Front. Earth Sci. 2016, 10(2): 205–221
DOI 10.1007/s11707-015-0525-4

RESEARCH ARTICLE

Mining spatiotemporal patterns of urban dwellers from taxi trajectory data

Feng MAO[1], Minhe JI (✉)[1,2], Ting LIU[3]

1 The GIScience Key Lab, Education Ministry of China, East China Normal University, Shanghai 200241, China
2 Research Center for East-West Cooperation in China, East China Normal University, Shanghai 200241, China
3 Institute of Remote Sensing and Earth Science, Hangzhou Normal University, Hangzhou 310036, China

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

Abstract With the widespread adoption of location-aware technology, obtaining long-sequence, massive and high-accuracy spatiotemporal trajectory data of individuals has become increasingly popular in various geographic studies. Trajectory data of taxis, one of the most widely used inner-city travel modes, contain rich information about both road network traffic and travel behavior of passengers. Such data can be used to study the microscopic activity patterns of individuals as well as the macro system of urban spatial structures. This paper focuses on trajectories obtained from GPS-enabled taxis and their applications for mining urban commuting patterns. A novel approach is proposed to discover spatiotemporal patterns of household travel from the taxi trajectory dataset with a large number of point locations. The approach involves three critical steps: spatial clustering of taxi origin-destination (OD) based on urban traffic grids to discover potentially meaningful places, identifying threshold values from statistics of the OD clusters to extract urban jobs-housing structures, and visualization of analytic results to understand the spatial distribution and temporal trends of the revealed urban structures and implied household commuting behavior. A case study with a taxi trajectory dataset in Shanghai, China is presented to demonstrate and evaluate the proposed method.

Keywords taxi trajectory, spatial clustering, spatiotemporal pattern mining

1 Introduction

The direct consequence of modern information and communication technologies, especially the widespread adoption of such location-aware equipment as smart phones and GPS loggers, is the generation of huge urban travel trajectory data on a daily basis. Characterized by timeliness in collection, high-precision, and wide-range applicability, these data provide powerful support for the study of microscopic individual behavior and macroscopic urban spatial structures and have become the most popular data source for GIS-based urban studies (Ahas et al., 2010). Accompanying advantages of such detailed mobility data is the great challenge in data processing and information mining due to the massive data volume and complexity. For example, such a data set may have millions of points/movements and unknown structures/patterns over space and time (Guo et al., 2012). Therefore, quantitative analysis and modeling of movement trajectory has become a hot topic in both transportation research and geographical information science, since it may help researchers to understand complex spatial and temporal phenomena in cities such as urban residents travel activities, urban population distribution, urban spatial organization and planning (Gao et al., 2013).

Compared with other spatiotemporal data sources (e.g., individual travel survey, mobile communication network data and social media data), GPS-enabled taxi trajectory data has some unique characteristics, which bring researchers both convenience and challenges in data analysis. First, with the improvement of urban infrastructure and widespread use of taxi GPS platforms, the travel survey samples of urban dwellers can cover much wider areas with lower costs. Second, with the 24/7 operation of a taxi, GPS-enabled taxi data may be collected even at nighttime and weekends, an important temporal coverage that any other survey tool may not provide. Generally, a taxi trajectory dataset has a finer spatial and temporal resolution compared with personal handheld GPS logger and mobile network dataset, because of the long-term

Received August 11, 2014; accepted March 27, 2015

E-mail: mhji@geo.ecnu.edu.cn

outdoor operation of taxis and their position data collection at a time interval of seconds. Finally, since the timestamp for a passenger to get on/off a taxi is recorded with the taxi meter system, the individual trips for different passengers along the taxi trajectory are accurately determined, saving the tedious task of trip endpoints identification and avoiding errors resulting from the uncertainty during GPS data collection.

Several technical challenges arise in order to make use of such data. First, the high temporal resolution during data collection results in a humongous volume for each dataset. When each of the 50,000 GPS-equipped taxis in Shanghai, China collects position data at 20-second intervals, for example, there will be a total of over 200,000,000 spatiotemporal points collected each day. It is extremely difficult for researchers to deal with such a large number of points in the analysis of long-term, large-scale spatiotemporal urban travel patterns. Second, due to their high sampling precision and the uncertainty of GPS receivers, these data points do not overlap with each other in space and bear no direct contextual meaning, even for those points recorded by the same taxi at the same location over time. It is an active research area to develop methodologies for aggregating trajectory points into meaningful geographic clusters and regions (Giannotti et al., 2007; Guo et al., 2010). Third, compared with the individual travel survey dataset and social media information, the GPS-enabled taxi data usually lack related text messages and thus provide little semantic information. So, one must be very careful when using such data to mine spatiotemporal patterns, and using multiple data sources for collaborative analysis when necessary is an alternative choice.

In this study, there are two types of information worthy of our major concern, each corresponding to a different analysis scenario: the first is the origin and destination (OD) pairs for each passenger ride, which are extracted from the passenger's getting on/off records, and the second is the movement trajectory of the taxis in the road network. The OD-pair information is a special type of trajectory data that only represents passengers' origin and destination locations but ignores the actual trajectory route. From a taxi passenger's perspective, the most relevant information for a taxi ride is the origin and the destination, while locations along the route are not important. So the relationship between human behavior and its spatial information is a core problem of the OD-pair study. The movement trajectory is another type of mobility information that records the sequence of positional points of moving vehicles, including not only origin and destination, but also all the points in between, which reflect the road traffic conditions (e.g., distance, travel speed, congestion). Therefore, the study of movement trajectories focuses on the process and direction of vehicles moving in the road network and other external information of transportation nodes such as traffic conditions, congestion and the state of intersections (Schäfer et al., 2002). However, little intuitive knowledge can be directly extracted from either OD pairs or movement trajectories due to lack of sufficient semantic information; more effective analysis tools and visualization methods are required to explore important urban dwellers' travel patterns.

This study focuses on the processing of point-based origin-destination pairs contained in the GPS-enabled taxi trajectory data to reveal their inherent urban spatiotemporal patterns. A new approach is proposed to analyze unidirectional origin-to-destination movements, which can be used to discover spatiotemporal patterns such as location characteristics and space-time trends in movements. This approach involves two processing stages: (i) spatial clustering of OD pairs based on a traffic grid partitioning; and (ii) discovering spatiotemporal patterns for urban commuting and jobs-housing balance. In the remaining text of this paper, a review of related research is first provided, which is followed by a detailed discussion of the design and implementation of the proposed method. A case study with a large dataset of taxi trajectories in Shanghai, China is then presented to demonstrate and evaluate the methodology.

2 Related research

Taxi trajectory data can be considered a type of Floating Car Data (FCD), which is a common data source used in Intelligent Transportation Systems (ITS) to obtain traffic information. FCD can be grouped into two types, i.e., active FCD and passive FCD. In the case of active FCD, vehicles are equipped with GPS to record real-time positions; whereas passive FCD are characterized by obtaining traffic information through road beacons (Schäfer et al., 2002). The urban traffic conditions such as the mean speed, traffic flow, and congestion can be analyzed by processing massive amounts of FCD (Kobayashi et al., 1999; Li et al., 2010). Taking into account the cost and scale, GPS-enabled taxis were used mostly as the travel survey vehicles to collect data, which is appropriate for macro-level urban traffic and land use research by deriving large-coverage and long-term characteristics (Li et al., 2011). Other researchers use the taxi trajectory as the most important data source to study human activity and mobility patterns. For instance, Jiang et al. (2009) analyzed trajectories of individuals, which were obtained from taxis of four cities in Sweden, and argued that the mobility pattern is determined by the street layout.

As the raw taxi GPS data lack enough semantic information, trajectory definition and data preprocessing are often needed before carrying out the actual analysis. A number of moving object oriented trajectory models have been proposed, including random way point, random direction, Brownian motion, random walk, and obstacle model for describing human movement (Lee et al., 2009). A semantic model is a common active approach that

combines background geographic information together with trajectory location (x, y), and the concept of stops and moves are often used to facilitate discovering and modeling trajectory patterns (Spaccapietra et al., 2008; Bogorny et al., 2014). Clustering analysis is commonly used to preprocess GPS data and group similar trajectories (Guo et al., 2012). For example, a partition-and-group approach is presented in Lee et al. (2007), which partitions each trajectory to generate sub-trajectories based on geometric characteristics, groups sub-trajectories into clusters, and then classifies trajectories based on the sub-trajectory clusters. Dodge et al. (2009) extracted local and global attributes of trajectories, such as speed, duration, curvature, and other descriptors and use them in classification of trajectories. Tietbohl et al. (2008) identified stops by using a density-based clustering algorithm and introducing a concept of "minimal stop durations", which takes into account the average periodicity of the trajectory time points.

