Density-based clustering for data containing two types of points

Tao Pei, Weiyi Wang, Hengcai Zhang, Ting Ma, Yunyan Du and Chenghu Zhou*

State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing 100101, China

(Received 10 July 2014; final version received 11 August 2014)

When only one type of point is distributed in a region, clustered points can be seen as an anomaly. When two different types of points coexist in a region, they overlap at different places with various densities. In such cases, the meaning of a cluster of one type of point may be altered if points of the other type show different densities within the same cluster. If we consider the origins and destinations (OD) of taxicab trips, the clustering of both in the morning may indicate a transportation hub, whereas clustered origins and sparse destinations (a hot spot where taxis are in short supply) could suggest a densely populated residential area. This cannot be identified by previous clustering methods, so it is worthwhile studying a clustering method for two types of points. The concept of two-component clustering is first defined in this paper as a group containing two types of points, at least one of which exhibits clustering. We then propose a density-based method for identifying two-component clusters. The method is divided into four steps. The first estimates the clustering scale of the point data. The second transforms the point data into the 2D density domain, where the x and y axes represent the local density of each type of point around each point, respectively. The third determines the thresholds for extracting the clusters, and the fourth generates two-component clusters using a density-connectivity mechanism. The method is applied to taxicab trip data in Beijing. Three types of two-component clusters are identified: high-density origins and destinations, high-density origins and low-density destinations, and low-density origins and high-density destinations. The clustering results are verified by the spatial relationship between the cluster locations and their land-use types over different periods of the day.

Keywords: Origins and destinations of taxicab trip; kth nearest distance; density-based cluster

1. Introduction

When one type of point is distributed over a region, an area containing points of high density can be seen as a cluster, indicating a meaningful anomaly. A number of density-based clustering methods have been developed to identify clusters from data containing only one type of point (Ester et al. 1996, Ankerst et al. 1999, Pei et al. 2006, Wan et al. 2012). However, when two different types of points coexist and overlap with various density combinations (say, origins and destinations (OD) of taxicab trips, crime venues and traffic accidents, restaurants and shops in a city), the meaning of a cluster of one type of point may change if points of the other type covered by this cluster have a different density (i.e., aggregation, random, or sparse). Here, we define a two-component cluster as
a group containing two types of points, at least one of which exhibits clustering. We then give some examples to illustrate the significance of identifying two-component clusters.

1. ODs of taxicab trips: A cluster with a high density of origins and destinations may indicate a transportation hub. One with a high density of origins and a low density of destinations in the morning (a hot spot where taxis are in short supply) may indicate a residential area with a dense population.

2. Traffic accidents and crime venues in a city: High levels of traffic accidents with sparse crime venues may indicate a flaw in the design of the local transportation system; those with higher densities of crime venues may indicate the mutual inducement of traffic accidents and crimes.

3. High-temperature events and precipitation events: Clustered high-temperature events aggregated with precipitation events may indicate a flood; those with a low density of precipitation events may indicate a drought.

From these examples, we can appreciate the importance of identifying two-component clusters. Regarding (1), taxi drivers may find more opportunities in the cluster of high-density origins and low-density destinations, whereas passengers wishing to find a taxi should move to the cluster of low-density origins and high-density destinations. Of course, a cluster with a high density of origins and destinations is also a good choice for both taxi drivers and passengers. In addition, the city government can locate hot spots where taxis are in short supply and those where they exceed demand. This information may shed light on the city dynamics and thereby assist urban planning. Regarding (2), the transportation system may require reconstruction around the cluster of high-density traffic accidents and sparse crime venues, while more policemen and surveillance monitors should be deployed to the area of aggregated crime venues and traffic accidents. For Point (3), different relief measures should be taken at clusters where temperature and precipitation events have different combinations. Nevertheless, most previous clustering methods can only deal with one type of point and are difficult to identify two-component clusters. To this end, it is important to develop a method for two-component clusters.

In recent years, significant progress has been made in density-based clustering methods. Ester et al. (1996) proposed Density Based Spatial Clustering of Applications with Noise (DBSCAN) to identify high-density point groups from point data. However, in DBSCAN, the parameters (e.g., eps, the distance used to separate clusters from noise) must be defined in an interactive way. To adjust the parameters visually, Ankerst et al. (1999) constructed an enhanced density-connected algorithm, referred to as Ordering Points To Identify the Clustering Structure (OPTICS). Many modifications of DBSCAN were subsequently proposed to enhance the cluster identification ability (Sander et al. 1998, Daszykowski et al. 2001, Roy et al. 2005, Pei et al. 2009, Wang and Huang 2009, Ashour et al. 2011, Tran et al. 2013). However, existing clustering models are still limited to one type of point and are unsuitable for clusters containing two (or more) types of points.

