International Journal of Geographical Information Science
Vol. 20, No. 2, February 2006, 153–168

Taylor & Francis
Taylor & Francis Group

Research Article

A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes

TAO PEI*†, A-XING ZHU†‡, CHENGHU ZHOU†, BAOLIN LI† and CHENGZHI QIN†

†State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, 11A, Datun Road Anwai, Beijing 100101, China

‡Department of Geography, University of Wisconsin Madison, 550N, Park Street, Madison, WI 53706-1491, USA

(*Received 4 June 2004; in final form 18 August 2005*)

When two spatial point processes are overlaid, the one with the higher rate is shown as clustered points, and the other one with the lower rate is often perceived to be background. Usually, we consider the clustered points as feature and the background as noise. Revealing these point clusters allows us to further examine and understand the spatial point process. Two important aspects in discerning spatial cluster features from a set of points are the removal of noise and the determination of the number of spatial clusters. Until now, few methods were able to deal with these two aspects at the same time in an automated way. In this study, we combine the nearest-neighbour (NN) method and the concept of density-connected to address these two aspects. First, the removal of noise can be achieved using the NN method; then, the number of clusters can be determined by finding the density-connected clusters. The complexity for finding density-connected clusters is reduced in our algorithm. Since the number of clusters depends on the value of k (the kth nearest neighbour), we introduce the concept of lifetime for the number of clusters in order to measure how stable the segmentation results (or number of clusters) are. The number of clusters with the longest lifetime is considered to be the final number of clusters. Finally, a seismic example of the west part of China is used as a case study to examine the validity of our method. In this seismic case study, we discovered three seismic clusters: one as the foreshocks of the Songpan quake (M=7.2), and the other two as aftershocks related to the Kangding-Jiulong (M=6.2) quake and Daguan quake (M=7.1), respectively. Through this case study, we conclude that the approach we proposed is effective in removing noise and determining the number of feature clusters.

Keywords: Nearest-neighbour; Feature; Noise; Cluster; Spatial point process; Poisson process; Spatial data mining

1. Introduction

When two processes with different rates are overlaid by occupying different areas (support domain), the one with the higher rate is usually displayed as clustering of points, while the other with a lower rate is often considered as background (Byers

*Corresponding author. Email: peit@lreis.ac.cn

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810500399654

and Raftery 1998). In general terms, background can be considered as noise, and clustering of points can be viewed as a feature. Noise can be defined as a homogeneous Poisson point process over the entire area while the feature is treated as a Poisson process but restricted to a certain area and overlaid on the noise. Revealing these point clusters allows us to further examine and understand the spatial point process in question, which may be referred to as a 'hotspot' in some studies (Brimicombe 2003). For example, revealing patterns of spatial distribution of prostate cancer mortality helps in understanding the potential causes of cancer (Jemal *et al.* 2002), and detecting the clustering of traffic accidents could shed light on problems associated with traffic planning (Steenberghen *et al.* 2004). Furthermore, this understanding of the spatial point process facilitates the prediction of future events. For example, the clustering of earthquake events may imply the outbreak of a strong earthquake or the existence of aftershocks of a strong earthquake (Reasenberg 1999, Umino *et al.* 2002).

Two important aspects in discerning spatial cluster features from a set of points are the removal of noise and the determination of the number of spatial clusters. The second aspect is the same as determining the subjection of each point and the size of each cluster. The presence of noise often obscures the patterns of clusters by changing the size, shape, and concentration of clusters produced from spatial point processes, and makes the detection of clusters difficult. In addition, distinct clusters usually represent different independent instances of a homogeneous spatial process, and each has approximately the same rate and is within a specific area. So, determination of the number of clusters and size of each cluster will be very helpful in understanding the spatial process in the research area.

To date, several methods have been established to detect clustered features from spatial point processes in the presence of noise. Banfield and Raftery (1993) initially put forward a statistical framework for non-Gaussian clustering and a means of incorporating noise in the form of Poisson process. But their model can only be applied for the special case where points are distributed uniformly along and tightly about a line segment in linear space. In order to apply the Expectation-Maximization (EM) algorithm to the cluster method, Fraley and Raftery (1998) built a mixture model in which each statistical component corresponds to a different cluster. Although their model can deal with spatial noise and clusters in varied geometry, it cannot yet accommodate clusters in arbitrary shape especially for the nonlinear distribution. Dasgupta and Raftery (1998) utilized a statistical model-based clustering method to detect minefields in linear or piecewise linear form in the presence of noise. With their method, the shape of the spatial point feature has to be presumed prior to detection. Allard and Fraley (1997) derived a maximum-likelihood estimator for a mixture of uniform point processes using the Voronoï Polygons method. He assumed that the feature in the point set is a single connected component, and estimated the boundary of feature using the maximum-likelihood estimator. Although this algorithm can remove the noise effectively, its precondition requires that the feature must be restricted to a single connected component without holes, and the boundary of features is of a specific geometry. Byers and Raftery (1998) developed the Nearest-Neighbour (NN) method for estimating features in spatial point process by removing the clutter (noise). The NN method is implemented by decomposing the distribution of kth nearest neighbours of each point and cannot yet be employed to determine the exact number of the features (clusters). Ester *et al.* (1996) and Sander et al. (1998) put forward the method of