In general, the major clustering methods can be technically classified into the following categories: partitioning, hierarchical, density based, and grid based. All of these methods have advantages and disadvantages in terms of computing resource cost, reliability, complexity and application scope. The k-means algorithm is the most well-known and commonly used partitioning method, which partitions a set of n objects into k clusters by maximizing both the resulting intra-cluster similarity and the inter-cluster difference. Partitioning methods in general are not suitable for discovering clusters with non-convex shapes or clusters with very different sizes, and they are sensitive to noise and outlier data points as well. The hierarchical methods such as BIRCH (Zhang et al., 1996) work by grouping data objects into a hierarchy or "tree" of clusters, which is usually effective for data objects in the form of a hierarchy. However, hierarchical clustering methods can generate low-quality clusters if merge or split decisions are not well chosen (Han et al., 2011). Moreover, both partitioning and hierarchical methods can only find spherically-shaped clusters rather than clusters of an arbitrary shape. The density-based clustering method can resolve this problem (Ester et al., 1996; Ankerst et al., 1999). The main strategy of this approach is to model clusters as dense regions in the data space, separated by sparse regions, and connects core objects and their neighborhoods to form dense regions as clusters. For this reason, the density-based clustering method is usually more suitable for geospatial clustering than other methods. The grid-based clustering method, which is different from the aforementioned methods, takes a space-driven approach by partitioning the embedding space into cells independent of the distribution of the input objects (Wang et al., 1997). The shapes of the resulting clusters are isothetic, which means that all the cluster boundaries are either horizontal or vertical, which may degrade the quality and accuracy of clustering.

Due to the nature of the point data distribution along the taxi trajectory, none of the aforementioned techniques alone is suited for taxi trajectory clustering. Therefore, this paper proposes a novel spatial clustering method by combining grid-based clustering and density-based clustering to improve clustering of taxi OD points for travel pattern identification.

Finding aggregated behavioral patterns and urban spatial structures is the main research direction in taxi trajectory data mining applications. It is reasonable to assume that the frequency of taxi visits is proportional to population density, which means clusters of taxi OD points generally reflect hot spots of urban activities (i.e., places of attraction to the public, known as point of interest- POI) to a certain extent. If proven true, this information can then be used in urban land assessment, urban planning and urban energy assessment as an assistant decision support dataset (Yue et al., 2009; Liu et al., 2012a, b; Zhang et al., 2013). On the other hand, the spatial structure in trajectories provide an intuitive and reliable data support for urban traffic management decision-makers to dynamically perceive the spatiotemporal movement patterns of all taxis on the city scale, and estimate the spatial distribution of traffic flow density, the taxis information in the morning and evening commuter peak periods, and road conditions such as traffic average speed (Gao et al., 2013; Kang et al., 2013). In addition, massive taxi trajectories also provide intelligent navigation services for general car drivers (Yuan et al., 2010, 2013). This paper focuses on the household travel behavior and the daily urban jobs-housing spatial structure. A new methodology is presented for extracting distributional patterns of urban jobs-housing areas and their adjacent relationships from massive taxi trajectories by spatial analysis and visualization approaches.

3 Methodology

3.1 Data Representation

Let $T = \{T_k\}$ be a taxi trajectory dataset and $N = |T|$ be the total number of trajectories so that $1 \leqslant k \leqslant N$, where $T_k = \langle (x_0^k, y_0^k, t_0^k), (x_1^k, y_1^k, t_1^k), \ldots, (x_n^k, y_n^k, t_n^k) \rangle$ is an ordered sequence in trajectory dataset, x_i, y_i is a pair of spatial coordinates, and t_i is the timestamp value with $i = 0,1,\ldots,n$ and $t_0 < t_1 < \cdots < t_n$. The origin and destination pair is the key points in a taxi trajectory, which is essential in travel description and spatial pattern mining. For the convenience of subsequent analyses, let $O = (x_0, y_0, t_0)$ be the origin of trajectory and $D = (x_n, y_n, t_n)$ be the destination of trajectory. Let $OD = \{OD_k\}$ be the OD pair extracted from taxi trajectory dataset T, where $OD_k = \langle O^k, D^k \rangle$, which presents a trajectory that starts at location (x_0^k, y_0^k) and time t_0^k and ends at location (x_n^k, y_n^k) and time t_n^k. Let $M = \langle (x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_{n-1}, y_{n-1}, t_{n-1}) \rangle$ be the sub-

trajectory between origin and destination, representing the move of the taxi. Finally, the T_k can be presented as a tuple of origin (O), destination (D) and move (M), that is $T_k = \langle O^k, M^k, D^k \rangle$.

The purpose of the taxi trajectory spatial clustering method is to group all OD points in the dataset into a number of clusters. Let $C = \{C_j\}$ be a set of clusters, and we denote $N_c = |C|$ as the total number of clusters, so that $1 \leqslant j \leqslant N_c$. Also let $C_j = \{O^p, ..., O^q, D^p, ..., D^q\}$ be a cluster, containing a set of OD points, labeled from p to q. Here we denote $N_{C_j} = |C_j|$ as the total number of OD points in C_j, known as the size of the cluster.

3.2 The OD pairs spatial clustering method based on traffic grid

A key step in this research is the spatial clustering of taxi OD pairs. There are three reasons to perform taxi OD points clustering. First, converting massive number of data points into a limited number of clusters can significantly reduce the data volume, so that other scale-sensitive methods can be used to analyze the data. Second, the taxi OD pairs with rich semantic information can be treated as a natural expression of origin and destination locations, and the polygon formed by clustered OD points represents a meaningful place. Finally, the clustering results can be expediently visualized to discover spatial and temporal mobility patterns.

Due to the nature of taxi trajectory data, it is extremely difficult to conduct a meaningful clustering with conventional methods. First, taxi trajectories of a given time period are highly unevenly distributed within the entire study area, commonly seen as a result of road network density, commuting demand and supply, and population distribution. Second, from a local perspective, the distribution of taxi trajectories is also highly uneven, as they only follow the linear orientation of a given road. Third, points along a taxi trajectory can appear anywhere in a discrete space, thus it is difficult to identify noise, even if a noise point in the conventional sense also has its semantic characteristics and cannot be arbitrarily eliminated. In order to improve the taxi OD points clustering, this paper develops a new clustering method based on a traffic grid using urban road data. The method rests on two observations, first, most of the cluster centers of taxi OD points are either located at or close to road intersections; second, there is a positive correlation between the distribution density of taxi OD points and the distribution density of roads.

The process of the traffic grid-based clustering method is shown in Fig. 1, including the following steps. 1) Finding the road intersections from the road network database. 2) Generating initial Thiessen polygons from the road

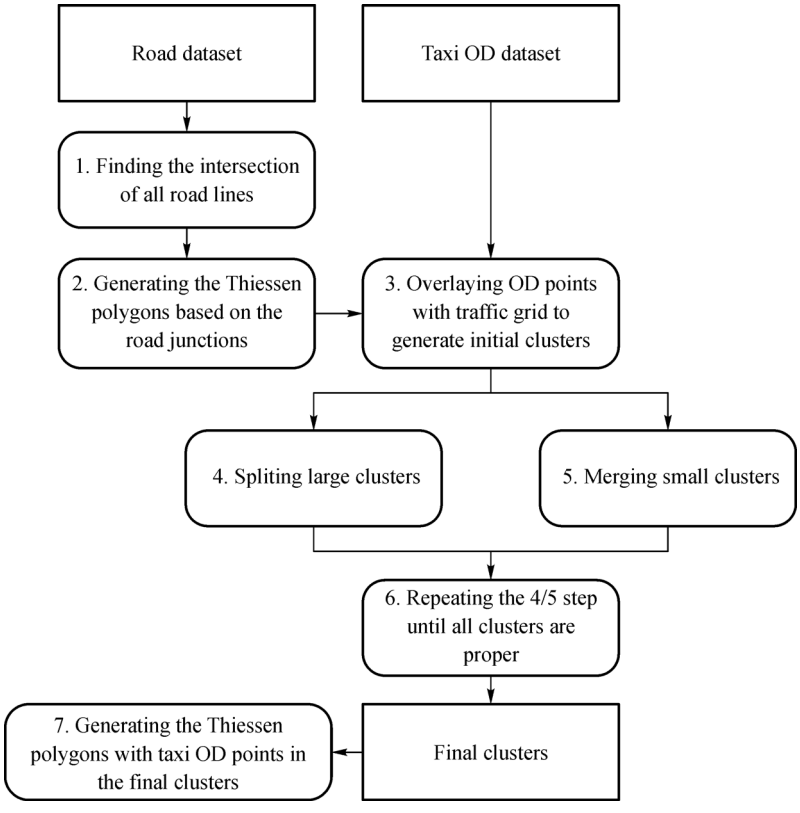

Fig. 1 The overall process of traffic grid-based clustering of taxi OD points.

junctions to divide the entire urban space into multiple zones known as "traffic grid". It is to ensure that there is only one road junction in each zone of the traffic grid. 3) Performing initial clustering to gather the OD points within the same zone as a cluster. 4) Partitioning oversized clusters that exhibit obvious inner density variations into sub-clusters. 5) Merging undersized clusters with their neighbors. 6) Repeating steps 4 and 5 until all clusters have met predetermined criteria. The resulting clusters are designated as final clusters. 7) Generating final Thiessen polygons by dissolving the boundaries of the original polygons that are within the area demarcated by the final cluster. Among all above steps, 4) and 5) are the most critical, since they determine which clusters need to split or merge, and the split and merge strategies and algorithms need to be designed.

For the initial clusters formed by overlaying OD points with traffic grid, two issues need to be resolved. First, if a cluster is so large that the inner density of contained OD points exhibits an obvious spatial variation, it needs to be divided into multiple clusters. Second, if a cluster is too small, it needs to be merged with its neighbors.

