敏, 程涛, 朱建军. 2008. 一种基于双重距离的空间聚类方法. 测绘学报, **37**(4): 482—487

李光强, 邓敏, 朱建军, 程涛, 刘启亮. 2009a. 一种顾及邻近域内实体间距离的空间异常检测新方法. 遥感学报, **13**(2): 197—202

李光强, 邓敏, 刘启亮, 程涛. 2009b. 一种适应局部密度变化的空间聚类方法. 测绘学报, **38**(3): 255—263

马荣华, 何增友. 2006. 从 GIS 数据库中挖掘空间离群点的一种高效算法. 武汉大学学报(信息科学版), **31**(8): 679—682

裴韬, 周成虎, 骆剑承, 韩志军, 汪闽, 秦承志, 蔡强. 2001. 空间数据知识发现研究进展评述. 中国图象图形学报, **6**(9): 854—860

郑旻琦, 陈崇成, 樊明辉, 叶东毅, 林甲祥. 2008. 基于 Delaunay 三角网的空间离群挖掘. 微型计算机应用, **29**(6): 76—82