Density-Based Spatial Clustering of Application with Noise (DBSCAN) to discover clusters from spatial database with noise. DBSCAN can address the number of cluster features in a spatial data set, but the *Eps* (Epsilon, the parameter of DBSACN to define the neighbourhood points of a given point) can only be determined through an interactive process by examining its N-dist graph. It is quite difficult and subjective for users to find the valley point of N-dist graph in order to determine the value of *Eps*. It is a particularly difficult task when the curve appears smooth. In order to discover clusters that are deemed acceptable to the user, they may need to run the algorithm many times while setting different parameter values for each run. Without any guidance in setting these parameters, a great deal of effort and time are needed in such a trail-and-error approach (Han *et al.* 2001). To improve the efficiency of density-based clustering, Ankerst *et al.* (1999) put forward an algorithm, referred to as Ordering Points To Identify the Clustering Structure (OPTICS). OPTICS provides a graphical and interactive method to help find the cluster structure by constructing an augmented cluster-ordering of the database objects and its reachability plot with respect to *Eps* (generating distance which is used to generate the reachability plot) and *MinPts*. The reachability plot is a one-dimensional plot which can display the structure of the clusters. Although the reachability plot is less sensitive to the input parameters *Eps* and *MinPts*, the clustering result is still dependent on manual determination of *Eps′* (clustering distance which is used to separate the feature from noise) and *MinPts*. Later on, Daszykowski *et al.* (2001) proposed a DBSCAN-based modification to look for natural patterns (NP) of arbitrary shape. They removed noise by separating a prescribed percentage of data points, distributed at the tail of the frequency curve. However, their method cannot provide an effective way to determine the *Eps*.

In this paper, we present an approach to spatial data mining for finding clustering features by removing noise and detecting the number of features (clusters) from a spatial point data set in the presence of noise. We assumed that only two point processes, i.e. the clustered points and the noise points, are dominant in the point set. The NN method is employed to detect the *Eps* and features. With this method, we can deal with the clusters with arbitrary shape. In order to determine the number of spatial clusters under each k (the kth nearest neighbour), a recursive connection method is developed based on the concept of density-connected, employed in the DBSCAN. Then, the concept of lifetime is introduced to measure the stability of number of clusters, with the final number of clusters taken as the number of clusters with the longest lifetime.

This paper is organized in five sections. Section 2 reviews the theory of nearest-neighbour method for noise removal. The recursive method based on the concept of density-connected and the concept of lifetime used for determining the proper number of clusters are described in section 3. The approach is evaluated in section 4 through a case study using seismic data to find foreshocks and aftershocks. Conclusions and future work are given in section 5.

2. Theory of nearest-neighbour

2.1. *Basic theory of the distribution of the kth nearest neighbours*

Given a point p_i in the 2D Poisson point set Y, its spatial probability distribution of kth nearest distance D_k (the distance between p_i and its kth nearest neighbour) can

be acquired through seeking the probability density function (pdf) of including 0, 1, 2,…, $k-1$ points within the circle of $A(p_i,x)$, in which p_i is the centre, and x is the radius.

$$P(D_k \geqslant x) = \sum_{m=0}^{k-1} \frac{\mathrm{e}^{-\lambda\pi x^2}\left(\lambda\pi x^2\right)^m}{m!} = 1 - F_{D_k}(x) \tag{1}$$

where k is the parameter referring to the ordinal number of nearest neighbour, $F_{D_k}(x)$ is the cumulative distribution function of D_k, λ is the rate of the Poisson point set. If D_k is larger than x, there must be 0 or 1 or 2…$k-1$ points within the circle $A(p_i,x)$, and their pdf $f_{D_k}(x)$ is the derivative of $F_{D_k}(x)$:

$$f_{D_k}(x) = \frac{\mathrm{d}F_{D_k}(x)}{\mathrm{d}x} = \frac{\mathrm{e}^{-\lambda\pi x^2} 2(\lambda\pi)^k x^{2k-1}}{(k-1)!} \tag{2}$$

were λ and k are the same as those in the equation (1). This pdf can be treated as a mixture pdf of Gamma according to the definition of the pdf of Gamma, that is, $Y \sim \Gamma(k,\lambda\pi)$, where $Y=(D_k)^2$.

In our research, the noise and the feature can be represented as two superimposed Poisson processes with different rates, for example, λ_1 and λ_2. In this way, the bimodal pdf of D_k can be expressed as:

$$D_k \sim p\Gamma^{\left(\frac{1}{2}\right)}(k,\lambda_1\pi) + (1-p)\Gamma^{\left(\frac{1}{2}\right)}(k,\lambda_2\pi) \tag{3}$$

where p is the proportion coefficient, and λ_1 and λ_2 are the rates for the two processes, respectively (Byers and Raftery 1998).

2.2. *EM algorithm to evaluate the parameters*

The EM algorithm can be employed to evaluate the parameters of λ_1, λ_2, and p. A detailed discussion of the EM algorithm is beyond the scope of this paper; interested readers are referred to Celeux and Govaert (1992), Moon (1996) and Byers and Raftery (1998) for more details. A summary of the algorithm is given below.

The E-step (the Expectation step) in this context is:

$$E\left(\hat{\delta}_i^{(t+1)}\right) = \frac{\hat{p}^{(t)} f_{D_k}\left(d_i;\hat{\lambda}_1^{(t)}\right)}{\hat{p}^{(t)} f_{D_k}\left(d_i;\hat{\lambda}_1^{(t)}\right) + \left(1-\hat{p}^{(t)}\right) f_{D_k}\left(d_i;\hat{\lambda}_2^{(t)}\right)} \tag{4}$$

while the M-step (the Maximization step) is:

$$\hat{\lambda}_1^{(t+1)} = \frac{k\sum_{i=1}^n \hat{\delta}_i^{(t+1)}}{\pi\sum_{i=1}^n d_i^2 \hat{\delta}_i^{(t+1)}} \text{ and } \hat{\lambda}_2^{(t+1)} = \frac{k\sum_{i=1}^n \left(1-\hat{\delta}_i^{(t+1)}\right)}{\pi\sum_{i=1}^n d_i^2 \left(1-\hat{\delta}_i^{(t+1)}\right)}$$

$$\text{with } p^{(t+1)} = \sum_{i=1}^n \hat{\delta}_i^{(t+1)} \Big/ n \tag{5}$$

where n is the number of points, and t is the time of iteration. If we define the component with λ_1 representing the feature, then points with $\hat{\delta}_i^{(t+1)} \geqslant 0.5$ belong to features, and points with $\hat{\delta}_i^{(t+1)} < 0.5$ can be viewed as noise.