The spatial scan statistic relates to two different types of points (i.e., case and control). The cases are seen as a single-point process occurring with low probability, whereas the controls are generally distributed across the research area (Kulldorff 1997, Kulldorff et al. 2006, Neill 2006). The spatial scan method looks for the most likely cluster by sweeping over all zones circumscribed by circles of varying radii centered at each polygon in the area. Later, modifications were made to adapt the method to arbitrarily shaped clusters (Duczmal and Assunção 2004, Janeja and Atluri 2008, Pei et al. 2011, Wan et al. 2012). The spatial scan method finds clusters with a high ratio between cases and controls. Nevertheless, a high-ratio area identified by the spatial scan method may not signify a real anomaly of cases (namely, false positive), because the number of controls could be quite
low. To better visualize a high-risk area, Guo et al. (2012) used the shared nearest distance to form the OD points of taxicab trips into a network, the nodes of which are actually similar-size clusters. The distribution of the net flow ratio can then be determined at each node. Although this method reveals anomalies of high ratios between inflow and outflow, smoothed by a window covering the same number of points instead of the same size, the anomalies identified are not those in which points are aggregated.

A literature review shows that few methods are able to identify clusters containing two types of points. To solve this problem, we construct a method for identifying two-component clusters. Inspired by the 2D Fourier transform, we first determine eps for constructing the density domain of point data. Second, we transform the point data into the 2D density domain, in which the x and y axes represent the local density of each type of point around each point, respectively. The thresholds for identifying two-component clusters are then determined. Points located between the thresholds are extracted to finally form clusters based on their density and connectivity. The rest of the paper is arranged as follows. Section 2 introduces some basic concepts of two-component clusters. In Section 3, the clustering algorithm is described in detail, and Section 4 presents clustering results with synthetic data. A case study using Beijing taxicab data is given in Section 5. Section 6 provides a summary of this paper, as well as directions for future research.

2. Basic concepts

As our method uses density-based clusters, some of the basic concepts relating to two types of points are modified and extended from those defined in DBSCAN.

(1) Mixed eps-neighborhood of Point p : n points, including \( n_1 \) of Type A \( \left(o_i (i = 1, 2, \ldots, n_1)\right) \) and \( n_2 \) of Type B \( \left(q_j (j = 1, 2, \ldots, n_2)\right) \), in the circle centered at p with radius eps, where \( n = n_1 + n_2 \). Here, Point p can be a point of A or B (Figure 1).

\[
N_{meps}(p) = \left\{ o_i, q_j | d_{(p, o_i)} < \text{eps}; d_{(p, q_j)} < \text{eps}, (i = 1, 2, \ldots, n_1, j = 1, 2, \ldots, n_2) \right\}
\]

(2) Mixed core point for high density of A and high density of B: Point p is a mixed core point with respect to (wrt) eps, MinPtA, and MinPtB if its mixed eps-neighborhood \( N_{meps}(p) \) contains at least MinPtA points of A and MinPtB points

![Figure 1. Mixed eps-neighborhood of Point p (dots represent points of Type A and circles represent points of Type B. There are three points of Type A and two points of Type B in the mixed eps-neighborhood).](image-url)
of B. In Figure 2, $p_2$ and $p_3$ are the mixed core points wrt $\text{eps} = e$, $\text{MinPtA} = 2$, and $\text{MinPtB} = 2$. Because no point of A is in the mixed eps-neighborhood of Point $p_1$, it cannot be a mixed core point. Note that we abbreviate ‘high density of A and high density of B’ as HA-HB hereafter. Similarly, ‘high density of A and low density of B’ is denoted by HA-LB, and ‘low density of A and high density of B’ is shortened to LA-HB.

3. Method

3.1. Two-component clustering method

The method is divided into four steps. First, we determine $\text{eps}$, which is used to construct the density domain of point data. Second, the point data are transformed into the density domain, which is displayed as a 2D image (Figure 3). Third, we estimate four thresholds for separating two-component clusters from noise. Specifically, one is for high-density points of Type A ($X_h$) (i.e., $\text{MinPtA}$ in Section 2), one is for low-density points of A ($X_l$) ($\text{MaxPtA}$), and similarly for points of Type B ($Y_h$).
and \( Y_1 \) (MinPtB and MaxPtB). Using these thresholds, the density domain is divided into nine regions (Figure 3). Fourth, we extract mixed core points (e.g., mixed core points for HA-HB clusters are contained in the area between \( X_h \) and \( Y_h \)) and identify different types of clusters from the core points based on the mixed density-connectivity mechanism. We now describe each step in detail.