In the first case, it is necessary to determine whether a cluster is too large, and whether the distinction among clusters is apparent. Two parameters are introduced for this purpose: the cluster-size alerting threshold θ and the maximum density factor λ. A cluster size greater than θ means that the cluster is too large and needs to be divided. The partitioning process is, however, not necessary for the cases where the OD points of an initial cluster are uniformly distributed and have an appropriate density. To measure the density of a cluster, a metric, known as ***cluster density factor*** which was first proposed by Birant and Kut (2007), is introduced as follows:

Definition 1 **(*Neighborhood*)** Given object p and search distance k, the set of objects whose distance from p is less than k is called the neighborhood of p.

Definition 2 **(*Min / Max Inner-cluster Distance*)** For a given object p, let minimum inner-cluster distance (marked as IDmin) be the minimum distance between p and its neighbor objects; similarly, let maximum inner-cluster distance (marked as IDmax) be the maximum distance between p and its neighbor objects. We then have

$$ID\min(p) = \min\{dist(p,q) | q \in C \cap dist(p,q) < k\}, \quad (1)$$

$$ID\max(p) = \max\{dist(p,q) | q \in C \cap dist(p,q) < k\}, \quad (2)$$

where $dist(p,q)$ is the distance from p to q, which can be the Euclidean distance, Manhattan distance, or any other predefined distance function, C is the set of elements of the cluster, and k stands for the search distance. As shown in Fig. 2, Euclidean distance is set as the distance function. With a given point p (the red point) and search distance k, q_1, q_2, q_3 and q_4 form the neighborhood of p, wherein q_1 is the nearest point from p while q_2 is the farthest point from p, thus $ID\min(p)$ is equal to $dist(p,q_1)$ and $ID\max(p)$ is equal to $dist(p,q_2)$.

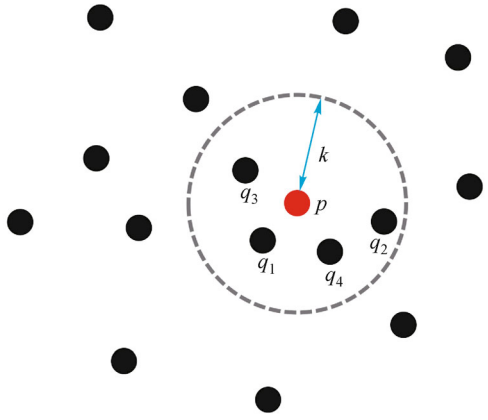

Fig. 2 Example dataset which contains the given point p and its neighborhood.

Definition 3 **(*Inner-cluster Distance*)** Inner-cluster distance is defined as the ratio of the maximum inner-cluster distance to the minimum inner-cluster distance.

$$ID(p) = ID\max(p)/ID\min(p). \quad (3)$$

There is an exception if the neighborhood of p is an empty set when the given k is too small so that p has no neighbors. Then $ID\max(p)$ and $ID\min(p)$ are both equal to 0, and $ID(p)$ is artificially set to 1. And if there is only one point in the neighborhood set of p, $ID\max(p)$ is equal to $ID\min(p)$, so that $ID(p)$ is also 1.

So far, the density factor of a cluster can be defined as follows.

Definition 4 **(*Cluster Density Factor*)** A density factor of a cluster C is defined as Eq. (4).

$$\rho = 1/\left[\frac{\sum_{p \in C} ID(p)}{|C|}\right], \quad (4)$$

where $|C|$ is the size of cluster C. For a nonuniformly distributed cluster with a given k, $ID\max(p)$ is much greater than $ID\min(p)$, forcing $ID(p)$ toward infinity, thus ρ is close to 0. For a uniformly distributed cluster, $ID\max(p)$ is usually close to $ID\min(p)$, resulting in a $ID(p)$ close to 1, thus ρ is close to 1. Therefore, ρ can be used as a good metric to indicate the distribution of a cluster. For a given threshold λ, if $\rho < \lambda$, the cluster is considered nonuniformly distributed and must be split into sub-clusters.

If an initial cluster is too small, it needs to be merged with one of its neighbors. The question then is which neighboring cluster should be chosen. This paper suggests that the smallest neighbor with a road connection is the best candidate for merging. Four steps are involved in searching for this best candidate: whether it is adjacent to

the target cluster; whether it has a direct road connection with the cluster; whether it contains minimal points among all neighboring clusters; whether it has the smallest areal coverage among all neighboring clusters. For the sake of efficiency, the topological structure of the Thiessen polygons is created in advance and stored as an adjacency matrix along with road connectivity information. For the convenience of indexing, this matrix can be represented as a long look-up table, in which the field *Src_Obj* denotes the source polygon, the field *Nbr_Obj* stores the id number of its neighboring polygon, and the field *RdConn* indicates direct road connectivity status between the source and its neighbor. In the process of searching for the smallest neighboring cluster to merge into, the identification and comparison of multiple neighbors of an undersized cluster are made easy with assistance of this table, as information about the number of OD points, the area of the containing polygon, and road connectivity is all available in one place.

A re-clustering algorithm was developed to implement the above splitting and merging tasks, see Appendix A for details.

3.3 Spatiotemporal pattern mining based on taxi trajectory

Commuting is the main purpose of urban residents' travel, which reflects the distribution of jobs and residencies and the relation between them. In this study, we focus on commuting behavior and mining its relevant time-spatial patterns from the taxi trajectories. We first introduce an index known as "job-residential factor (JFR)" to describe the functional characteristics of urban living areas as follows.

Definition 5 **(*Job-residential Factor*)** The job-residential factor of an area is defined as the ratio of the difference between the product of morning-inflows and evening-outflows and the product of morning-outflows and evening-inflows to the product of total morning-flows and total evening-flows. It is expressed as

$$JRF = \frac{inflow_m \times outflow_e - inflow_e \times outflow_m}{totalflow_m \times totalflow_e}. \quad (5)$$

The variable $totalflow_m$ and $totalflow_e$ in Eq. (5) are defined as:

$$totalflow_m = inflow_m + outflow_m,$$

$$totalflow_e = inflow_e + outflow_e,$$

where $inflow_m$ and $inflow_e$ denote the number of incoming taxis in the morning and in the evening, respectively, while $outflow_m$ and $outflow_e$ represent the number of outgoing taxis in the morning and in the evening, respectively.

When a place has a large amount of morning-outflows and evening-inflows, JRF takes values within [−1, 0), indicating that the place tends to have residential characteristics. The closer JRF approaches −1, the stronger residential characteristics the place exhibits. Conversely, a place associated with a large amount of morning-inflows as well as evening-outflows will render JRF within (0, 1], implying that the place is associated with jobs. The closer JRF approaches 1, the stronger the place is job-related. In addition, various spatial visualization methods can be used to reveal urban jobs-housing characteristics from taxi trajectory data, as demonstrated in the next section.

On the other hand, the inter-region connectivity structure hidden in massive taxi OD points can be detected through visualizing the relationship between clusters along with road network orientation. The OD points of each taxi trajectory represent a link between two adjacent regions. It is essential to find those regions that have the strongest connection with other regions or region pairs with highest connectivity in order to understand the taxi mobility patterns and urban spatial structures.

4 Case study

4.1 Study area and data

Shanghai, which is the largest city in China, was selected as the study area in this paper (Fig. 3). There are currently over 50,000 taxi cabs in the city, providing transport services to over 20 million urban residents. As a service standard, taxi cabs in Shanghai are equipped with GPS and pricing systems and periodically transmit such data as vehicle number, timestamp, latitude and longitude, operational status to the dispatching center, forming massive data sources.

The taxi trajectory data adopted in this research is acquired from four major taxi companies in Shanghai. The data collection time was May 15, 2009 0:00–24:00, recording an operating day's trajectories of 18,976 taxis with a total of 56,597,588 location points. After data preprocessing, erroneous records and noisy data were removed from the raw data, resulting in a usable dataset of 9,349 taxis with a total of 35,856,715 records.

After the OD pairs were extracted from the trajectories dataset, further data processing was performed to ensure the data validity in this study, including deleting OD records that have a travel time less than 30 seconds or more than 180 minutes. The former case is mainly caused by the driver's misoperation, whereas the latter is mostly caused by missing GPS signals for a long period of time. Another kind of OD record being considered as invalid are those related to inter-provincial travel, i.e., either the origin or the destination is located beyond the administrative boundary of Shanghai. After removing all invalid records, a final total of 418,314 OD records were obtained (Fig. 4). Figure 4(a) displays the mere distribution of all these OD points, which is highly correlated with the road network patterns. Figure 4(b) visualizes these points with a street-based

Fig. 3 Study area: Shanghai, China with 17 districts and a total of nearly 50,000 taxis in daily operation.

aggregation, revealing that the central portion of the city has the highest density of OD points, with several OD concentrations at a much smaller scale in the northwest and west parts of the city. This pattern is highly compatible with the spatial variation of socioeconomic activities in the city.

The next step is to generate a Thiessen-polygon based traffic grid system for initial OD point clustering. The grids were generated from 5,207 road intersections, which were extracted from 5,494 road records comprising Shanghai's road network. The Thiessen polygons were then established from the intersection points and treated as the traffic grids. The initial clusters were generated by gathering the OD points that fell within the same traffic grids. The total number of initial clusters was the same as the number of traffic grids (Fig. 5).

