Geocarto International

ISSN: 1010-6049 (Print) 1752-0762 (Online) Journal homepage: http://www.tandfonline.com/loi/tgei20

A Network Distance and Graph-Partitioning-Based Clustering Method for Improving the Accuracy of Urban Hotspot Detection

Pengxiang Zhao, Xintao Liu, Jingwei Shen & Min Chen

To cite this article: Pengxiang Zhao, Xintao Liu, Jingwei Shen & Min Chen (2017): A Network Distance and Graph-Partitioning-Based Clustering Method for Improving the Accuracy of Urban Hotspot Detection, Geocarto International, DOI: 10.1080/10106049.2017.1404140

To link to this article: http://dx.doi.org/10.1080/10106049.2017.1404140

Publisher: Taylor & Francis

Journal: *Geocarto International*

DOI: http://doi.org/10.1080/10106049.2017.1404140

A Network-Constrained and Graph-Partitioning-Based Clustering Method for Improving the Accuracy of Urban Hotspot Detection

Pengxiang Zhao [a], Xintao Liu [a,*], Jingwei Shen [b], Min Chen [c, d, e]

[a] *Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong*
[b] *Chongqing Key Laboratory of Karst Environment, School of Geographical Sciences, Southwest University, Chongqing, 400715, PR China*
[c] *Key Laboratory of Virtual Geographic Environment, Ministry of Education of PRC, Nanjing Normal University, Nanjing, 210023, PR China*
[d] *State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, PR China*
[e] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, PR China*

[*] *Corresponding author's email: xintao.liu@polyu.edu.hk*

A Network Distance and Graph-Partitioning-Based Clustering Method for Improving the Accuracy of Urban Hotspot Detection

Abstract: Clustering is an important approach to identifying hotspots with broad applications, ranging from crime area analysis to transport prediction and urban planning. As an on-demand transport service, taxis play an important role in urban systems, and the pick-up and drop-off locations in taxi GPS trajectory data have been widely used to detect urban hotspots for various purposes. In this work, taxi drop-off events are represented as linear features in the context of the road network space. Based on such representation, instead of the most frequently used Euclidian distance, Jaccard distance is calculated to measure the similarity of road segments for cluster analysis, and further, a network distance and graph-partitioning-based clustering method is proposed for improving the accuracy of urban hotspot detection. A case study is conducted using taxi trajectory data collected from over 6,500 taxis during one week, and the results indicate that the proposed method can identify urban hotspots more precisely.

Keywords: Network space; graph-partitioning-based clustering; hotspot detection; taxi trajectory; spatiotemporal variations.

1. Introduction

Nowadays, taxis in modern cities have been equipped with GPS-enabled devices for both safety monitoring and navigation. These devices record real-time digital trajectories at an unprecedented large scale on a daily basis. Such taxi-based GPS trajectories become increasingly available and provide a good opportunity for us to "sense" how people interact with urban infrastructures and further to understand the mobility pattern as well as its underlying dynamics. Pick-up and drop-off events from the GPS trajectories are actually origin and destination (OD) pairs of the passengers, which represent a time-geographic dimension of human activities to some extent. Because of this, pick-up and drop-off events have been widely used in urban hotspot detection, with broad applications ranging from crime area analysis to transport prediction and urban planning. On the other hand, pick-up and drop-off events are constrained to the road network in comparison with other data sets such as land use and points of interest (POIs). Therefore, in addition to the commonly used cluster analysis based on the Euclidean distance threshold, how to improve the accuracy of urban hotspot detection using such pick-up and drop-off events has been drawing the attention of researchers from different fields around the world.

Spatial clustering is probably the dominant approach to detect hotspots using taxi trajectory data (Yue et al., 2009; Chang et al., 2010; Li et al., 2011; Zhao et al., 2017). The purpose of clustering analysis is to find groups of objects, which are more similar to each other than to those in any other groups or clusters. One of the most popular notions of spatial clusters is to use Euclidean distance to include objects with small distances, which means that Euclidean distance is used to measure similarity among spatial objects. As mentioned above, pick-up or drop-off events are