### 3.2. Choosing \( \varepsilon \) by determining the clustering scale

Given a point, its location in the density domain is determined by the local density of Type A (x-axis) and that of Type B (y-axis) (Figure 3). Specifically, the local density of A can be represented by the number of points of Type A in its \( \varepsilon \)-neighborhood, and similarly for the local density of B. Differing from the estimation of \( \varepsilon \) in the previous literature (Ester et al. 1996, Ashour et al. 2011, Jiang et al. 2011), which is used to separate the clustered points from noise after fixing MinPts, in this context, \( \varepsilon \) is the only parameter to transform spatial points into the density domain. In practice, the value of \( \varepsilon \) is related to the clustering scale in the data and should be in a certain range. On the one hand, if \( \varepsilon \) is too large (say, larger than the scale of a cluster), the local density of a point in the cluster will be underestimated because more noisy points are included in its \( \varepsilon \)-neighborhood (Pei et al. 2007). On the other hand, if \( \varepsilon \) is too small, the estimation of the local density of a point could be unstable because of the inhomogeneous distribution of points over space (Pei et al. 2006). To determine the relationship between \( \varepsilon \) and the clustering scale, which may result in a good clustering result, we first estimate the clustering scale using H-function. Then, we estimate the clustering scale factor between \( \varepsilon \) and clustering scale (note that \( \varepsilon \) is the product of the estimated clustering scale and the clustering scale factor).

Previous studies have shown that the clustering scale can be estimated by Ripley’s K-function and H-function (Ripley 1976, Besag 1977). The K-function is defined as

\[
K(h) = \lambda^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\delta_{ij}(h)}{n}, \quad (i,j = 1, 2, \ldots, n; i \neq j)
\]  

Figure 3. Density domain for data containing two types of points.
where \( \delta(h) = \begin{cases} 1 & (h_{ij} \leq h) \\ 0 & (h_{ij} > h) \end{cases} \), \( h_{ij} \) is the distance between Point \( i \) and Point \( j \), \( n \) is the number of points, and \( \lambda \) is the density of points in the research area. The H-function is defined as

\[
H(h) = \sqrt{\frac{K(h)}{\pi}} - h
\]  

Nevertheless, because the estimation of the clustering scale is significantly affected by the distance between clusters (i.e., domains in Kiskowski et al. (2009)), Kiskowski et al. (2009) used the derivative of the H-function to estimate the clustering scale (in detail, the clustering scale is the value yielding the minimum of the derivative function). In this paper, we use the derivative of Ripley’s H-function to estimate the clustering scale. Simulations under different scenarios also support this theory (see the supplemental document for details).

To determine the appropriate values of clustering scale factor, we generated different synthetic data sets. Experiments with synthetic data have indicated that when \( \varepsilon \) is around a quarter of the clustering scale (i.e., the scale of cluster(s) in a point data set), the classification error rate is minimized (for details of the experiments, please see the supplemental document). For this reason, we set the clustering scale factor to 1/4 when performing the clustering method.

3.3. Density domain of two types of points

With \( \varepsilon \), we can transform data containing two types of points into the density domain. If we use two thresholds for each type of point (one is the low-density threshold and the other is the high-density threshold), then there are four thresholds (specifically, \( X_h, X_l, Y_h \), and \( Y_l \)). The density domain is divided into nine regions (Figure 3). Thus, we call Region 1 the LA-LB region and name each region accordingly up to Region 9, which is the HA-HB region.

3.4. Determination of the thresholds and two-component clusters

As discussed in Section 3.3, four thresholds (i.e., \( X_h, X_l, Y_h \), and \( Y_l \)) must be estimated to divide the density domain. Next, we take \( X_h \) and \( X_l \) as examples to illustrate the determination of the thresholds. Threshold \( X_h \) is used to define high-density clusters of \( A \), whereas \( X_l \) is used to define low-density groups of \( A \). These two thresholds are determined under the assumption that all points of \( A \) are randomly distributed over the region (namely, the points obey a spatial Poisson process). As a result, before determining the thresholds, we must estimate the density \( (\lambda) \) of the points of \( A \) as follows:

\[
\lambda = \frac{k}{\pi E(d^2_{i,k})} = \frac{nk}{\pi \sum_{i=1}^{n} (d^2_{i,k})}
\]  

where \( d_{i,k} \) is the distance of the \( k \)-th closest point to Point \( i \), and \( n \) is the number of points (Pei et al. 2007). Here, we set \( k = 1 \).