3. Determining the proper number of clusters

3.1. *Concept of density-connected*

Although we can apply the NN method to clean the data sets that may contain a large amount of noise, the number of clusters is still unknown. This problem can be solved by employing the concept of density-connected, the core concept of DBSCAN (Ester *et al.* 1996). There are two terms in respect of density-connected, i.e. the $N_{Eps}(p_i)$ and *MinPts*, which should be defined first.

Of a point set D, the $N_{Eps}(p)$ $(p \in D)$ represents the *Eps*-neighbourhood of a point p and is defined by $N_{Eps}(p) = \{q \in D | dist(p,q) \leqslant Eps\}$. The *MinPts* refers to the minimum number of points in the *Eps*-neighbourhood.

So, a point p is density-connected to a point q with respect to (wrt) *Eps* and *MinPts* if there is a chain of point $p_1, p_2, \ldots, p_n, p_1 = q, p_n = p$ such that $p_{i-1} \in N_{Eps}(p_i)$ $(i = 2,3,\ldots,n)$ and $N_{Eps}(p_i)$ $(i = 2, 3, \ldots, n-1)$ must contain at least *MinPts* points. In fact, $N_{Eps}(p_1)$ and $N_{Eps}(p_n)$ can also contain points amounting to or more than *MinPts*. A cluster is defined to be a point set M that any point $p_i \in M$ is density-connected to point $p_j \in M$ $(p_i \neq p_j)$. A point $p_i \in M$ is called a core point if $N_{Eps}(p_i) \geqslant Minpts$, otherwise, it is called a border point. Thus, noise is the set of points not belonging to any given clusters. We must note that a cluster M wrt *Eps* and *MinPts* contains at least *MinPts* points including core points and border points. For more details, please see Ester *et al.* (1996) and Sander *et al.* (1998).

3.2. *Determination of Eps*

In the algorithm of DBSCAN, the *Eps* is the key parameter to detect the number of clusters and can only be determined through a graphical-interactive way. That is, for a given k, each point in the point set is mapping to the distance from the point itself to its kth nearest neighbour (D_k). The sorted k-dist graph is constructed by sorting the points in the point set in descending order based on their D_ks. The threshold point is the first point in the first valley of the sorted k-dist graph, and the *Eps* is the distance between the threshold point to its kth nearest neighbour.

The simulated points are displayed in figure 1(a), in which the point set is made up of noise and five clusters with an arbitrary shape. The *Eps* determined using this visual method may be quite subjective due to the different criteria used by different users. This will significantly influence the determination of the number of clusters. Figure 1(b) shows the N-dist graphs of simulated data at $k = 2$ to 8. From figure 1(b), it is difficult to locate the valley and determine the value of *Eps*. Therefore, the number of clusters is difficult to determine. In fact, this method of determining the number of clusters might be viable only when the density difference between noise and cluster is dramatically large.

The subjectiveness and the impreciseness of the interactive method call for an automated method in determining the proper number of clusters. If the kth nearest neighbour of feature and noise can be transformed into the mixture distribution function, then the NN method, which is viewed as non-parameter method, can be used to define the *Eps*. The *Eps* can be calculated by applying the EM algorithm. The formula is:

$$Eps = \sqrt{\frac{\ln\frac{1-\rho}{\rho} + k\ln\frac{\lambda_2}{\lambda_1}}{\pi(\lambda_2 - \lambda_1)}} \quad \rho = (1-p)/p \tag{6}$$