4.2 The clustering results and parameter optimization

The initial clustering results may contain oversized and undersized clusters, which need to be further processed using Algorithm 1 presented in Appendix A. The size of the final clusters is controlled using four parameters: threshold of minimum cluster size (ε), warning value of cluster size (θ), maximum density factor (λ), and search distance (k). Parameter ε is used to constrain the size of a cluster to its theoretical minimum, which means that the greater the value of ε, the larger the final cluster's size will be, and vice versa. Similarly, the size of big clusters is controlled by θ. The size of most final clusters can be effectively limited to $[\varepsilon, \theta]$, but there may still be a few clusters with size outside this range. The clusters with a size less than ε are generated when they have already been partitioned but cannot be merged into their neighboring clusters; while the clusters with a size greater than θ result when their density factors do not meet the partition requirements. Overall, there is no optimal value for the parameter ε and θ, for the inputs depend on the actual requirements of clustering resolution.

Figure 6 represents the final clusters and their details

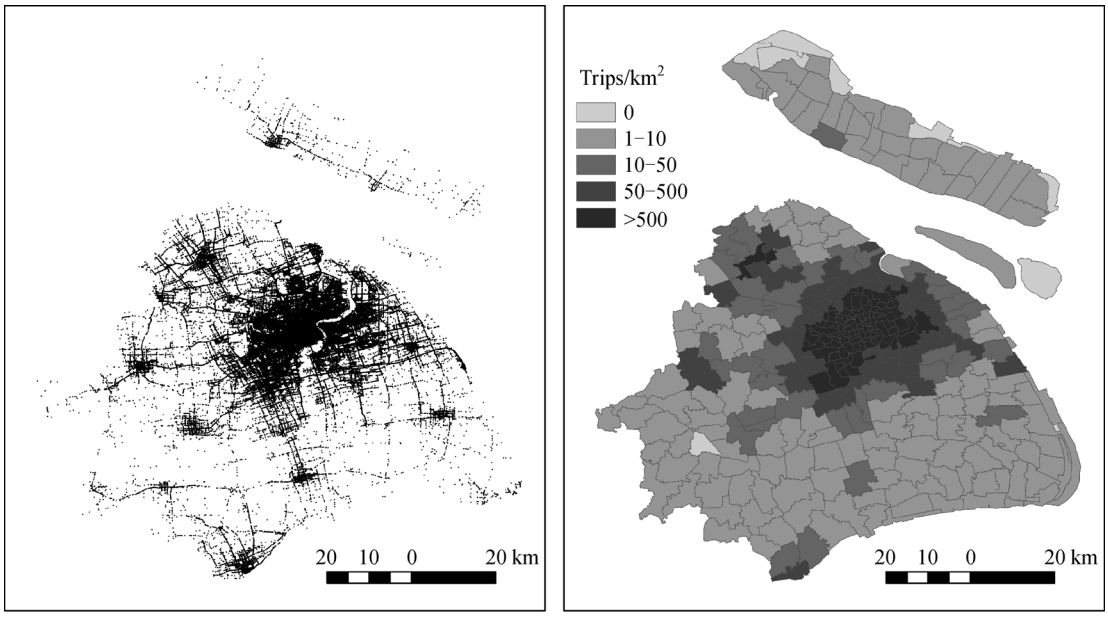

Fig. 4 The spatial distribution of taxi trajectories in Shanghai. (a) All valid taxi OD points in Shanghai. (b) The taxi travel density of each counties.

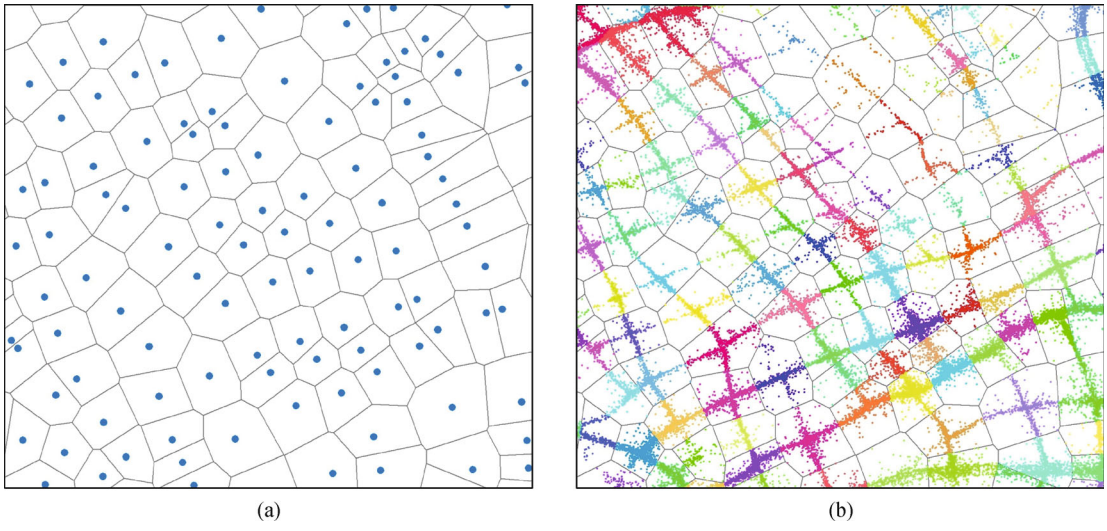

Fig. 5 An example of initializing clusters. (a) The intersections and traffic grid. (b) The initial clusters generated by traffic grid.

with the same λ and k (λ= 0.03, k = 100) and different ε and θ (ε = 100, θ = 200 and ε = 250, θ = 500). The re-clustering for the latter resulted in 2,850 final clusters from 5,207 initial clusters, whereas the number of final clusters for the former is much smaller. There is no significant difference in shape between the two clustering results. With the same λ and k, clusters of a larger ε and θ are derived by merging clusters of a smaller ε and θ, so that the number of clusters is less and the size of single cluster is larger. As a result, a small change in ε and θ does not affect the final clustering results.

On the other hand, how to set λ and k is a quite complicated task. As implied in Algorithm 1, the density factor of each cluster is strongly correlated with k, the search distance for accounting OD points within the cluster. Some descriptive statistics (i.e., the mean, the median, and the quartile) for each initial cluster's density factor were generated to study this correlation, as shown in Fig. 7. The mean density factor for each k is much larger than the median, which is mainly due to the fact that the density factor of several clusters equals 1 as they are small and the distance between points is quite disperse. To some extent, the median can better reflect the overall situation. Furthermore, this relationship seems to follow a power

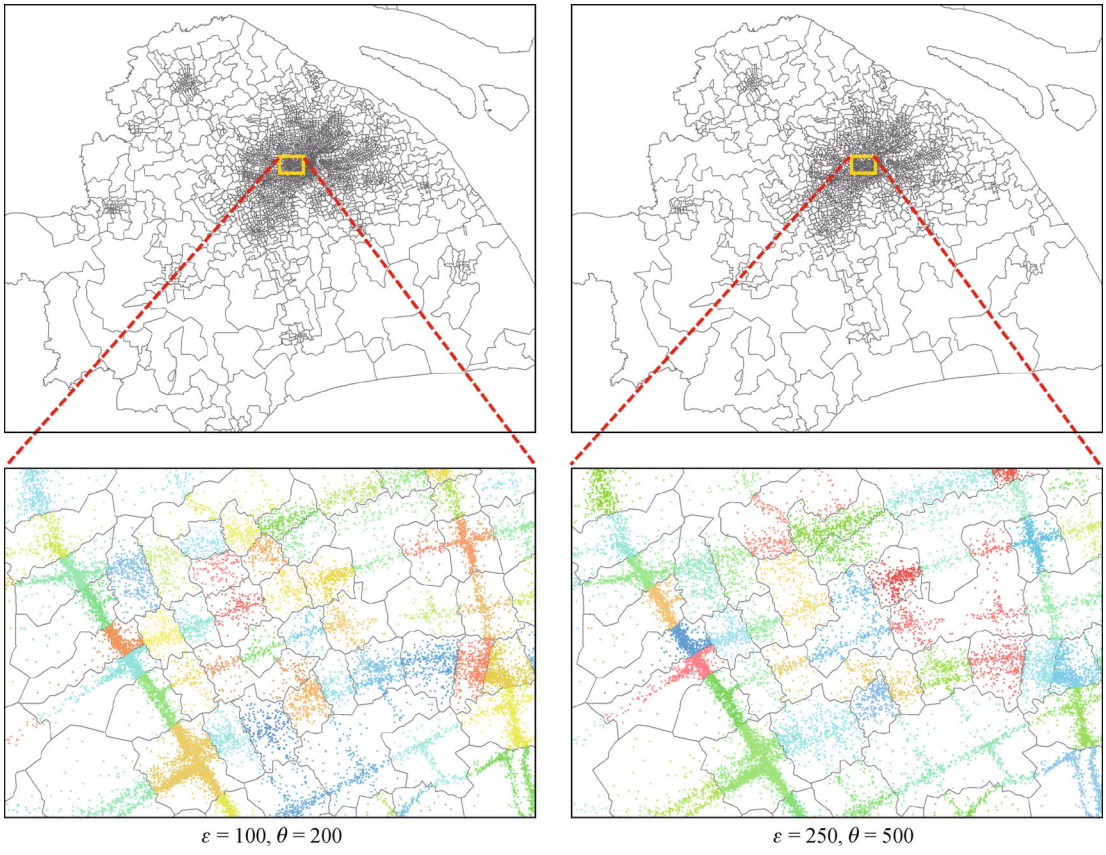

Fig. 6 Comparison of clustering results with different ε and θ.