The thresholds are then determined from the probability density function of the Poisson process (i.e., \( P(x = k) = \frac{e^{-\lambda \varepsilon^2} (\lambda \varepsilon^2)^k}{k!} \)) at a certain significance level, where \( \varepsilon \) has
already been determined based on the estimation of the clustering scale, and \( \lambda \) is the density of the points of A. We set the significance level to 0.2 and then estimate \( X_h \) and \( X_l \) such that 10% of the points are greater than \( X_h \) and 10% are less than \( X_l \). As we know, the significant level is traditionally set to 0.1, 0.05, and 0.01. In the context, the larger value indicates that more clusters will be found. Here, we set the significant level to a larger value (i.e., 0.2) to ensure that all possible clusters will be identified even if some false positives are generated. The other parameters (i.e., \( Y_h \) and \( Y_l \)) are estimated in the same way. Once all thresholds have been estimated, they are used to divide the density domain into nine regions.

3.5. Generating density-connected clusters

With the estimated parameters, different types of clusters can be generated from corresponding regions of the density domain. Note that only three types of clusters (HA-HB, HA-LB, and LA-HB) are considered in this paper. HA-HB clusters can be generated from the points in HA-HB region of the density domain based on the mixed density-connected mechanism. Similarly, HA-LB (LA-HB) clusters can be produced from points in HA-LB (LA-HB) region. Here, we only describe how an HA-HB cluster is generated, and other types of clusters can be identified in the same way. We first randomly choose a mixed core point (it can be either a point of Type A or that of Type B) from HA-HB region and assign it a cluster ID. Then, we find the next mixed core point for HA-HB in its neighborhood and assign the same cluster ID. This process proceeds until no unclassified mixed core point for HA-HB is found in the neighborhoods of the classified points of the same ID. All points with the same ID thus form a cluster. Another cluster begins with an unclassified mixed core point for HA-HB. All HA-HB clusters are identified until no unclassified mixed core point for HA-HB can be found in the whole data set.

Similar to the relationship between 1D Fourier and 2D Fourier transform (specifically, 2D Fourier transform can be realized by performing 1D Fourier transform horizontally, followed by the same transform vertically), our method can be accomplished by overlaying the results of clustering two types of points separately. We omit the details for short. With the same idea, our method can be extended to those multi-component data, which contain more than two types of points.

3.6. Complexity of the algorithm

The complexity of the clustering algorithm is the sum of the complexities of the following procedures: the determination of eps, estimation of the density thresholds, and the connection process of mixed core points. Because the major step in determining eps is the calculation of the \( k \)th nearest neighbor, the complexity of this step is \( O(n \log(n)) \), where \( n \) is the number of points (Pei et al. 2006). The major step in estimating the thresholds is the same as that for eps. Therefore, its complexity is \( O(n \log(n)) \). The complexity of the density-connected process is again \( O(n \log(n)) \) (Pei et al. 2006). Hence, the overall complexity of the algorithm is \( O(n \log(n)) \).

4. Experiments using synthetic data

To validate the method, we use the synthetic data shown in Figure 4a, where squares symbolize points of Type A and triangles symbolize points of Type B. We generate five clusters: the inverse ‘T’ (top, HA-LB), ‘cross’ (near middle, LA-HB), ‘square’ (right-hand side, HA-HB),
Figure 4. Clustering result of synthetic data (a) original data; (b) clusters (squares symbolize points of Type A and triangles symbolize points of Type B; in (b), HA-HB clusters are in red; HA-LB clusters are in green; LA-HB clusters are in blue).
‘line’ (bottom, HA-LB), and ‘circle’ (left-hand side, LA-HB). In this data set, the clusters are of
different types (including HA-HB, HA-LB, and LA-HB), sizes, shapes, and densities.

To determine the clusters, we need to estimate eps in light of the clustering scale. The derivative of the H-function is shown in Figure 5, indicating that the clustering scale
is 13.01. As indicated in Section 3.2, we set the clustering scale factor to 1/4 (namely, eps is a quarter of the clustering scale, or 3.25). The density domain is shown in
Figure 6. According to the method described in Section 3.4, the thresholds of $X_h$, $X_l$,
$Y_h$, $Y_l$ are estimated to be 14, 6, 14, and 6, respectively. The regions of different
clusters are then generated using the thresholds. With these thresholds, we can determine
the HA-HB, HA-LB, and LA-HB clusters, as displayed in Figure 4b. The error rate of
classification is 2.25%, indicating that our method has successfully identified the two-
component clusters in the data set.