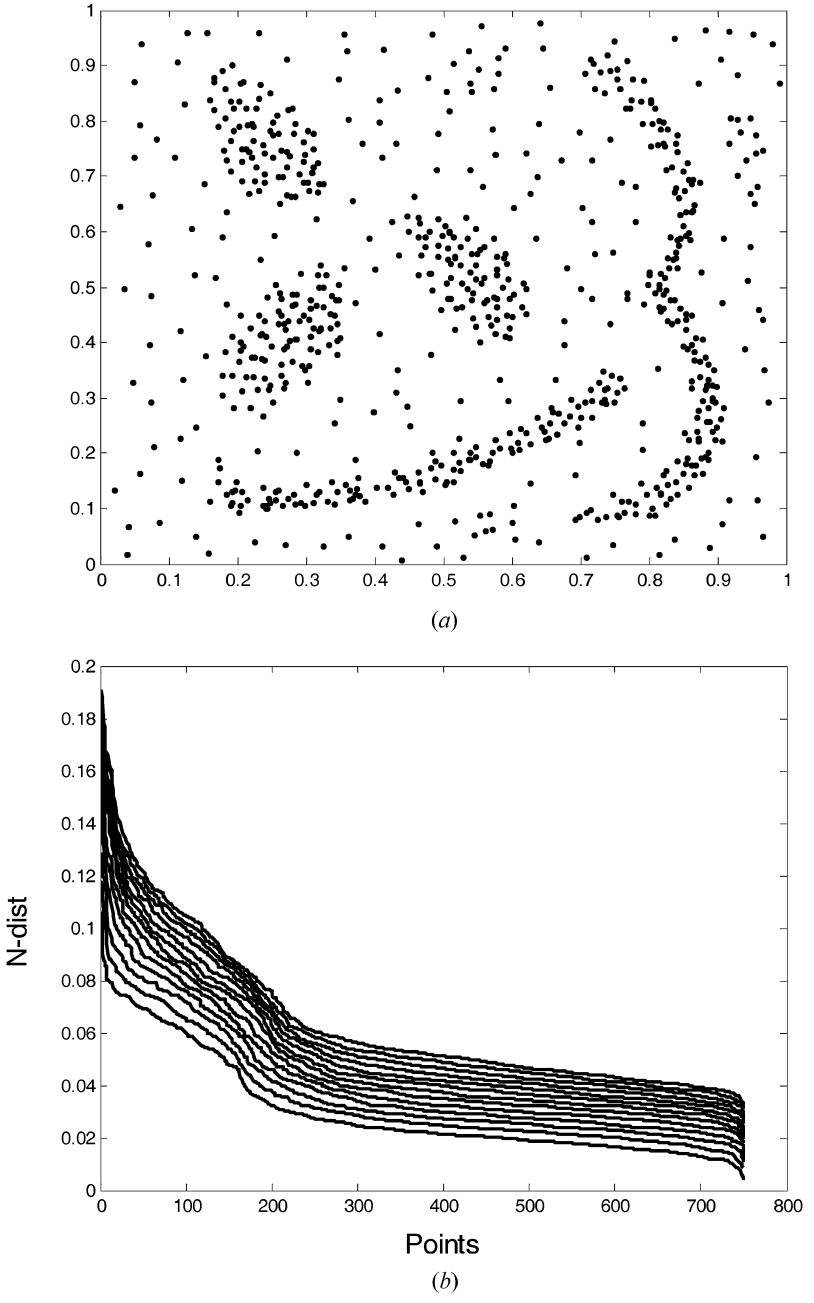


Figure 1. Simulated data and their N-dist graph. (*a*) Spatial distribution of simulated data; (*b*) N-dist graph for $k=2$ to 8. The valley which indicates the location of *Eps* cannot be easily determined from the N-dist graph.

where λ_1 and λ_2 are the rate of each component (feature and noise), respectively, p is the proportion of the first component (these parameters can be acquired from equations (4) and (5)), and k equals the value of *MinPts*.

From equation (6), we can deduce that the *Eps* depends on the choice of k and the proportion of noise and feature points. The histogram of the fifth nearest neighbour

of the simulated data (figure 1(*a*)) is displayed in figure 2. The reason we choose the fifth nearest neighbour is that the histogram from the smaller nearest neighbour cannot clearly display the bimodal characteristic. From this histogram, we clearly found a mixture histogram including two components which represent feature and noise, respectively. The EM algorithm can be applied to compute the *Eps*. The result is 0.046.

After obtaining the value of *Eps* at a given *k*, we can find the distinct clusters and thus the number of clusters at that *k* by running the following algorithm:

```
FeatureDetect(SpatialPoints, k): Boolean
[FeatureSet, NoiseSet, Eps] :=FindingFeatureByNN(SpatialPoints, k);
Id :=1;
FOR i=1 To FeatureSet.size(i) Do
  Point :=FeatureSet.get(i);//get each point from FeatureSet
    IF Point.Id==UNCLASSIFIED THEN
      IF ExpandCluster(FeatureSet,Point,Id,Eps) THEN
Id :=Id+1;
      END IF
    END IF
END FOR
FindingBorderPoints(NoiseSet, FeatureSet, Eps);
Return true;
END;//FeatureDetect
```

Function FeatureSet.get(i) returns the *i*th element in the feature points. The function of FindingFeatureByNN separates noise points from feature points using the NN method and returns the *Eps*, noise points (NoiseSet) and the feature points (FeatureSet). The returned feature points are not yet decomposed into distinct clusters. In order to decompose the feature points into distinct clusters, the function of ExpandCluster is applied, as described below.

```
ExpandCluster(FeatureSet, Point, Id, Eps, k) : Boolean;
  Seeds:=FeatureSet.regionQuery(Point,Eps);
```

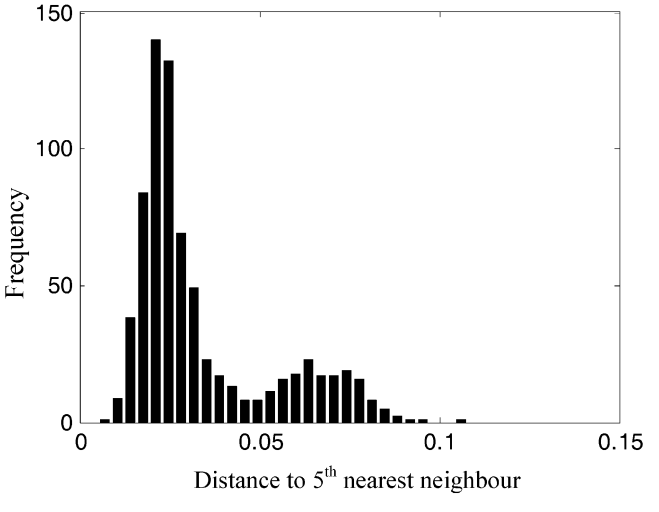


Figure 2. Histogram of the simulated data with its fifth nearest neighbour.

```
    FeatureSet.changeIds(Seeds,Id);
    Seeds.delete(Point);//delete current point of Seeds
    WHILE Seeds <> Empty DO
      CurrentP :=Seeds.first();//get the first point of Seeds
      Result :=FeatureSet.regionQuery(CurrentP,Eps);
      FOR i FROM 1 TO Result.size DO
        ResultP :=Result.get(i);//get ith point of Result
        ChangeId(ResultP,Id);//change ID of ResultP to Id
        Seeds.append(ResultP);//append ResultP to Seeds
      END FOR;
      Seeds.delete(CurrentP);//delete CurrentP from Seeds
    END WHILE;//Seeds <> Empty
    RETURN True;
  END;//ExpandCluster
```

The function of FindingBorderPoints is to find border points of each cluster, as described below. The result of the function FindingBorderPoints is to allocate *–Id* to the border points, whose IDs are different from those of core points of the same cluster.

```
  FindingBorderPoints(NoiseSet, Eps, FeatureSet) : Boolean;
    FOR i FROM 1 TO NoiseSet.size DO
      ResultP :=NoiseSet.get(i);//get ith point from NoiseSet
      Result :=FeatureSet.regionQuery(ResultP,Eps);
      IF Result.IsContainFeatureP Then
      Id :=Result.GetFeatureID();//get the ID number of feature point in Result
      ChangeId(ResultP,-Id);//change ID of ResultP to -Id
      END IF
    END FOR;
    RETURN True;
  END;//FindingBorderPoints
```

The function of the GetFeatureID return the ID of feature points in Result. If two clusters M_1 and M_2 are close enough to each other, it might be that Result contains points which belong to two different clusters. In this case, function GetFeatureID will return the ID of feature point discovered first.