constrained to the road network space, and it would be oversimplifying to just use Euclidean distance to detect hotspots without taking into account the topology of network space. Further, in most of the existing related studies, drop-off or pick-up events are normally represented as point events, which are defined as the last point or the first point of each occupied trajectory respectively (Tang et al., 2015). Ideally, a pick-up or drop-off event occurs at the exact location where the operation status of taxi changes from "occupied" to "empty" or vice versa. However, the actual drop-off position of passengers is uncertain considering the continuous movement of taxis and the sampling time interval of a GPS device, which is normally one minute or so. Therefore, it is more appropriate to represent a pick-up or drop-off event as linear features between the statuses of "occupied" and "empty". Based on such linear representation, a pick-up or drop-off event is a sort of "probability" event, which could happen on more than one or more adjacent road segments. Based on this concept, a kind of clustering method based on network distance should be proposed to improve the accuracy of hotspot detection. For instance, a railway station of a city is usually an urban hotspot. However, a railway station normally has two or more entrances and passenger volumes at each differ. Therefore, it is significant to identify exact road segments with large passenger flow around the railway station.

In this study, a network distance and graph-partitioning-based clustering method is proposed to detect urban hotspots in the road network space using taxi-based GPS trajectory data. Noticeably, drop-off events are extracted from taxi trajectories and represented as linear features. There are four steps in the proposed method: first is to calculate Jaccard distance as a similarity measure for road segments instead of Euclidean distance; second, to construct a similarity graph based on topological relations and similarity measures among road segments; third, partitioning the similarity graph from the second step into road segment clusters, which will be used for the detection of hotspots; and fourth, calculating the number of drop-off events in each cluster to define the cluster's hotspot intensity. Before conducting the case study, we compare the proposed method with other existing hotspot detection approaches based on point events to verify its effectiveness. After that, we apply this method to taxi-based GPS trajectory data obtained from over 6,500 taxis in one week in Wuhan, China. Based on the generated results, we visually explore and quantitatively analyse the spatio-temporal distribution of urban hotspots and dynamic patterns of people's daily travel. The results demonstrate that the proposed method effectively improves the accuracy of urban hotspot detection using taxi-based trajectory data.

The remainder of this paper is organized as follows: Section 2 briefly reviews related work on urban hotspot detection based on trajectory clustering. In Section 3, taxi-based GPS trajectory data and pre-processing are introduced. Section 4 presents the framework of the proposed method, followed by validation of the effectiveness of this method in Section 5. Finally, summarization and future work are provided in Section 6.

2. Related work

As mentioned above, clustering analysis using trajectory data has been increasingly gaining interest in urban hotspot detection in recent years. For instance, Yue et al. (2009) extracted the pick-up and drop-off locations as geometric points from taxi trajectory data during different periods, which were used to generate clusters to explore time-dependent hotspots using a single

linkage clustering algorithm. Chang et al. (2010) proposed a four-step method to predict the areas with potential passenger demand from contexts and past history, in which clustering methods were applied to identify locations with high passenger density as potential hotspots. Li et al. (2011) clustered pick-up and drop-off events to find hotspots with a grid-based clustering method. The study area was partitioned into grids with equal intervals and the top 99 busiest regions were selected as the hotspots according to the count of pick-up and drop-off events in each region. Pei et al. (2015) proposed a density-based method for identifying two-component clusters, which was applied to cluster origins and destinations of trips from taxi trajectory data in Beijing to detect hotspots. The clustered results were eventually verified by the spatial relationship between cluster locations and their land-use types. Shen et al. (2015) proposed a grid adaptive DBSCAN (GADBSCAN) algorithm to cluster passenger-loading (unloading) points and presented the loading (unloading) hotspot distribution in the map. Zhao et al. (2017) proposed a trajectory clustering approach based on a decision graph and data field for detecting hotspots in taxi trajectory data. However, these studies all represent the pick-up and drop-off events as geometric points in the hotspot detection process. Moreover, the researchers mainly focus on identifying densely distributed areas of pick-up and drop-off points, and neglect the constraints of the road network. Due to vehicle movement being restricted by the road network, the road network's constraints are crucial to similarity measurement of taxi trajectories. For instance, for two drop-off events located on two parallel roads, their network distance may be far greater than the Euclidean distance.