To further validate our method, we calculate the error rates at different values of the
clustering scale factor, ranging from 1/16 to 1 (the first line in Table 1). The result shows
that the minimum error rate is acquired when the clustering scale factor is 1/4, which
confirms our strategy of estimating eps. We then generate 199 further data sets for further
verification. The mean error rates at different clustering scale factors are shown in Table 1
(the second line). The results demonstrate that low error rates are found when eps is
between 3/16 and 3/8 of the clustering scale. Specifically, the lowest error rate is found at
1/4. Experiments with four different types of synthetic data in the supplemental document
are also in agreement with the result.

5. Case study

Although the spatial distribution of taxi trip ODs has often been considered (Lee et al.
2008, Andrienko et al. 2009, Yue et al. 2011, Guo et al. 2012, Liu et al. 2012, Zhang
et al. 2012, Pan et al. 2013), few studies have sought to identify two-component clusters. In this section, we apply our method to Beijing taxicab data. The data include the ODs of more than 18,000 taxis in Beijing on Monday 18 April 2011. We consider two periods (7am–10am and 4pm–7pm), representing the morning and afternoon rush hours, to examine the distribution of two-component clusters. The ODs of each period are displayed in Figure 7.

### 5.1. Results

The value of eps for generating the density domain was estimated at 150 m. The thresholds of $X_h, X_l, Y_h, Y_l$ for the morning period were determined to be 24, 13, 21, and 10, and those for the afternoon were 26, 14, 19, and 9, respectively. Note that $\lambda$ is calculated using Equation (3). The traditional method calculates $\lambda$ by dividing the number of ODs ($n$) by the area of the whole region ($S$). This may underestimate $\lambda$, because taxis cannot reach some regions (e.g., buildings and lakes) that are still included in the area (i.e., $S$). Pei et al. (2007) suggest using a method based on the nearest distance to reduce the estimation bias. This is

Table 1. Error rates generated at different clustering scale factors (%).

| Coefficients for eps | 1/16 | 3/32 | 1/8 | 3/16 | 1/4 | 3/8 | 1/2 | 3/4 | 1 |
|----------------------|------|------|-----|------|-----|-----|-----|-----|---|
| Selected data set    | 37.85| 27.16| 24.20| 2.50 | 2.03| 4.22| 6.15| 19.72| 23.14|
| Total                | 35.47| 27.18| 20.22| 2.90 | 2.85| 5.66| 8.16| 23.39| 28.42|
because, as indicated in Equation (3), $\lambda$ can be thought of as the ratio between the number of points (if $k = 1$) and the sum of the areas of the circles, each of which is centered at a data point and has a radius of the distance to its nearest neighbor. Thus, areas that are not available for taxis’ arrival can be excluded when calculating $\lambda$ (because no such circle can
be generated in those areas). For details, refer to Pei et al. (2007). Figures 8 and 9 show the clustering results of the morning and afternoon periods, respectively.

5.2. Discussion

Previous studies have shown that the distribution of taxi trip ODs is closely related to land-use types (Guo et al. 2012, Liu et al. 2012, Zhang et al. 2012, Pan et al. 2013). We seek to explain the clustering results based on the relationship between the cluster locations and their land-use types. The land-use map of Beijing is shown in Figure 10. The clusters are assigned to different land-use types through the following method. In each cluster, every point is assigned to the specific land-use type of its nearest cell. The nearest cell is determined based on the distance between the point and the cell centers. The cluster is assigned to the land-use type of the most points in the cluster. Note that in Figures 8 and 9, the color of each cluster represents its land-use type. The numbers and
proportions (i.e., the proportion of the total assigned to each cluster type) of different clusters are listed in Tables 2 and 3.

From Table 2, we see that in the morning, the greatest number of ‘Residential’ clusters is of type HO-LD, followed by LO-HD and HO-HD. The proportion of HO-LD is also larger than that of LO-HD. In the afternoon (Table 3), the numbers of clusters labeled as ‘Residential’ are, in descending order, of types LO-HD, HO-HD, and HO-LD. At this time of day, the proportion of LO-HD is larger than that of HO-LD. The spatiotemporal variation of clusters in the residential areas indicates that, in the morning, hot spots that are short of taxis are dominant, whereas in the afternoon, hot spots generating a surplus of taxis are more common.