The function of regionQuery can be implemented by spatial access methods, such as R-tree (Beckmann *et al.* 1990), whose runtime is $O(\log n)$. Function IsContainFeatureP is used to check if the Result contains the core points.

Apparently, the efficiency of the algorithm is determined by the efficiency of the function of regionQuery. The complexity of regionQuery is $O(\log n)$, where n is the number of points. Hence, the algorithm of DBSCAN has the runtime of $O(n * log\ n)$. In our approach, the complexity of the algorithm is reduced. This is because we have removed the noise before determining the number of clusters, that is, the original data have been divided into feature set and noise set. If we assume that the number of feature points is m_1 and the number of noise point is m_2, then the complexity of FeatureSet.regionQuery reduces to $O(log\ m_1)$. The total complexity of the algorithm (the total complexity of function ExpandCluster and FindingBorderPoints) reduces to $O(n * log\ m_1)$, where $m_1 + m_2 = n$. Therefore, it is reducing the complexity of the function regionQuery that improves the efficiency of the algorithm. The next section will discuss the influence of k on the clustering result.

3.3. *Lifetime of number of clusters*

From the analysis above, we think that the number of clusters may depend on the value of k. Figure 3 shows the clustering results with $k=5$, 9, 18. When k is small, some small features (the three small clusters in figure 3(a)) appear, which may be viewed as pseudo-features which disappear as $k>8$. When $k=9$, small features are filtered out. As k increases, some features, which should be viewed as distinctive clusters, may merge into a large one. For instance, in figure 3(c) the one on the right side and the one at the bottom are united. The effect of k on the number of clusters makes it difficult to determine the correct number of clusters in a spatial point set.

The concept of lifetime of number of clusters is introduced in this paper to assist in the determination of the correct number of clusters in a spatial point set. The lifetime of the number of clusters is defined as the period a given number of clusters exists over k, under the condition that no existing clusters disappear and no new clusters emerge during this period. That is to say, not only does the number of clusters need to be stable but also the clusters must remain the same ('alive') during the lifetime of a specific number of clusters. The number of clusters can be readily determined at different k, whereas the assessment of aliveness of a cluster may not be easily achieved.

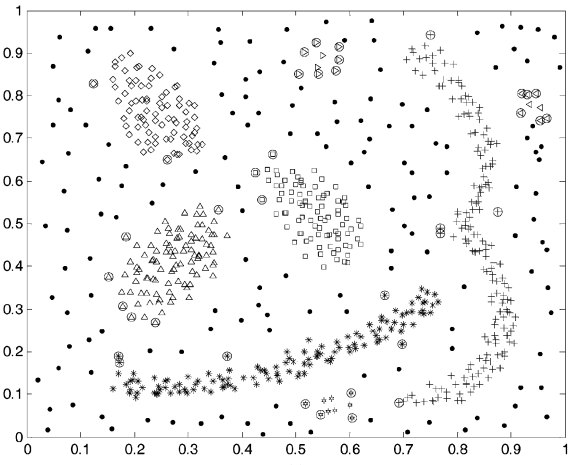
For this reason, we define aliveness as follows. When considered as 'alive', a cluster must satisfy two conditions: (1) for a given cluster (A1) at k, there must be a cluster (A2) at $k + 1$ which shares more than 50% core points of itself with A1; (2) A1 shares more than 50% core points of itself with A2. The threshold of 50% applied in the paper will be sufficient for identifying the newly emerged cluster and the disappeared cluster. In that way, A1 and A2 can be assumed as the same, though they may be different in the number of points. Hence, the lifetime of a given number of clusters can be acquired based on the condition that each cluster must be confirmed as 'alive' during this period, and this can be easily implemented by counting common core points between clusters at different k.

The concept of lifetime can be used to measure the stable degree of the segmentation result. For example, the lifetime of five clusters simply means the total number of times producing five clusters in the data set within the scope of k ($k=1 \ldots l$).

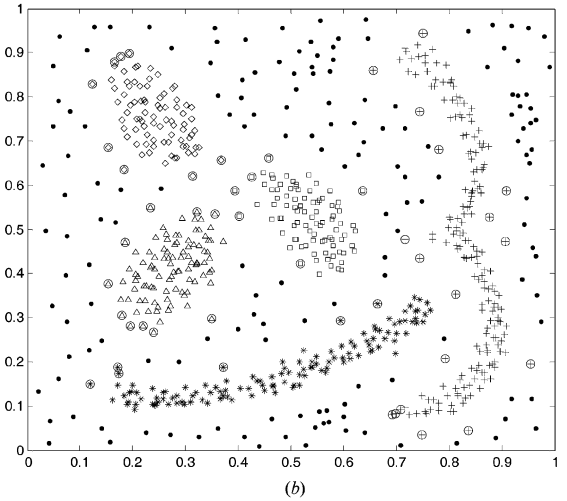
Table 1 lists the cluster number along with different k. We found that the cluster number equals 5 when $k=9$, 10, …, 17 and each of the 5 clusters is 'alive' during this period. In other words, no new clusters emerged during this period. Consequently, the lifetime during which 5 clusters stay 'alive' is 9. This means that the most stable state of the simulated data should be divided into 5 clusters. Besides 5, the 8 and 4 clusters have the lifetime of 3 after the aliveness of each cluster was checked. These two results are the sub-stable states. Hence, 5 clusters will be considered as the proper number of clusters in the simulated data set according to the lifetime of number of clusters.