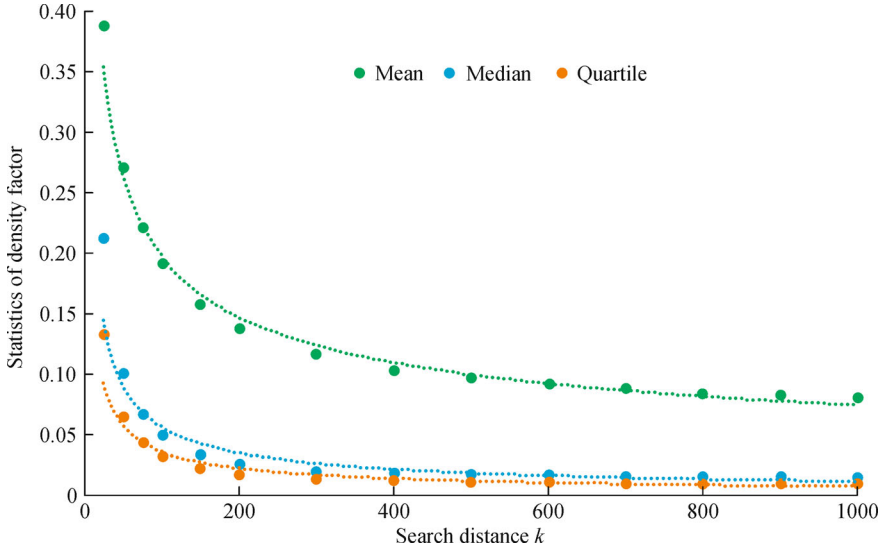

Fig. 7 The mean, median and quartile of density factor of all initial clusters with different k.

distribution characterized by a very steep drop of density for a change in k in the beginning and rapidly leveling off when k reaches 200.

This research used power functions to fit k with such density factor statistics as mean, median, and quartile. The regression models were established as follows:

$$y_1 = 1.39x^{-0.423},\ R^2 = 0.989, \tag{6}$$

$$y_2 = 1.28x^{-0.680},\ R^2 = 0.932, \quad (7)$$

$$y_3 = 0.81x^{-0.677},\ R^2 = 0.941, \quad (8)$$

where y_1, y_2, y_3 stand for the mean, median, and quartile density factors, respectively, and x is the search distance k. Taking the derivative of x in Eq. (6), the first derivative of the search distance versus the mean density factor is obtained. As shown in Fig. 8, when $k > 200$, the impact of search distance on density factor began to diminish rapidly. Particularly, the impact after $k > 500$ was extremely weak. The analysis on median and quartile density factors also led to the same observation. This means that a larger k makes the density factor more stable. On the other hand, it will cost a lot more computing resources when a larger k is adopted. Theoretically, the definition of density factor dictates a four-time increase of points when the search distance is doubled. To obtain the density distance of a cluster, the distance between any two points in the search area must be calculated. For the time complexity of this process, $O(n^2)$, the performance loss can be significant due to the increasing k value. Based on the aforementioned analysis and the actual data of study area (the average area of traffic grids in the study area is 1.26 km^2, so if the traffic grids are considered as approximate circles, the average radius of which is about 600 m), it is reasonable to suggest a k value between 100–300 m for this study. After the k value is determined, maximum density factor λ can be assigned to the median or the quartile of population. Half of large initial clusters will be reclustered when the median is assigned to λ, while three quarters of large initial clusters will be reclustered when the quartile is assigned to λ.

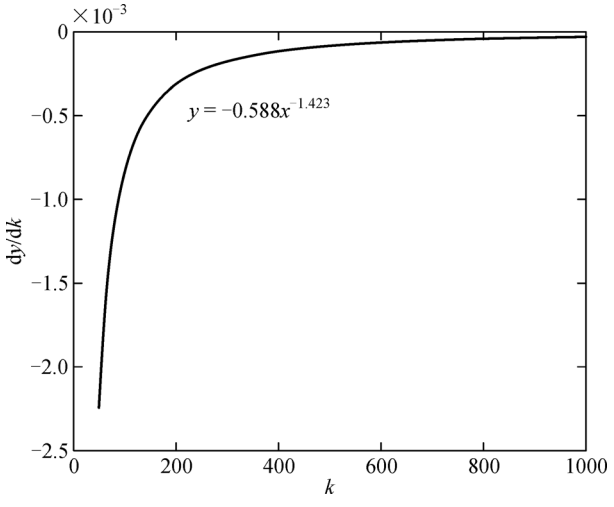

Fig. 8 The first derivative of search distance with respect to the mean of density factor.

4.3 The Urban Jobs-Housing Patterns

Journeys between jobs and residency comprise the majority of urban travels during the weekdays. When these journeys occur is also dictated by the institutional factors such as work schedules conventionally adopted by employers. According to the fifth Shanghai residents travel survey in 2009, most urban working dwellers commute between 6 and 9 am and 4 and 7 pm, respectively, and the peak traffic hour is from 8 to 9 am and from 5 to 6 pm. For a single commuting duration using taxi as the mode of transportation of about 30 minutes, the research extended 30 minutes before and after the commuting concentration travel periods as the job-residential factor's time period; i.e., 5:30−9:30 a.m. as the morning interval, 3:30−7:30 p.m. as the evening interval. The taxi trajectories during these two time periods account for 43% of the total.

In this study, the inflow, outflow, and total flow of each cluster were measured, and the job-residential factor was calculated. Interestingly, the calculated J-R factors of all clusters in Shanghai follow a normal distribution with a mean of −0.05 and a standard deviation of 0.2 (Fig. 9). The normal distribution of all J-R factors is consistent with our assumption, which proves the effectiveness of the factor for description of urban jobs-housing. First, the number of workplaces in Shanghai is roughly equal to the number of residential places (the mean is approximately equal to 0), and their distribution of significant degree is also similar. Second, most of the areas in Shanghai do not represent an obvious characteristic of workplace or residential place, for the land use of an area can contain both commercial buildings, industrial land use, and residential land use types. Actually, the scale of the study area has a great impact on the result.

According to Definition 5, the cluster whose J-R factor is less than zero represents the characteristics of a residential place, and the closer to −1 the J-R factor, the more obvious the characteristic is. Conversely, the cluster whose J-R factor is larger than zero represents the characteristics of a work place, and the closer to 1 the J-R factor, the more obvious the characteristic is. In this study, all clusters are mapped into five functional zones according their J-R factor values. The area with J-R factor greater than 0.2 is set to be a work place, the area with J-R factor less than −0.2 is set to be a residential place, the area with J-R factor in [0.1, 0.2] interval is set to be a quasi-workplace, the area with J-R factor in [−0.2, −0.1] interval is set to be a quasi-residential place, while the rest are set to be a neutral zone. The areas with various functional characteristics are visualized in Fig. 10(a), and the spatial distribution of jobs-housing regions is clearly mapped in Fig. 10(b), with all neutral zones, quasi-workplaces, and quasi-residential places being hidden.

Furthermore, adjacent places of the same kind were merged together to form job or residential blocks. These blocks are visualized as colored nodes in Fig. 11, with red representing workplaces and blue, residential places. The size of the node is proportional to the travel density of the place (i.e., the ratio of the times of travel to the area of the

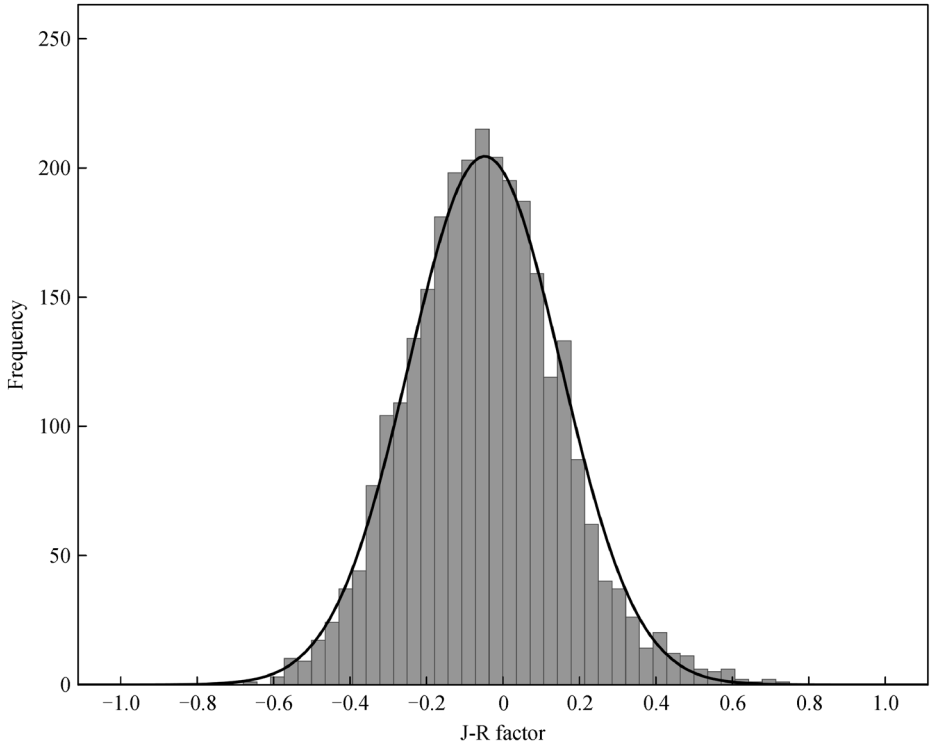

Fig. 9 The histogram of J-R factor.