Unlike ‘Residential’ zones, the numbers and proportions of HO-LD clusters are smaller than those of LO-HD around ‘Public facilities’ and ‘Educational institutions’ in the morning, while in the afternoon the numbers and proportions of HO-LD clusters are greater than

Figure 9. Clustering results for data from 4pm–7pm (a) HO-HD; (b) HO-LD; (c) LO-HD.
those of LO-HD. These spatiotemporal variations indicate the inverse pattern from that in ‘Residential’ areas. The above phenomena agree with the common-sense notion that more residents go to work (i.e., to public facilities and educational institutions) in the morning and leave for home in the afternoon. In addition, the number of HO-HD clusters is the highest for ‘Educational institutions’ in both morning and afternoon. The reason for ‘Educational institutions’ generating many HO-HD clusters is that in China many ‘Educational institutions’ also include residential areas mainly for staffs and their families, so they are potentially hot places for both origins and destinations no matter what the period is. As a result, it is not strange to see many HO-HD clusters appearing in ‘Educational institutions’.

Note that in commercial areas, besides HO-HD clusters, the number of LO-HD clusters is larger than that of HO-LD in both periods. The existence of a high number of HO-HD clusters is caused by the high mobility of residents in commercial areas. Because of the high density of buildings and population in the ‘Commercial’ zone, taxis are not usually allowed to wait for passengers. This reduces the number of HO-LD clusters to some extent. A similar phenomenon is found in the ‘Industrial’ areas. This is because most industrial areas are located in the suburbs of Beijing, and taxis tend to leave soon after dropping off (because it is difficult to find a fare in areas far from the city center), which causes the number of HO-LD clusters to be consistently less than that of LO-HD.

To sum up, in the morning, the proportion (i.e., more HO-LD and fewer LO-HD and HO-HD) of three types of cluster for ‘Commercial,’ ‘Public facilities,’ ‘Industrial,’ and ‘Educational institutions’ shows inverse pattern (i.e., fewer HO-LD and more LO-HD and HO-HD) to ‘Residential’, which indicates the interaction between ‘Residential’ and other land-use types. Nevertheless, in the afternoon, due to ‘more HO-HD clusters for “Educational institutions” ’ and ‘fewer HO-LD cluster for “Commercial” and “Industrial”’ (see explanations above), the inverse pattern is not significant.

We can also study the spatial relationship between transportation centers and HO-HD clusters to verify the clustering result. If the distance between the transportation center and its nearest cluster is less than 500 m, we state that the cluster is related to
Table 2. Numbers and proportions of different clusters between 7am and 10am.

| Cluster type | Residential | Public facilities | Commercial | Educational institution | Sports | Industrial | Railway and airport | Water body | Agriculture and green | others |
|--------------|-------------|-------------------|------------|-------------------------|--------|------------|---------------------|-----------|-----------------------|--------|
| HO-HD        | 55          | 17                | 37         | 12                      | 0      | 6          | 3                   | 0         | 16                    | 7      |
|              | 0.360       | 0.111             | 0.242      | 0.0784                  | 0      | 0.0392     | 0.0196              | 0         | 0.105                 | 0.0458 |
| HO-LD        | 74          | 3                 | 17         | 5                       | 0      | 2          | 0                   | 2         | 13                    | 4      |
|              | 0.617       | 0.0250            | 0.142      | 0.0417                  | 0      | 0.0167     | 0                   | 0.0167    | 0.108                 | 0.0333 |
| LO-HD        | 59          | 26                | 52         | 12                      | 2      | 11         | 4                   | 1         | 18                    | 3      |
|              | 0.314       | 0.138             | 0.277      | 0.0638                  | 0.011  | 0.0585     | 0.0213              | 0.0053    | 0.0957                | 0.0160 |
## Table 3. Numbers and proportions of different clusters between 4pm and 7pm.

| Cluster type | Residential | Public facilities | Commercial | Educational institution | Sports | Industrial | Railway and airport | Water body | Agriculture and green | Others |
|--------------|-------------|------------------|------------|--------------------------|-------|------------|---------------------|-----------|-----------------------|--------|
| HO-HD        | 73          | 41               | 8          | 0                        | 3     | 5          | 0                   | 0         | 0                     | 5      |
| HO-LD        | 26          | 9                | 15         | 0                        | 2     | 3          | 0                   | 0         | 0                     | 4      |
| LO-HD        | 90          | 8                | 29         | 0                        | 6     | 1          | 0                   | 0         | 0                     | 8      |

HO-HD = HO-HD, HO-LD = HO-LD, LO-HD = LO-HD

T. Pei et al.
the center. The cluster statistics are shown in Table 4. In the morning, 11 HO-HD clusters (a proportion of 0.5) are found to be related to transport centers, and in the afternoon 14 HO-HD clusters (a proportion of 0.583) are the closest to a transportation center. Both the number and proportion of HO-HD clusters are higher than those of HO-LD and LO-HD. This indicates that in a metropolis like Beijing, to save time over a long journey, passengers commonly take a taxi to their nearest transport center, and then transfer to the subway to reach their destination, or take a taxi to their destination after transferring from a transportation center.