3.4. *Restriction of the algorithm*

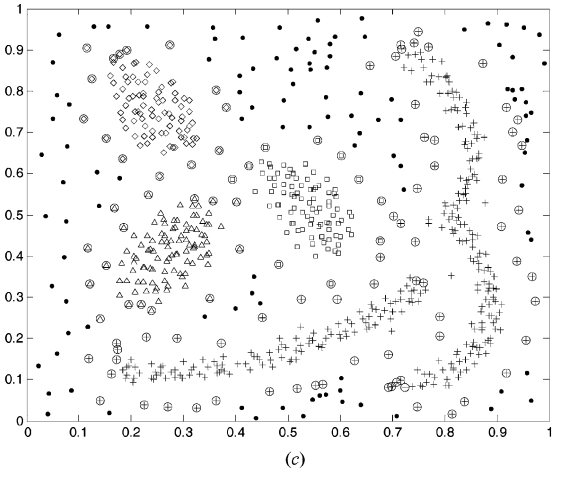
Although this algorithm can find the clustered points and detect the number of cluster features, it still has one limitation: this algorithm can only be applicable to those point sets, in which only two point processes are dominant, i.e. the feature and the noise. Otherwise, the λ calculated cannot be used to represent each cluster feature.



(*a*)



(*b*)



(*c*)

Figure 3. Clustering result of simulated data. (*a*) $k=5$; (*b*) $k=9$; (*c*) $k=18$. The number of clusters decreases, and the number of border points increases as k increases (encircled symbols represent border points; solid points represent noise).

Table 1. Number of clusters of simulated data for different values of k.

Value of k	1	2	3	4	5	6	7	8	9	10
Cluster number	*47*	*11*	*9*	*9*	*8*	*8*	*8*	*6*	*5*	*5*
Value of k	11	12	13	14	15	16	17	18	19	20
Cluster number	*5*	*5*	*5*	*5*	*5*	*5*	*5*	*4*	*4*	*4*

4. Case study: determining spatial clusters of earthquakes

4.1. *Clusters of earthquakes*

Clustered earthquakes are usually considered as the foreshocks or aftershocks of a strong earthquake. Clustered earthquakes are perceived to be foreshocks if a strong earthquake breaks out after the clustered earthquakes and are viewed as aftershocks if a strong earthquake breaks out before the clustered earthquakes. Thus, clustered earthquakes could serve as a primary clue to predict earthquakes if the possibility of their being aftershocks of some strong earthquakes can be excluded (Chen *et al.* 1999, Ripepe *et al.* 2000). In addition to foreshocks, aftershocks of strong earthquakes can also be considered as clustered earthquakes, namely a seismic sequence (Wu *et al.* 1990). The determination of the area of a seismic sequence is very useful for understanding its changing trend and the mechanism of strong earthquakes. Therefore, the determination of clustered earthquakes is attracting more and more attention in the seismic-research community.

As we known, foreshocks and aftershocks are not only spatially clustered but also temporally clustered. Hence, the temporal segmentation can sometimes help to pick out the spatially clustered earthquakes. However, in most cases, some background earthquakes simultaneously break out with those spatially clustered earthquakes, and in particular, aftershocks and (or) foreshocks within different areas may happen in the same period, so the temporal segmentation may only provide the temporally clustering information of earthquakes and cannot effectively determine the spatially clustered earthquakes. In order to locate the spatially clustered earthquakes, background earthquakes should be removed using spatially concentrated nature of clustered earthquakes (in the following text, 'clustered earthquakes' are referred to as 'spatially clustered earthquakes'). The background earthquakes are referred to as several small earthquakes that break out at a stable rate within a certain area (Diao *et al.*, 1994, Wyss and Toya 2000, Pei *et al.* 2003). In this regard, background earthquakes and clustered earthquakes can be deemed as two overlaid spatial Poisson processes with different rates of λ. The separation of these two types of earthquakes can be used as a test case for evaluating the proposed method of detecting spatial clusters from a point data set.

4.2. *Study area and seismic data*

4.2.1. Study area. The study area is located between 100–107° E and 27–34° N (figure 4). It contains the east part of Tibet, south part of Sichuan and Chongqing, north part of Yunnan, and west part of Guizhou. From geotectonic point of view, this area is the transition from the Tibetan plateau to the Yangtze Platform. It is one of the areas with the most intensive seismicity in China. Sixteen devastating earthquakes ($M \geqslant 6.0$) occurred in this area from 1970 to 2002 (Feng and Huang 1980, 1989, China Seismograph Network Data Management Center 2004). The seismic records in this area may not only indicate the law of seismicity but also

Figure 4. Location of the research area.

provide evidence to help us to understand the tectonic movement of this area (Wang *et al.* 2003).

4.2.2. Seismic dataset. All of the catalogue data are from the Seismic Catalog of West China (1970–1975, $M \geqslant 1$) (Feng and Huang 1980) and Seismic Catalog of West China (1976–1979, $M \geqslant 1$) (Feng and Huang 1989). The selected seismic data are from 15 February 1975 to 15 August 1976 and larger than 2 (M). Thus, 236 epicentres are obtained altogether. Because the devastating Xingtai quake ($M=7.2$) caused serious losses in 1966, the Chinese government paid more attention to monitoring and predicting the seismicity, and began to set up the seismograph network across the whole country. In the following few years, more than 400 seismograph stations were founded, and the monitoring ability for seismicity has been greatly promoted in terms of the measurement precision and reaction speed (Jiao *et al.* 1990). According to Jiao *et al.* (1990), in the research area the floor limit of seismic monitoring and measuring in this area has been lowered to 2 (M). In addition, the errors of the epicentres ($3 > M \geqslant 2$) were less than 15 km, and the errors of the epicentres ($M \geqslant 3$) were less than 5 km. Therefore, the integrality and the quality of this dataset satisfy the requirement of this research.

4.3. *Results of the case study*

The number of clusters with different values of k is listed in table 2. We found that there are two stable states in the sequence of number of clusters, 4 and 3, after

Table 2. Number of earthquake clusters in western China for different values of k.