Meanwhile, research on hotspot detection methods in network space has substantially progressed during the past years. It has mainly focused on the fields of traffic crashes and POIs (Xie and Yan, 2008; Okabe et al., 2009; Xie and Yan, 2013; Mohaymany et al., 2013; Nie et al., 2015; Rui et al., 2015; Yu et al., 2015). In regard to the spatial cluster of discrete events in network space, kernel density estimation (NKDE) and local indicators of spatial association (LISA) may be the two most typical methods. For instance, Xie and Yan (2008) presented a novel network kernel density estimation method to estimate the traffic accident events. The results indicated that the new NKDE is superior to standard planar KDE for analysis of traffic accidents. Okabe et al. (2009) developed a kernel density estimation method for assessing the density of point events in a network and applied it to the density estimation of traffic accidents on streets to identify 'hotspots' of traffic accidents. Rui et al. (2015) utilized network kernel density estimation and network K-function to explore retail service hotspots and the spatial clustering patterns of a local retail giant in Nanjing, China. Furthermore, Tang et al. (2015) proposed a novel Network Kernel Density Estimation method for Linear features (NKDE-L) to analyse the spatial distribution of linear events over network space, which was further used to cluster pick-up events during different periods to study space-time patterns of hotspots. Although NKDE is very useful for network-constrained spatial cluster analysis, the local maximums and boundary effects caused by the derivation of the kernel function is an inevitable problem (Nie et al., 2015). Meanwhile, compared with detection of hotspots, NKDE is better for visualization purposes due to lack of quantitative statistical inference assessment. Moreover, these researches are also sensitive to the length of linear units and geometric searching bandwidth.

Next, local spatial statistics has also been widely used in hotspot detection. For instance, Steenberghen et al. (2004) employed a local Moran index to compute spatial clusters of accidents based on proximity and connectivity characteristics of locations. Moons et al. (2009) identified

accident hotspots using a local indicator of spatial association (LISA), which takes into account the distribution characteristics of road accidents along the network. Shen et al. (2017) utilized Moran's I index to detect the spatial distribution of pickup and drop-off locations and identify statistically significant spatial clusters of hot and cold spots using taxi trajectory data. The results indicate that the number of pickup and drop-off locations gradually diminish from the downtown areas to the outer suburbs. In addition, the network-constrained LISA was proposed to detect large-scale clustering in spatial context constrained by a network space (Yamada and Thill, 2010), and has been used to identify high-risk road segments (Nie et al., 2015). Comparing with clustering algorithms, Moran's I is used to measure the aggregation or dispersion characteristics of spatial objects, and local spatial statistics are mainly used to identify "hot spot" areas and "cold spot" areas, namely, areas of very high or very low values that occur near one another. However, LISA fails to distinguish different hotspot areas, hence the ranges of detected hotspots are normally too large. Clustering algorithms are proficient in identifying different clusters.

Despite a few studies on network-constrained clustering methods of hotspot detection (Steenberghen et al., 2010; El Mahrsi and Rossi, 2012; Han et al., 2015), the related work is still in its early stage in terms of accurate assessment. There is an urgent necessity to investigate a new network distance and graph-partitioning-based clustering method to precisely detect hotspots based on linear representation of pick-up or drop-off events extracted from taxi trajectories.

3. Taxi trajectory data and data preprocessing

Taxi-based GPS trajectory data, which includes a sequence of consecutive geo-referenced coordinates, corresponding timestamps and a set of attributes such as speed, operation status, etc., records taxis' movement in space and time. A taxi trajectory can be denoted by a time series of triples (x, y, t), in which x and y represent the longitude and latitude coordinates of the taxi at time t. Let $Tr_{T_1}^{T_2}(i) = \{(x, y, t) | T_1 \le t \le T_2\}$ denote the trajectory of taxi i from time T_1 to T_2 . In this study, the taxi trajectory data are collected from more than 6,500 taxis operating within the third ring road in Wuhan City, China from 5 May (Monday) to 11 May (Sunday), 2014. The status of taxis is automatically sampled approximately every 60s. The road network and a typical 2-hour trajectory of one taxi in the study area are shown in Figure 1.

For data preprocessing, the main work focuses on extracting drop-off events from the taxi trajectory data in the study area. First, points in the trajectory are matched to road segments using a map-matching method. Second, drop-off events are extracted as follows: for a taxi's drop-off event, it is defined as a sub-trajectory, of which the start and end points are consecutive points with operation status of "occupied" and "empty", respectively. As shown in Figure 2(a), the representation of a drop-off event is displayed as a black bold solid line in the road network space. Figure 2(b) displays the extracted drop-off events at 8:00-9:00, 5 May, 2014, which are represented as green lines. It is straightforward to calculate the number of drop-off events that occurred in each road segment. The road segments without drop-off events occurring are believed to be abnormal elements to remove.