The cluster results have been verified by analyzing the relationship between the temporal variation of hot spots, land-use types, and transport centers. Our method effectively locates two-component clusters that may have a completely different meaning from ‘traditional’ ones. For example, an HO-HD cluster does not indicate a shortage of taxis, whereas an LO-HD cluster implies an urgent demand for taxis, although both would be seen as a cluster of destinations using traditional methods (in other words, they are treated as the same type of cluster if using the traditional methods).

### Table 4. Transport centers and their nearest clusters.

| Transport center | Related clusters | Proportion | Related clusters | Proportion | Related clusters | Proportion |
|------------------|------------------|------------|------------------|------------|------------------|------------|
| HO-HD            | 11               | 0.500      | 1                | 0.0556     | 8                | 0.216      |
| HO-LD            | 1                | 0.0556     | 2                | 0.0883     | 8                | 0.195      |
| LO-HD            | 8                | 0.216      | 8                | 0.195      | 8                | 0.195      |

The cluster results have been verified by analyzing the relationship between the temporal variation of hot spots, land-use types, and transport centers. Our method effectively locates two-component clusters that may have a completely different meaning from ‘traditional’ ones. For example, an HO-HD cluster does not indicate a shortage of taxis, whereas an LO-HD cluster implies an urgent demand for taxis, although both would be seen as a cluster of destinations using traditional methods (in other words, they are treated as the same type of cluster if using the traditional methods).

### 6. Conclusion

In this paper, we first defined the concept of a two-component cluster. In such clusters, there are two types of points, at least one of which exhibits clustering. We proposed a method of identifying three types of two-component clusters (i.e., HA-HB, HA-LB, and LA-HB). Our method requires five parameters, namely eps, \(X_h\), \(X_l\), \(Y_h\), and \(Y_l\). Experiments on synthetic data showed that eps should be around 1/4 of the clustering scale of the data. The other thresholds (i.e., \(X_h\), \(X_l\), \(Y_h\) and \(Y_l\)) for separating clusters from noise were determined by setting the significance level under the assumption that points are distributed randomly across the whole domain. Two-component clusters can then be found in the HA-HB, HA-LB, and LA-HB regions of the density domain, which are generated from these thresholds. The method was applied to the identification of clusters of taxicab trip ODs. The significant relationship between the locations of the clusters and their land-use types over different periods verifies the efficacy of the two-component cluster method.

Two-component clustering provides a new insight into point data anomalies, in which the ground truth may be related to two types of points. Compared to traditional clustering methods for only one type of point, the two-component method provides more information about the clusters. In the case study of taxi data, our method was
able to locate places with an urgent demand for taxis, as well as those with a surplus of taxis, by analyzing the information of origins and destinations. This result, although very useful to urban planning, would be difficult to determine based on previous conceptions and methods. Besides trip OD data, this method can also be applied to other two-component data – for example, crime venues and traffic accidents, extreme precipitation and temperature events, and bank and restaurant locations. Moreover, this method can be extended to analyze multi-component data, say, a data set of OD points and Point Of Interests (POIs).

Acknowledgments
This study was funded through supports from the National Natural Science Foundation of China (Project Number: 41171345 and 41231171), a grant from National Key Technology R&D Program of China (No. 2012AA12A403).

Supplemental data
Supplemental data for this article can be accessed here.