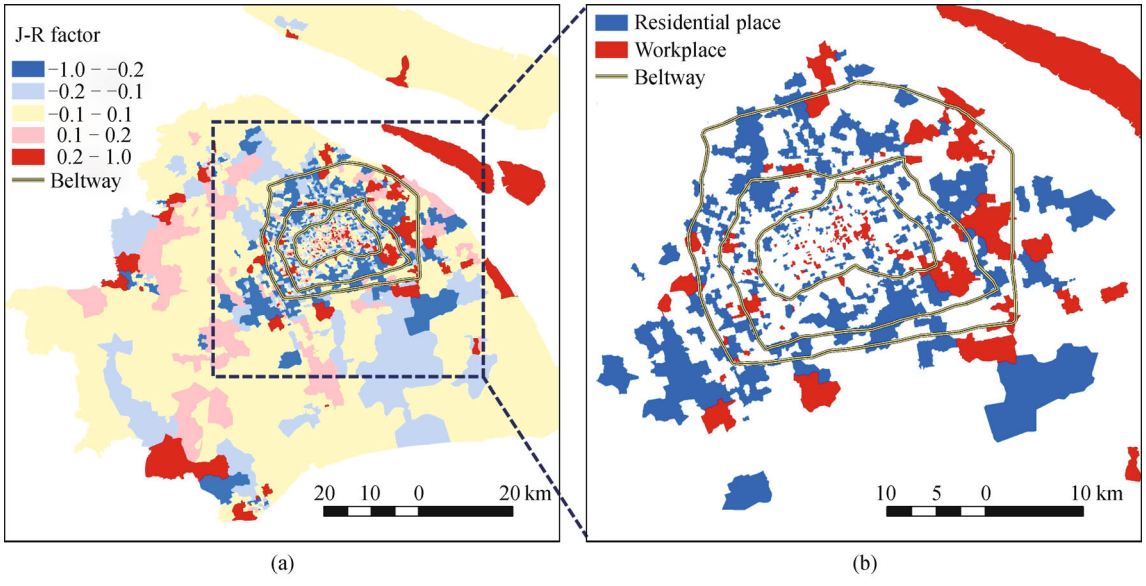

Fig. 10 The J-R factor maps. (a) The choropleth map of J-R factor in Shanghai; (b) the map of jobs-housing region distribution.

region). The edges connecting nodes represent journeys between regions, and a darker and thicker edge shows a more frequent flow. Several significant facts are observed from Fig. 11:

• As the most important transportation hub, the airports and railway stations showed clear characteristics of workplace (i.e., strong inflows in the morning and outflows in the evening), even though the travelers are mostly passengers rather than staff of the airports or train stations.

• The taxi-based traffic pattern seems able to characterize the old versus new urban structures of the city very well. Highly intense taxi-based traffic tends to occur

between newly developed large-sized regions with distinctive and complementary urban functions. These newly developed regions are the result of recent urban planning, and they are featured by either a large pool of jobs or massive residential districts. For example, Pudong New District, including several big residential areas (e.g., Jinqiao) and important workplaces (e.g., Lujiazui and Zhangjiang High-Tech zone), manifests itself as a self-contained jobs-housing system (shown as the yellow box in Fig. 11). Another example is Minhang District (as enclosed in the green box in Fig. 11), including Hongqiao, Xinzhuang, Zhuanqiao and some other areas. Unlike many old urban districts surrounding the city center and CBD west of Huangpu River, which were left over from the first part of the 20th century, both Pudong and Minhang districts were more recently developed, reflecting the urban design idea of a modern megalopolis. Conversely, jobs-housing commuting within the old urban districts located west of Huangpu River seems to rely much less on taxi services, since the well-developed city bus system provides sufficient transportion, and short-distance commuting could be easily accomplished by bicycles and even walking. On the other hand, it is extremely rare to see any jobs-housing related commuting that occurs across the entire city, except for traveling to (or from) airports or railway stations for out-of-town travelers.

- The intensity of taxi-based jobs-housing commuting is not symmetric in terms of morning-leaving and evening-returning flows. The residential-to-job flows in the morning are more intensive and direct, whereas the job-to-residential flows in the evening are more diverse and round-about. This pattern seems to suggest that morning

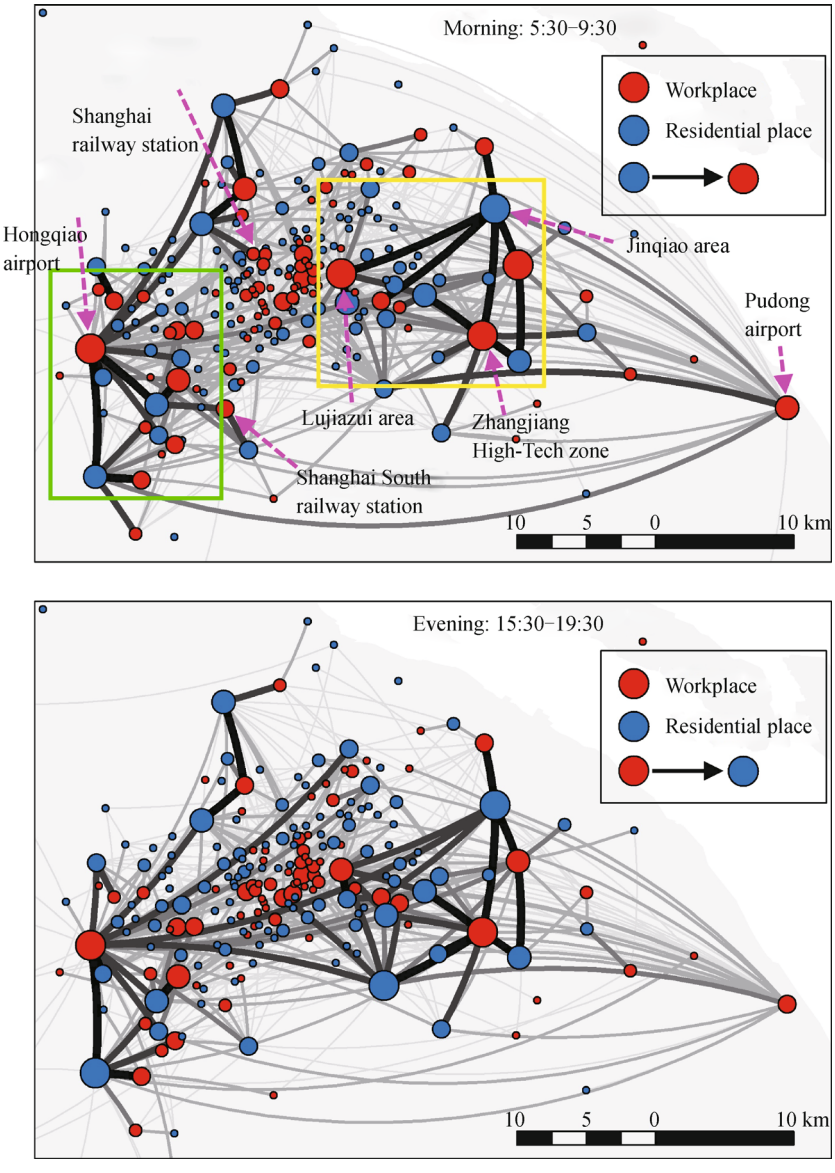

Fig. 11 The travel density and connectivity in the morning and evening.

travelers are constrained by the fixed work schedule and therefore are inclined to go straight to work, whereas evening travelers have more freedom of choice in terms of social activities (e.g., shopping, entertainment, visiting friends) and when and where these activities take place, thus resulting in a more diversified spatiotemporal taxi traffic pattern.

4.4 Result evaluation

The accuracy of J-R labeling generated from the road-intersection and Thiessen-polygon based clustering requires some quantitative assessment before the above interpretation of taxi-travel patterns can be made sensible. Three approaches were employed in this study to evaluate the J-R identification results. The first method compared a 2006 land use dataset of Shanghai to the resulting J-R map (Fig. 10). Due to the highly dynamic nature of suburban areas, we only chose the area enclosed by the outer-ring road as the study object. The land use map was recoded so that commercial, public facilities, public buildings, and industrial land use types were combined into workplace parcels, and all residential land use types were labeled as residential parcels. By overlaying the J-R map with the recoded land use map in a GIS, we obtained a map combining two definitions of jobs-housing land uses. The accuracy of the 2,261 final clusters generated by the traffic grid-based clustering method was assessed using the following equation:

$$LU = (area\_j - area\_r)/(area\_j + area\_r), \quad (9)$$

where LU is the land use type value of the cluster region, $area\_j$ is the area of workplace parcel in the final cluster, and $area\_r$ is the area of residential place parcel in the final cluster. The values of LU are in the interval of −1 to 1. The closer LU approaches −1, the stronger residential characteristics the place exhibits, and the closer LU approaches 1, the stronger the place is job-related. A correlation analysis between cluster LU and J-R factor was made to verify the spatial patterns discovered in this paper. The analytical result showed that the LU value is correlated to the J-R factor with Pearson correlation coefficient of 0.44, and the correlation is significant at the 0.01 level (Fig. 12).