Although the sampling time interval for taxi trajectory data is 60s on average, there still exists loss and incompleteness within several taxis' trace data due to GPS signal, equipment failure, etc. ,

which enlarges the range of a minority of drop-off events. If the distance between two consecutive GPS points corresponding to one drop-off event is too long (i.e., greater than a distance threshold), then the location of the taxi dropping off passengers can't be effectively determined during this period. Therefore, the drop-off events with long distance should be removed through deciding an appropriate threshold. To determine the distance threshold, we conduct statistical analysis on distances of extracted drop-off events in Figure 2(b) using a histogram. As shown in Figure 3(a), the distances tend to be stable around 1,000 meters and the percentage of distances that are less than 1000 meters is 93.8%. Due to the limited speed of vehicles downtown, 40-60 kilometers per hour, travel distance during one minute is normally no more than 1 kilometer. Therefore, it is consistent with the actual situation that we select 1 kilometer as the threshold of distance. Figure 3(b) displays the spatial distribution of preprocessed drop-off events.

4. Methodology

4.1 Framework

This study proposes a systemic framework and detailed implementation method for detecting urban hotspots in network space using taxi-based GPS trajectory data. The overall framework is displayed in Figure 4. The proposed method includes four steps: *Similarity Measurement*, *Construction of Similarity Graph, Clusters of Road Segments Generation*, and *Detection of Hotspots*. A brief description of each step is presented as follows:

- **Similarity Measurement**: The first step is to measure similarity between road segments based on drop-off events that occurred. This paper proposes to use Jaccard distance as a similarity measure, assuming that the number of drop-off events co-appearing in two road segments represents their similarity. The output of this step is a distance matrix containing pair-wise Jaccard distances between road segments, denoted by D.
- **Construction of Similarity Graph**: In this step, a similarity graph is constructed based on topological relations and similarity measures between road segments. We utilize a dual graph to model the similarity graph, in which nodes represents road segments, edges stand for topological relations between road segments, and weights of edges are set by matrix D. The goal of this step is to convert the road segments clustering problem into a graph partition problem. The output of this step is an undirected weighted graph, denoted by G.
- **Clusters of Road Segments Generation**: This step is responsible for partitioning similarity graph G into road segment clusters, which will be used for the detection of hotspots. In this step, an Info map algorithm is utilized to generate clusters of road segments, which simultaneously takes into account the weights of nodes and edges. The output of this step is a collection of road segment clusters, denoted by C.
- **Detection of Hotspots**: After obtaining clusters of road segments, we can detect hotspots by calculating the count of drop-off events in each cluster, which is defined as the cluster's hotspot intensity. Considering that only a minority of clusters have higher intensity, the clusters are classified into hotspots and non-hotspots according to their hotspot intensity

using a head/tail breaks scheme.

Next, we present implementation methods in detail for each step.

4.2 Similarity measure for road segments

In the context of similarity measure, the similarity of road segments can be defined based on the spatial distance between them, e.g., minimum-distance-based, maximum-distance-based, centroid-distance-based, Hausdorff distance (Huttenlocher et al., 1993), etc. The Hausdorff distance is mainly used to measure the distance between two point sets, which refers to the maximum distance of a point in one set to the nearest point in the other set. One road segment can be regarded as a point set. In this study, we focus on how often road segments co-appear in an identical drop-off event (El Mahrsi and Rossi, 2012). We measure the similarity of two road segments by investigating their spatial relations with drop-off events, which means if a linear drop-off event overlaps with these two road segments at the same time. The method used is to calculate the number of concomitant appearances of both road segments in drop-off events. Here, each road segment is mapped to one binary vector. For set D including n drop-off events and road segments sets S, road segment i can be represented as a one by n vector. If drop-off event j is intersected with road segment i, $v_{ij} = 1$; otherwise, $v_{ij} = 0$.

The methodology in this research is based on the Jaccard similarity coefficient proposed by Jaccard (1901), which measures the similarity between two sets by comparing similarity and diversity of the selected sets. Given two n dimensional vectors V_i and V_j corresponding to road segments i and j respectively, the Jaccard similarity coefficient between them can be denoted as follows:

$$Jaccard\left(V_i, V_j\right) = \frac{|F_{11}|}{|F_{01}| + |F_{10}| + |F_{11}|} \tag{1}$$

Where sets satisfy $F_{11} = \left\{k \mid v_{ik} = 1, v_{jk} = 1, k = 1, 2, \ldots, n\right\}$, $F_{01} = \left\{k \mid v_{ik} = 0, v_{jk} = 1, k = 1, 2, \ldots, n\right\}$, $F_{10} = \left\{k \mid v_{ik} = 1, v_{jk} = 0, k = 1, 2, \ldots, n\right\}$, the range of Jaccard similarity coefficient is between 0 and 1. The larger is the number of concomitant appearances of 1 occurring in the same position, the more they are considered similar.