References
Andrienko, G., et al., 2009. Interactive visual clustering of large collections of trajectories. In: IEEE symposium on visual analytics science and technology, 12–13 October, Atlantic City, NJ, 3–10.
Ankerst, M., et al., 1999. OPTICS: ordering points to identify the clustering structure. In: Proceedings of ACM-SIGMOD'99 international conference on management data, 1–3 June, Philadelphia, PA, 46–60.
Ashour, W., et al., 2011. Multi density DBSCAN. Lecture Notes in Computer Science, 6936, 446–453. doi:10.1007/978-3-642-23878-9_53
Besag, J.E., 1977. Comments on Ripley's paper. Journal of the Royal Statistical Society B, 39 (2), 193–195.
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, 83–92. doi:10.1016/S0169-7439(01)00111-3
Duczmal, L. and Assunção, R., 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics and Data Analysis, 45, 269–286. doi:10.1016/S0167-9473(02)00302-X
Ester, M., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd international conference on knowledge discovery and data mining, 2–4 August, Portland, OR, 226–231.
Guo, D.S., et al., 2012. Discovering spatial patterns in origin-destination mobility data. Transactions in GIS, 16 (3), 411–429. doi:10.1111/j.1467-9671.2012.01344.x
Janeja, V.P. and Atluri, V., 2008. Random walks to identify anomalous free-form spatial scan windows. IEEE Transactions on Knowledge and Data Engineering, 20, 1378–1392. doi:10.1109/TKDE.2008.96
Jiang, H., et al., 2011. A new hybrid method based on partitioning-based DBSCAN and ant clustering. Expert Systems with Applications, 38, 9373–9381. doi:10.1016/j.eswa.2011.01.135
Kiskowski, M.A., Hancock, J.F., and Kenworthy, A.K., 2009. On the use of Ripley’s K-function and its derivatives to analyze domain size. Biophysical Journal, 97 (4), 1095–1103. doi:10.1016/j.bpj.2009.05.039
Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics and Theory Methods, 26, 1481–1496. doi:10.1080/03610929708831995
Kulldorff, M., et al., 2006. An elliptic spatial scan statistic. Statistics in Medicine, 25, 3929–3943. doi:10.1002/sim.2490
Lee, J., Shin, I., and Park, G.L., 2008. Analysis of the passenger pick-up pattern for taxi location recommendation. In: Proceeding of the fourth international conference on networked computing and advanced information management, Vol. 1, 2–4 September, Gyeongju, 199–204.

Liu, Y., et al., 2012. Urban land uses and traffic ‘source-sink areas’: evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106, 73–87. doi:10.1016/j.landurbplan.2012.02.012

Neill, D.B., 2006. Detection of spatial and spatio-temporal clusters. Thesis (PhD). University of South Carolina.

Pan, G., et al., 2013. Land-use classification using taxi GPS traces. IEEE Transactions on Intelligent Transportation Systems, 14, 113–123. doi:10.1109/TITS.2012.2209201

Pei, T., et al., 2006. A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes. International Journal of Geographical Information Science, 20, 153–168. doi:10.1080/13658810500399654

Pei, T., et al., 2007. Delineation of support domain of feature in the presence of noise. Computers & Geosciences, 33, 952–965. doi:10.1016/j.cageo.2006.11.010

Pei, T., et al., 2009. DECODE: a new method for discovering clusters of different densities in spatial data. Data Mining and Knowledge Discovery, 18, 337–369. doi:10.1007/s10618-008-0120-3

Pei, T., et al., 2011. Detecting arbitrarily shaped clusters using ant colony optimization. International Journal of Geographical Information Science, 25, 1575–1595. doi:10.1080/13658816.2010.533674

Ripley, B.D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability, 13, 255–266. doi:10.2307/3212829

Roy, S., Bhattacharyya, D.K., and Chakraborty, G., 2005. An approach to find embedded clusters using density based techniques. Lecture Notes in Computer Science, 3816, 523–535. doi:10.1007/11604655_59

Sander, J., et al., 1998. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery, 2, 169–194. doi:10.1023/A:1009745219419

Tran, T.N., Drab, K., and Daszykowski, M., 2013. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometrics and Intelligent Laboratory Systems, 120, 92–96. doi:10.1016/j.chemolab.2012.11.006

Wan, Y., et al., 2012. ACOMCD: a multiple cluster detection algorithm based on the spatial scan statistic and ant colony optimization. Computational Statistics & Data Analysis, 56 (2), 283–296. doi:10.1016/j.csda.2011.08.001

Wang, X. and Huang, D., 2009. A novel density-based clustering framework by using level set method. IEEE Transactions on Knowledge and Data Engineering, 21, 1515–1531. doi:10.1109/TKDE.2009.21

Yue, Y., et al., 2011. Identifying shopping center attractiveness using taxi trajectory data. In: International workshop of trajectory data mining and analysis (TDMA), 18 September, Beijing, 31–36.

Zhang, W., Li, S., and Pan, G., 2012. Mining the semantics of origin–destination flows using taxi traces. In: Proceedings of the 4th international workshop on LBSN, 8 September, Pittsburgh, PA.