The second evaluation compared the real workplace/residential place and the resulting J-R map. The real workplace/residential place is defined as follows:

$$RW = \begin{cases} true, (area\_j > area\_r) \cap \left(area\_j + area\_r > \frac{1}{2} area\_total\right), \\ false, \text{else} \end{cases} \quad (10)$$

$$RR = \begin{cases} true, (area\_j < area\_r) \cap \left(area\_j + area\_r > \frac{1}{2} area\_total\right), \\ false, \text{else} \end{cases} \quad (11)$$

where RW and RR are the real workplace and the real residential place, respectively, and $area\_total$ is the final cluster's area, $area\_j$ is the area of workplace parcel in the final cluster, and $area\_r$ is the area of residential place parcel in the final cluster. These real places were then compared with the J-R factor results in section 4.4. There are 222 clusters were identified as workplace with J-R factor, and 141 of these are real workplace, with an identification accuracy of 64%, and there are 519 clusters were identified as residential place, and 364 of these are real residential place, with an identification accuracy of 70%.

The third evaluation compared the clustering algorithm proposed in this paper to the traditional DBSCAN clustering method (Ester et al., 1996) and another clustering method in Guo et al. (2012). As the result of

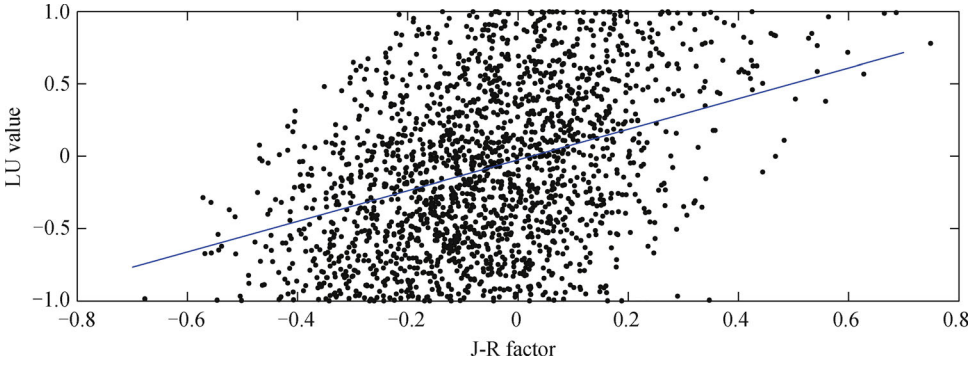

Fig. 12 Relationship of J-R factor and LU values.

DBSCAN clustering is very sensitive to its parameters ε and MinPts, an accurate clustering result depends on selection of proper parameters. After several experiments, we suggested $\varepsilon = 25$ and MinPts = 3 to be ideal values for the dataset. Compared with the results from our algorithm (Fig. 13(a)), many points are marked by DBSCAN as noise points (the black dots in Fig. 13(c)), resulting in many blank areas after dissolving parcels, especially in regions that OD points sparsely distributed. On the other hand, the DBSCAN method grouped a large number of points in a dense area so that we were not able to identify clusters of different point densities.

A measure called silhouette coefficient was used to assess the clustering quality in this paper. For each object o in a data set, the silhouette coefficient of o is defined as

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}, \tag{12}$$

where $a(o)$ is the average distance between o and all other objects in the cluster to which o belongs and $b(o)$ is the

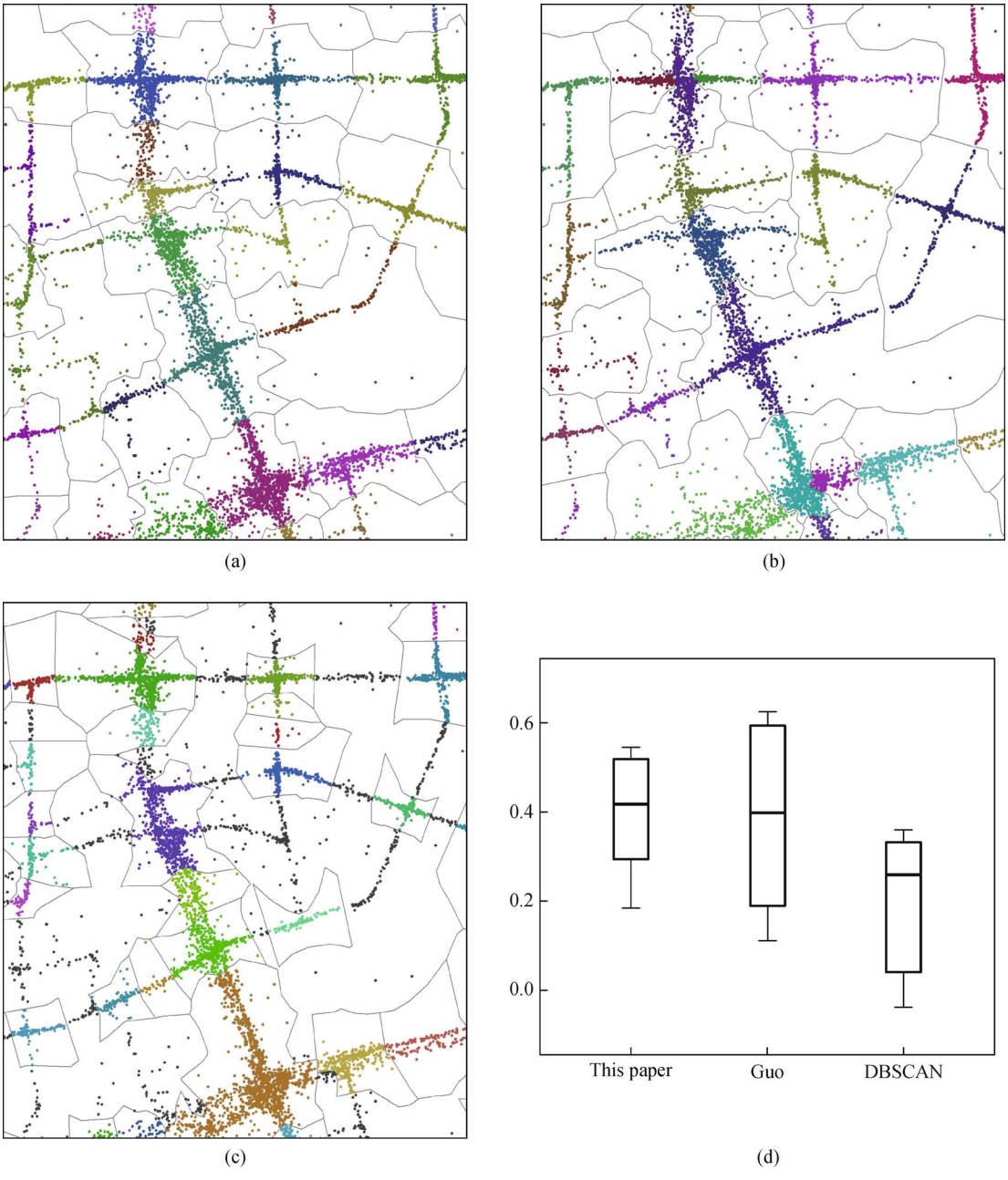

Fig. 13 The comparison of clustering method proposed in this paper and other clustering methods. (a) The result of traffic grid-based clustering method proposed in this paper. (b) The result of clustering method proposed by Guo et al. (2012). (c) The result of DBSCAN clustering method. (d) The boxplot of statistics of silhouette coefficient for above three clustering methods.

minimum average distance from o to all clusters to which o does. The detail of the method could be found in Kaufman and Rousseeuw (1990). In general, the value of the silhouette coefficient is between −1 and 1. The closer the silhouette coefficient value of o approaches 1, the more compact the cluster containing o and the more separated o is from other clusters. Figure 13(d) represented the statistics of silhouette coefficient for each clustering method. The result showed that our method has a better clustering quality on average.

5 Conclusions

We presented a novel approach to discover spatiotemporal patterns of household travel from the taxi trajectory dataset with a large number of point locations. The approach involves three critical steps: spatial clustering of taxi OD based on urban traffic grids to discover potentially meaningful places, identifying threshold values from statistics of the OD clusters to extract urban jobs-housing structures, and visualization of analytic results to understand the spatial distribution and temporal trends of the revealed urban structures and implied household commuting behavior. The case study provides a preliminary evaluation of the presented approach.

The presented spatial clustering method for taxi OD has the following advantages: First, it is more efficient. Second, it can create clusters using traffic grid with different size to adapt to different density surface. This means that there are enough clusters to be generated in dense areas to retain data resolution and fewer clusters to be created in sparse areas to reduce data redundancy, and thus the impact of the uneven distribution of taxi trajectories will be reduced. Third, this method is shape-insensitive, which means it is able to detect clusters of arbitrary shape. Due to these advantages, this method can improve the taxi OD points clustering and be well applied to taxi trajectory data analysis.