Conversely, the Jaccard distance measures dissimilarity between two sets, which is defined as the difference between one and the Jaccard similarity coefficient. Therefore, the Jaccard distance between road segments i and j can be further defined as Equation (2):

$$Jaccarddistance\left(V_i, V_j\right) = 1 - Jaccard\left(V_i, V_j\right) \tag{2}$$

Jaccard distance has been widely used in text clustering as an alternative to traditional Euclidean distance to deal with objects with multidimensional attributes (Ferdous, 2009; Patel, Vaishali and Rupa, 2012). In this study, we further extend Jaccard distance to network space to measure similarity between road segments.

4.3 Constructing similarity graph

This section describes the construction of a similarity graph based on road segments. We model the similarity graph using an undirected, weighted graph $G=(V,E,W)$. Graph G is composed of nodes and edges, in which V is the set of all nodes, E is the set of edges connecting nodes, and W stands for weights of edges. Primary graph and dual graph are two widely used modelling methods in network analysis. Considering that road segments are used as research units in this study, we select dual graph to construct a similarity graph. Each road segment is represented as a node in graph G, the topological relations between road segments are mapped to edges, and the Jaccard distance is assigned as a weight to the corresponding edge. Figure 5 displays an example of a similarity graph based on road segments. As shown in Figure 5(a), a simple sketch map of the road network is presented, where each segment corresponds to a road segment. The corresponding similarity graph is shown in Figure 5(b), where nodes stand for road segments and edges represent their topological relations. The weight of each edge equals the Jaccard distance between its two endpoints. The constructed similarity graph considers not only spatial adjacency of road segments but also co-occurrence of drop-off events.

In order to improve operation efficiency of the proposed method, we remove the road segments without occurrence of drop-off events while constructing the similarity graph. Here, we take the drop-off events data during 8:00-9:00, as shown in Figure 3(b), as an example to illustrate the construction of similarity. After measuring the similarity of road segments and removing the road segments without occurrence of drop-off events, the obtained road network including 11,277 road segments is displayed in Figure 6(a). Furthermore, the corresponding similarity graph can be obtained, as shown in Figure 6(b), which is a weighted undirected graph.

4.4 Partitioning similarity graph

Once the similarity graph of road segments is constructed, the next step is to partition it into clusters. In network science, community detection methods can be used to partition an entire network into tightly connected sub-networks, namely communities, which reveal the characteristics of dense connections between edges (Girvan and Newman, 2002). Considering that the constructed similarity graph is weighted and contains a considerable number of nodes, the Info map algorithm is utilized to partition the similarity graph in this study, which performs well for the large and weighted network graph (Rosvall and Bergstrom, 2008).

The Info map algorithm considers the code length of a random walk in a map as the objective function to be optimized, and then community detection in network science is transformed into an information compression coding problem. Specifically, for a given network, the purpose is to minimize the description length of a random walk's movements on the network. In order to understand the movement of random walk on the network, each node is represented by the assigned binary Huffman code. After coding the nodes, the process of identifying communities is equivalent to finding an optimally compressed description path of how information flows on the network. Normally, information flows can be quickly and easily aggregated in a well-connected module. The more the nodes are linked with each other, the more the walker will stay within them

and thus form a community. Hence, the partition with the shortest description length corresponds to the scheme that best identifies the community structure of the network. Given a graph G including n nodes, assuming that the graph is partitioned into m modules, according to Shannon's source coding theorem (Shannon et al., 2002), description path length function can be defined as follows:

$$L(G) = qH(Q) + \sum_{i=1}^{m} p^i H(p^i) \quad (3)$$

Where q stands for the probability that information flow enters each module, $H(Q)$ is the entropy of information moving between modules, $H(p^i)$ is the entropy of information moving within module i, and p^i represents the probability that information flows within module i.

The similarity graph is partitioned using a weighted Info map algorithm, as shown in Figure 7. Various colours represent different clusters. The algorithm simultaneously takes into account the weights of nodes and edges, which corresponds to the number of drop-off events that occurred in road segments and the Jaccard distances between road segments respectively. The modularity of the weighted Info map algorithm reaches 0.946, which is utilized to partition the similarity graph in this study. Eventually, 937 clusters are generated.