Value of k	1	2	3	4	5	6	7	8	9	10
Cluster number	*29*	*20*	*12*	*8*	*6*	*5*	*4*	*4*	*4*	*4*
Value of k	11	12	13	14	15	16	17	18	19	20
Cluster number	*4*	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*2*	*2*

checking the aliveness of each cluster. Moreover, 3 clusters have the longest lifetime. According to the lifetime trend and the spatial distribution of clustered earthquakes (figure 5), we think that the features in figure 5(*b*) are more likely the clustered earthquakes in the research area.

By running the algorithm with $k=12$, we obtain three clustering features (taking $k=12$ is to reduce the number of border points). We mark them as feature A, B, and C, respectively.

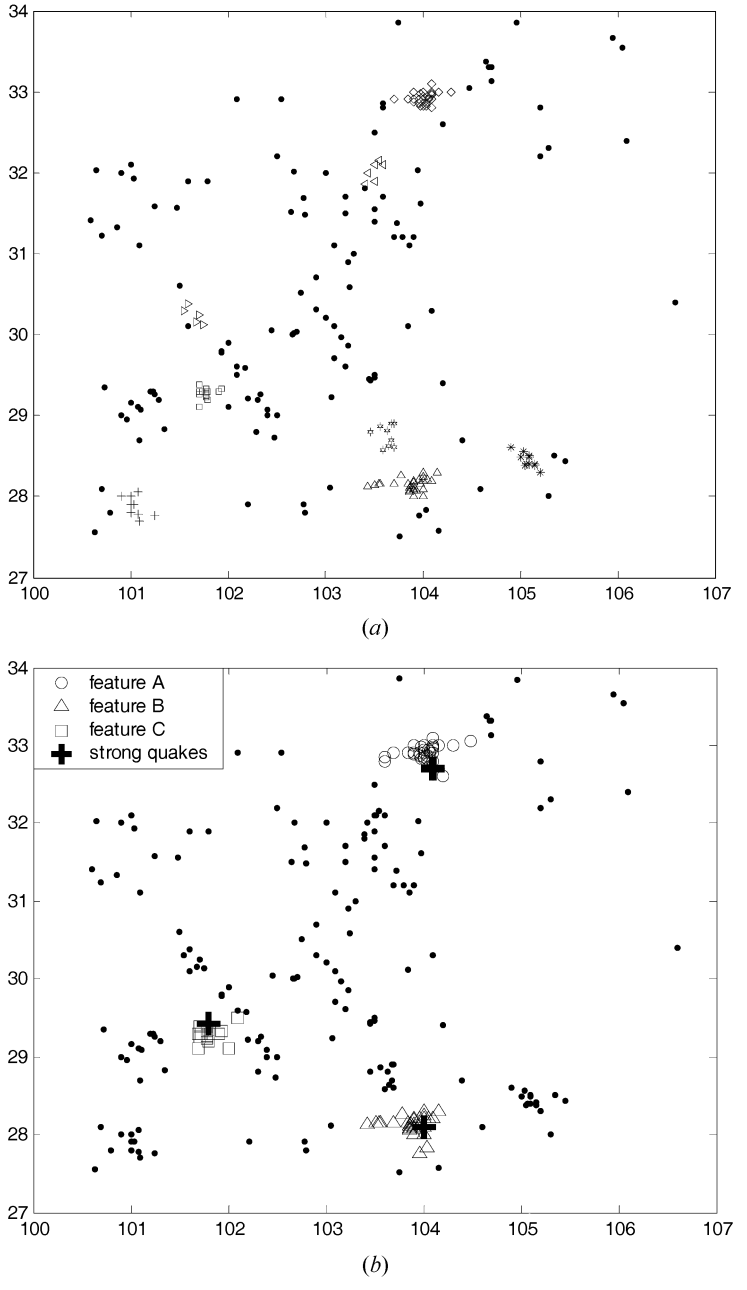


Figure 5. Clustered earthquakes with different k. (*a*) $k=4$ (8 clusters); (*b*) $k=12$ (3 clusters).

4.4. *Analysis and discussion of the case-study results*

Do these clustered earthquakes really belong to aftershocks or foreshocks of strong earthquake? The seismic records allow us to answer this question. For features B and C, they can be viewed as the aftershocks of strong earthquakes. According to Zhang (1986), feature B is the aftershocks of Kangding-Jiulong quake (M=6.2), which broke out at 29° 26′ N, 101° 48′ E on 15 January 1975. Feature C is the aftershocks of Daguan quake (M=7.1), which broke out at 28° 06′ N, 104° 00′ E on 11 May 1974. Feature A can be viewed as the foreshocks of Songpan quake. The strong earthquake of Songpan (M=7.2) broke out at (32°42′ N,104°06′ E) on 16 August 1976 (Zhang 1990). The main shock was located not in the centre of feature A but in the south part of the area. Feature A apparently indicated the outbreak of the Songpan strong quake (note that the seismic data used in this case study span from 15 February 1975 to 15 August 1976).

5. Conclusions and future work

Finding clustering features and determining their number from the spatial database are major challenges in spatial data mining. Most methods cannot achieve both of these in an automated way. In this paper, we present a new approach based on the NN method to accomplish the task. We also employed the concept of density-connected and the concept of the lifetime of a number of clusters to determine the proper number of clusters. Our approach requires only one parameter k and reduces the run time. Although the method was applied to seismic data for discerning spatial patterns of clustered earthquakes, it can be easily applied to other areas of point processes such as landslide and spatial distribution of cancers.

As discussed in section 3.4, this algorithm is limited to those point sets in which only two points processed are dominant; future theoretical work will focus on the analysis for the overlaid process, which may contain more than two point processes.

Acknowledgements

This study was funded through support from a grant (Project Number 2006CB701305) from National Key Basic Research and Development Programme of China, the 'Hundred Talents' Programme of Chinese Academy of Sciences, a grant (Project Number: 40225004) from National Natural Science Foundation of China. Support from the University of Wisconsin-Madison is also appreciated.

References

ALLARD, D. and FRALEY, C., 1997, Nonparametric maximun likelihood Estimation of features in spatial point process using Voronoi tessellation. *Journal of the American Statistical Association*, **92**, pp. 1485–1493.