In the study case, we focus on commuting behavior and mining its relevant time-spatial patterns from the taxi trajectories. Two methods were presented to help discover these patterns in mobility: First, we used an index known as "job-residential factor" to identify the functional characteristics of urban living areas and visualize spatial distribution of jobs-housing structures. Second, linking each region pair with commuting connection, a commuting network was established to help us to understand the inter-region connectivity structure hidden in massive taxi OD points.

In general, there are several limitations of the method proposed in this paper. First, the method is highly dependent on the road network data. The accuracy and timeliness of the road dataset affects the analysis results. One solution is to use a community-based free editable map such as OpenStreetMap (OSM) to update the road dataset iteratively. Secondly, the proposed method is only applicable to taxi GPS trajectory data, while the applicability and effectiveness for other urban sensor data such as smart phones and mobile communications has not been evaluated yet.

In fact, the problem of sample bias cannot be resolved completely in the analysis of urban space using only taxi trajectory data. Methods combined multi-sensor detectors with traditional household travel survey, such as web-based or smartphone apps, are considered. The approach for functional analysis of urban regions using multi-source urban spatial data is a growing trend in future city management, which will assist future Smart City research, social public service and urban refined management.

Acknowledgements This research is sponsored by the National High Technology Research and Development of China (No. 2013AA12A402), the National Natural Science Foundation of China (Grant Nos. 40771138, 41101371, and 41301484) and the Zhejiang Province Key Scientific and Technological Project (No. 2013C01124). Thanks to Dr. Zhongwei Deng for providing taxi trajectory data of Shanghai, China.

References

Ahas R, Aasa A, Silm S, Tiru M (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. Transp Res, Part C Emerg Technol, 18 (1): 45–54

Ankerst M, Breunig M M, Kriegel H P, Sander J (1999). Optics: ordering points to identify the clustering structure. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Philadelphia: ACM, 49–60

Birant D, Kut A (2007). ST-DBSCAN: an algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1): 208–221

Bogorny V, Renso C, de Aquino A R, de Lucca Siqueira F, Alvares L O (2014). CONSTAnT— A conceptual data model for semantic trajectories of moving objects. Trans GIS, 18(1): 66–88

Dodge S, Weibel R, Forootan E (2009). Revealing the physics of movement: comparing the similarity of movement characteristics of different types of moving objects. Comput Environ Urban Syst, 33 (6): 419–434

Ester M, Kriegel H P, Sander J, Xu X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 1996 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Portland: AAAI 226–231

Gao S, Wang Y, Gao Y, Liu Y (2013). Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ Plann B Plann Des, 40(1): 135–153

Giannotti F, Nanni M, Pinelli F, Pedreschi D (2007). Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 330–339

Guo D, Liu S, Jin H (2010). A graph-based approach to vehicle trajectory analysis. J Locat Based Serv, 4(3–4)183–199

Guo D, Zhu X, Jin H, Gao P, Andris C (2012). Discovering spatial

patterns in origin—Destination mobility data. Trans GIS, 16(3): 411–429

Han J, Kamber M, Pei J (2011). Data mining: concepts and techniques (3rd Edition). Boston: Morgan Kaufmann, 457–458

Jiang B, Yin J, Zhao S (2009). Characterizing the human mobility pattern in a large street network. Phys Rev E Stat Nonlin Soft Matter Phys, 80(2): 021136

Kang C, Sobolevsky S, Liu Y, Ratti C (2013). Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago: Association for Computing Machinery, 1

Kaufman L, Rousseeuw P J (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley & Sons, 28–37

Kobayashi T, Shinagawa N, Watanabe Y (1999). Vehicle mobility characterization based on measurements and its application to cellular communication systems. IEICE Trans Commun, 82(12): 2055–2060

Lee J G, Han J, Whang K Y (2007). Trajectory clustering: a partition-and-group framework. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing: ACM, 593–604

Lee K, Hong S, Kim S J, Rhee I, Chong S (2009). Slaw: a new mobility model for human walks. In: 2009 Proceedings IEEE INFOCOM. Rio de Janeiro: IEEE, 855–863

Li Q, Zhang T, Wang H, Zeng Z (2011). Dynamic accessibility mapping using floating car data: a network-constrained density estimation approach. J Transp Geogr, 19(3): 379–393

Li X, Li X, Tang D, Xu X (2010). Deriving Features of Traffic Flow around An Intersection from Trajectories of Vehicles. Beijing: IEEE, 1–5

Liu Y, Kang C, Gao S, Xiao Y, Tian Y (2012a). Understanding intra-urban trip patterns from taxi trajectory data. J Geogr Syst, 14(4): 463–483

Liu Y, Wang F, Xiao Y, Gao S (2012b). Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. Landsc Urban Plan, 106(1): 73–87

Schäfer R P, Thiessenhusen K U, Wagner P (2002). A traffic information system by means of real-time floating-car data. In: Proceedings of the 9th ITS world congress. Chicago, 1–8

Spaccapietra S, Parent C, Damiani M L, De Macedo J A, Porto F, Vangenot C (2008). A conceptual view on trajectories. Data Knowl Eng, 65(1): 126–146

Tietbohl A, Bogorny V, Kuijpers B, Alvares L O (2008). A clustering-based approach for discovering interesting places in trajectories. In: Proceedings of the ACM Symposium on Applied Computing. Fortaleza: ACM, 863–868

Wang W, Yang J, Muntz R (1997). STING: a statistical information grid approach to spatial data mining. In: 23rd International Conference on Very Large Data Bases. Athens, 186–195

Yuan J, Zheng Y, Xie X, Sun G (2013). T-Drive: enhancing driving directions with taxi drivers' intelligence. IEEE Transactions on Knowledge and Data Engineering, 25(1): 220–232

Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y (2010). T-drive: driving directions based on taxi trajectories. In: GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems. San Jose: ACM, 99–108

Yue Y, Zhuang Y, Li Q, Mao Q (2009). Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: 2009 17th International Conference on Geoinformatics. Fairfax: IEEE, 1–6

Zhang F, Wilkie D, Zheng Y, Xie X (2013). Sensing the pulse of urban refueling behavior. In: UbiComp 2013- Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Zurich: ACM, 13–22

Zhang T, Ramakrishnan R, Livny M (1996). BIRCH: an efficient data clustering method for very large databases. In: SIGMOD Record (ACM Special Interest Group on Management of Data). Montreal: ACM, 103–114

Appendix A

Re-clustering

A re-clustering algorithm as shown in Fig. A1 was developed to implement the splitting and merging tasks. It receives a set of initial clusters, D, as the input, and generates a set of final clusters, C, as the output. In order to save the final clusters, the algorithm initializes C as an empty set and sorts D in the ascending order so that the process starts from the minimum cluster (i). All the initial clusters in D need to be processed and removed one by one untill D is empty. When D is not empty, the following steps are repeated (ii): First, retrieving the current element D_i from D and compare its size to minimum threshold ε and/or maximum threshold θ. If it is smaller than ε, then the merge action is taken; if it is larger than θ, and the density factor of D_i is larger than threshold λ with search radius k, then split D_i; otherwise, save D_i in C.

In the merging step, the candidate cluster should be found first (iii). The *GetNbrMergeCluster*(D_i) function searches all *Nbr_Obj* values which meet the search criteria of $Src\_Obj = D_i$ and $RdConn = 1$ from the TrafficGrid table, and retrieve the minimum cluster D_k from the search results. D_k is merged with D_i to form a new cluster D', $|D'| = |D_i| + |D_k|$, and insert D' into set D in order (v).

The split process is essentially a clustering procedure, so that in theory any proper clustering method could be applied here. For the sake of simplicity, this research chooses the k-means method and sets the number of clusters to 2 (i.e., split the target cluster into two sub-clusters). To reduce the randomness, the two furthest points in the cluster are taken as the initial cluster centers instead of arbitrarily chosen. The *kmeans* () function returns two sub-clusters D_x and D_y (vi), which are then inserted into set D in order.

In the above split-merge process, a special case of endless looping might be encountered, which means, by the criteria, the two clusters resulting from the split (or

merge) process are merged (or split) again. Thus a special provision is made that two clusters should not be merged if they are generated from the same initial cluster (iv), and vice versa (vi).

```
Algorithm 1. Reclustering of Taxi OD pairs
   // Inputs:
      // D = {D1, D2, ..., Dn} Set of clusters
      // ε: Minimum cluster size
      // θ: Warning value of cluster size
      // λ: Maximum density factor value
      // k: Search distance
      // ε < θ
  // outputs:
      // C = {C1, C2, ..., Cm} Set of clusters

C = {}
Sort(D)                          //(i)
While D != {} Do                 //(ii)
   Di = put(D)
   If |Di| < ε Then
     Dk = GetNbrMergeCluster(Di)                 //(iii)
If Dk and Di are formed from split process of initial
clusters Then      //(iv)
         Push(C, Di)
    D' = Merge(Dk, Di)
    Insert(D, D')                                 //(v)
Else If |Di| > θ and DD(Di,k) < λ Then
    {Dx,Dy} = kmeans(Di,2)                        //(vi)
    Insert(D, Dx)
    Insert(D, Dy)
Else
    Push(C, Di)
```

Fig. A1 The re-clustering algorithm of taxi OD points.