4.5 Detecting hotspots from clusters

Once the clusters of road segments are obtained, the next step is to detect hotspots according to the number of drop-off events that occurred in each cluster. For a given cluster of road segments, we define the number of drop-off events that occurred in it as its hotspot intensity. In this study, we propose to detect hotspots from clusters of road segments based on the concept of head/tail breaks. As a new classification scheme, head/tail breaks can be used to identify inherent class and hierarchical structures with heavy-tailed distribution, which reflects the scaling patterns of far fewer large things than small ones (Jiang, 2013; Liu and Ban, 2013; Jiang, 2016). Specifically, the values of geographic objects can be divided into two parts, namely a low percentage of large objects in the head and a high percentage of small objects in the tail, by the average of all values, while the values follow a heavy-tailed distribution (Jiang and Liu, 2012). Considering the proportion of hotspots is far lower than that of non-hotspots in urban areas, following the characteristic of the power law of geographic space, it is appropriate to detect hotspots from clusters according to each cluster's hotspot intensity using head/tail breaks. Although there are a range of data classification methods based on objects' attribute values, such as local Moran's I, quantiles and Jenks's natural breaks, they are unsuitable in this study. For instance, Moran's I divides data objects into four types according to the aggregation degree of each object with surrounding objects, namely high-high (HH), low-low (LL), high-low (HL), and low-high (LH) (Anselin, 1995). If several clusters with high intensity are adjacent to each other, they will be regarded as one hotspot using local Moran's I index. The schemes of quantiles and natural breaks require determining the number of classes and the class intervals first.

Based on this division rule, we measure the distribution of all clusters' hotspot intensity. The clusters are ranked according to their intensity. The ranking is plotted on the x-axis and the

corresponding intensity values on the y-axis, as shown in Figure 8(a). Figure 8(b) displays the cumulative frequency distribution of clusters' intensity and corresponding log-log plot, which indicates that clusters' hotspot intensities approximately exhibit a power-law distribution.

We further examine the hierarchy of the intensity values. The first arithmetic mean (34.4) divides the clusters into two groups: the values greater than the mean and the values smaller than the mean. The former including 248 clusters corresponds to the head and the latter including 689 clusters corresponds to the tail. Moreover, the head part is partitioned into two categories by the second mean (103.9) of these 248 clusters. Likewise, the obtained head part can be further divided until it no longer follows power-law distribution. Eventually, four hierarchical levels are obtained through three partitions, the result of which is displayed in Table 1. Thus, 27 clusters have been detected as shown in Figure 9.

5. Experiment

5.1 Validation of the proposed method

In this section, to verify its effectiveness, the proposed method is compared with three other existing hotspot detection approaches using real trajectory data: 1) single-link clustering (Gower and Ross, 1969; Yue et al., 2009); 2) density based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996; Gui and Yu, 2014); and 3) trajectory clustering approach based on decision graph and data field (TCDGDF) (Zhao et al., 2017). The real trajectory data selected as the experimental data consists of drop-off events collected by over 6,500 taxis within the third ring road in Wuhan at 8:00–9:00 on 5 May 2014. Here, we use a POI data set to validate the proposed method, which includes typical places of interest attracting large passenger flow. The POI data set covers different types of POI, such as railway station, passenger station, business center, hospital, subway station, scenic spot, and so on. Visual inspection of overlap degree between the detected hotspots and POIs is used to examine the accuracy of hotspot detection. If hotspots have a higher overlap degree with POIs, the identified hotspots can be considered more accurate. Hence, POI dataset is suitable as an evaluation baseline.

Figure 10 displays the clustering results. The original drop-off events and drop-off point data are presented in Figures 10(a) and (b) respectively. Figure 10(c) displays the clustering result of the proposed method based on the drop-off events. The proposed method maps the drop-off events to corresponding road segments. Furthermore, road segments are clustered to detect hotspots. Considering that the ranges of urban hotspots are various, several large hotspots are divided into two or three adjacent clusters, such as railway station, commercial zone, etc. Although 27 clusters are generated, 21 hotspots are eventually identified, which are presented with different colours. It can be attributed to various sizes of urban hotspots. Three comparative approaches detect hotspots based on the drop-off point data in Figure 10(b). Figure 10(d) presents the clustering result of TCDGDF, in which parameter σ is set as 0.124 kilometer. This method selects 800 meters as the range of hotspot based on prior knowledge. By comparing the clustering results in Figures 10 (c) and (d), it is found that most of the detected hotspots are identical based on these two approaches. Figure 10(e) displays the clustering result of DBSCAN. The minimum number of points is set as