ANKERST, M., BREUNIG, M., M., KRIEGEL, H.-P. and SANDER, J., 1999, OPTICS: Ordering points to identify the clustering structure. In *Proceedings of ACM-SIGMOD'99 International Conference on Management Data* (Philadelphia, PA), pp. 46–60.

BANFIELD, J.D. and RAFTERY, A.E., 1993, Model Based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, pp. 803–821.

BECKMANN, N., KRIEGEL, H.P., SCHNEIDER, R. and SEEGER, B., 1990, The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 322–331 (Atlantic City, NJ).

BRIMICOMBE, A.J., 2003, A variable resolution approach to cluster discovery in spatial data mining. *Lecture Notes in Computer Science*, **2669**, pp. 1–11.

BYERS, S. and RAFTERY, A.E., 1998, Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, **93**, pp. 557–584.

CELEUX, G. and GOVAERT, G., 1992, A classification EM algorithm and two stochastic versions. *Computational Statistics and Data Analysis*, **14**, pp. 315–332.

CHEN, Y., LIU, J. and GE, H.K., 1999, Pattern characteristics of foreshock sequences. *Pure and Applied Geophysics*, **155**, pp. 395–408.

China Seismograph Network Data Management Center 2004, China Seismograph Network (CSN) Catalog. Available online at: http://www.csndmc.ac.cn (accessed 2005).

DASGUPTA, A. and RAFTERY, A.E., 1998, Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, pp. 294–302.

DASZYKOWSKI, M., WALCZAK, B. and MASSART, D.L., 2001, Looking for natural patterns in data Part 1. Density-based approach. *Chemometrics and Intelligent Laboratory Systems*, **56**, pp. 83–92.

DIAO, S.Z., GUO, A.X. and WANG, W.H., 1994, Characteristics of early background seismicity and predictive of strong earthquake risk region with $M \geqslant 7$. *Earthquake Research in Plateau*, **6**, pp. 40–46 (in Chinese).

ESTER, M., KRIEGEL, H.P., SANDER, J. and XU, X.W., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 1996 International Conference on Knowledge Discovery and Data Mining*, pp. 226–231 (Portland, OR).

FENG, H. and HUANG, D.Y., 1980, *A Catalogue of Earthquakes in Western China (1970–1975, $M \geqslant 1$)* (Beijing: Seismological Press) (in Chinese).

FENG, H. and HUANG, D.Y., 1989, *Earthquake Catalogue in West China (1976–1979, $M \geqslant 1$)* (Beijing: Seismological Press) (in Chinese).

FRALEY, C. and RAFTERY, A.E., 1998, How many cluster? Which clustering method? Answer via model-based cluster analysis. *Computer Journal*, **41**, pp. 578–588.

HAN, J.W., KAMBER, M. and TUNG, A.K.H., 2001, Spatial clustering methods in data mining. In *Geographic Data Mining and Knowledge Discovery*, H.J. Miller and J.W. Han (Eds), pp. 188–217 (London: Taylor & Francis).

JEMAL, A., KULLDORFF, M., DEVESA, S.S., HAYES, R.B. and FRAUMENI, J.F., 2002, A geographic analysis of prostate cancer mortality in the United States, 1970–89. *International Journal of Cancer*, **101**, pp. 168–174.

JIAO, Y.B., WU, K.T. and YANG, M.D., 1990, An assessment about capability and quality of our country earthquake observation network. *Earthquake Research in China*, **6**, pp. 1–7 (in Chinese).

MOON, T.K., 1996, The Expectation-Maximization algorithm. *IEEE Signal Processing Magazine*, **13**, pp. 47–60.

PEI, T., YANG, M., ZHANG, J.S., ZHOU, C.H., LUO, J.C. and LI, Q.L., 2003, Multi-scale expression of spatial activity anomalies of earthquakes and its indicative significance on the space and time attributes of strong earthquakes. *Acta Seismologica Sinica*, **3**, pp. 292–303.

REASENBERG, P.A., 1999, Foreshock occurrence rates before large earthquakes worldwide. *Pure and Applied Geophysics*, **155**, pp. 355–379.

RIPEPE, M., PICCININI, D. and CHIARALUCE, L., 2000, Foreshock sequence of September 26th, 1997 Umbria-Marche earthquakes. *Journal of Seismology*, **4**, pp. 387–399.

SANDER, J., ESTER, M., KRIEGEL, H.P. and XU, X.W., 1998, Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, **2**, pp. 169–194.

STEENBERGHEN, T., DUFAYS, T., THOMAS, I. and FLAHAUT, B., 2004, Intra-urban location and clustering of road accidents using GIS: a Belgian example. *International Journal of Geographical Information Science*, **18**, pp. 169–181.

UMINO, N., OKADA, T. and HASEGAWA, A., 2002, Foreshock and aftershock sequence of the 1998(M 5.0) Sendai, northeastern Japan, earthquake and its implications for earthquake nucleation. *Bulletin of the Seismological Society of America*, **92**, pp. 2465–2477.

WANG, C.Y., HAN, W.B., WU, J.P., LOU, H. and BAI, Z.M., 2003, Crustal structure beneath the Songpan–Garze orogenic belt. *Acta Seismologica Sinica*, **3**, pp. 237–250.

WU, K.T., JIAO, Y.B., LŰ, P.L. and WANG, Z.D., 1990, *Panorama of Seismic Sequence* (Beijing: University Press) (in Chinese).

WYSS, M. and TOYA, Y., 2000, Is background seismicity produced at a stationary poissonian rate. *Bulletin of the Seismological Society of America*, **90**, pp. 1174–1187.

ZHANG, Z.C., 1986, *Earthquake Cases in China (1966–1975)* (Beijing: Seismological Press) (in Chinese).

ZHANG, Z.C., 1990, *Earthquake Cases in China (1976–1980)* (Beijing: Seismological Press) (in Chinese).