50 within the radius of 0.4 km. On the basis of this setting, 9 hotspots are detected from the drop-off points. With regard to Single-link, the only parameter involved is distance threshold d, which is set as 0.14 km in this study. While detecting hotspots, the clusters are regarded as noise if the number of drop-off points in each of them is less than 60. It can ensure that there emerges at least one drop-off point in the cluster less than every 1 minute on average, which is dense enough to consider the cluster as a hotspot. The obtained hotspots are shown in Figure 10 (f).

By comparing the results in Figures 10(c)-(f), the clustering results of the other three approaches include situations in which several hotspots are misclassified and omitted. As shown in Figure 10(c), some common urban hotspots are detected based on the proposed method, including railway stations, commercial zones, industrial parks, parks, etc. Due to the selected period being during a workday, a typical hotspot during this period is Optical Valley Software Park, which is the largest software and service industrial park in the central and western regions of China. Tourist attractions and places of entertainment are not presented as hotspots, such as Moshan hill, zoo, forest park, and happy valley. In addition, 8:00-9:00 am is the morning rush hour. In Figure 10(d), TCDGDF can obtain relatively satisfactory results with an appropriate σ value. However, several adjacent hotspots are misclassified, such as Wuhan Square and Zhongshan Park. We conjecture that it can be attributed to the inaccuracy of using drop-off points to represent the locations where the passengers get off. Due to TCDGDF obtaining the clusters by identifying cluster centers firstly, the identified cluster centers determine the number of clusters. For instance, two hotspots close to each other may be recognized as one since some drop-off points cannot accurately reflect passengers' exact getting off locations. In Figures 10(e) and (f), DBSCAN and single-link methods are not required to find cluster centers while detecting hotspots. However, adjacent clusters may be grouped in one cluster or several hotspots may be omitted when the selected parameters are inappropriate, such as at the software park.

We also provide a quantitative analysis and comparison of the results. A series of traditional clustering evaluation indexes are applicable only if the dataset meets with the prerequisite that the classifications for each object are known, such as precision rate, recall rate, F measure, entropy, and so on. However, as a special dataset, trajectory data is difficult to manually classify in advance. Therefore, the selected quantitative evaluation baseline is that the preferred hotspot detection method meets the requirements of high number and density of drop-off events and low size of cluster areas (Chang et al., 2009; Zhao et al., 2017). We leverage three measurements to evaluate the trajectory clustering results: the size of the cluster, the number of drop-off events per cluster, and the density of drop-off events in one cluster. Here, we utilize the total length of road segments in one cluster to represent its size S. For the clustering results in Figures 10 (d)-(f), we defined the size of cluster as the total length of road segments intersected with drop-off points in the detected hotspots. Then, we further calculate the number of drop-off events in each cluster N and obtain the density of drop-off events per cluster ED, which is denoted as follows:

$$ED = \frac{N}{S} \tag{4}$$

The comparison results are shown in Table 2. From the results in Table 2, pros and cons of the hotspots detection methods can be clearly observed. Table 2 indicates that the proposed method exhibits higher density of drop-off events than the other three approaches. Although the average number of events is not the highest for the proposed method, it is closely associated with lower average size of clusters. In addition, the clustering results of lower number and density of

drop-off events for the TCDGDF method is probably attributed to the range of hotspots. Although the TCDGDF method is able to effectively determine cluster centers, it regards 800 m buffer areas around cluster centers as the ranges of hotspots. Actually, the sizes of hotspots should be various according to people's travel intensity in different areas.

Figures 10(c)-(f) shows that all the four methods can reflect the spatial distribution of urban hotspots to a certain extent. However, the proposed method characterizes urban hotspots in a different way compared with the other three methods. Here, we take Hankou Railway Station as an example to further compare the hotspot distribution of the methods at a local scale, as shown in Figure 11. The figure shows that all of the four methods are capable of identifying this hotspot. Furthermore, the proposed method can exactly identify the spatial distribution of the hotspot, the range of the hotspot is accurate to the exact road segments. However, the ranges of hotspot determined by the other three comparing methods correspond to the areas that cover the drop-off points in each cluster. Especially, while drop-off points fail to accurately represent passengers' actual drop-off locations, the obtained range of hotspot will be inaccurate. Moreover, the range of hotspot is sensitive to the parameter settings of each approach.

Figure 11(a) displays the spatial distribution of hotspot in terms of road segments. In Figures 11(b)-(d), the spatial distributions of hotspot are represented as coverage of drop-off points, which are closely associated with the parameters of the methods. TCDGDF, DBSCAN and single-link approaches all involve selecting parameters based on prior knowledge or through multiple attempts to detect urban hotspots using GPS trajectory data. Besides, it can be observed that drop-off events mainly focus on the Second Ring Road segment (within blue ellipse) and branch roads around the railway station for the four results in Figure 11. We conjecture that heavy traffic jams on the main roads in rush hour cause the passenger flow to extend to branch roads since 8:00-9:00 is rush hour.

We further conduct quantitative analysis and comparison based on the single hotspot, as shown in Table 3. From the table we can observe that a higher number and density of drop-off events accompany the proposed method. In addition, the proposed method obtains the hotspot with greatest size. We conjecture that this is mainly due to two reasons. On the one hand, the railway station is a larger hotspot, which attracts tremendous passenger flow every day. On the other hand, a large railway station normally provides two or more entrances in order to facilitate passenger flow. Therefore, the entrances from different directions may simultaneously present the hotspot during one period.

Overall, by comparing the results of the proposed method to those of TCDGDF, DBSCAN, and single-link based on taxi trajectory data, we find the following conclusions: (1) the proposed method more precisely identifies hotspots than the other three hotspot detection methods. (2) The proposed method detects hotspots on the basis of drop-off line data, which more accurately show the locations where passengers leave the taxi. (3) The proposed method does not involve selecting parameters by experience.

5.2 Spatio-temporal variations of hotspots

In this section, we apply the proposed method to analyse the spatio-temporal variations of hotspots based on taxis' drop-off events. The taxi trajectory data during a workday (5 May 2014) are

selected to detect the hotspots and their dynamic patterns. The drop-off events are divided into 24 groups based on the time they occur. The drop-off events of each period are clustered using the proposed method, and the hotspot distributions are obtained. Hotspots of different time spans can be recognized by overlaying the extracted regions on the digital map of Wuhan.

It is predictable that hotspots are mainly located in railway stations and residential communities from 0:00 to 6:00, as displayed in Figure 12(a). Specifically, the Changqing Garden Community is a typical hotspot in this time span. Since it is located far from the central city and has a population of over 100,000, numerous commuters living here have to spend considerable time transferring from/to downtown areas. In addition, the number of hotspots during this period is relatively small compared with that of other periods, being determined by the daily travel volume by taxi during the periods. The overall travel volume during 0:00-6:00 is the minimum in the whole day. Between 6:00 and 12:00, Tongji Hospital is a representative hotspot impacted by its opening hours to some extent since attending physicians begin treatment at 8:00. Besides, most expert physicians are generally absent from hospital on Saturday and Sunday. Hence, more people will go to hospital on a Monday.

Figure 12(b) shows that several commercial and entertainment places gradually become hotspots between 12:00 and 18:00, including Optical Valley, Wuhan Square, Xu Dong, Jianghan Road, Zhongnan Road, Hanzheng Street and Hankou Marshland, etc. Specifically, Optical Valley, as one of the typical hotspots, has a long duration. It is a comprehensive service center that integrates science and technology shows, commodity trading, product development, technical exchanges, business negotiations, cultural entertainment, tourism, etc., which attracts a large amount of traffic flow every day. Between 18:00 and 24:00, commercial hotspots, such as Optical Valley and Wuhan Square, gradually disappear by 22:00. Generally speaking, commercial malls are closed before 22:00. Hotspots gradually transfer to the residential community.

Overall, the distribution of the hotspots during these periods displays a certain spatio-temporal pattern. For instance, some regions are constant hotspots that have a large volume of travel activities for a long time whereas others appear with large traffic flow only during certain periods. This observation is in accordance with the conclusions in previous studies (Yue et al., 2009; Zhao et al., 2017). Constant hotspots are mainly concentrated at Hankou Railway Station, Wuchang Railway Station, Optical Valley and Wuhan Square, which are remarkably influenced by the huge volume of passenger flow. Temporary hotspots are affected by one or more factors, such as an irregular event (e.g. star concert), festival or holiday, etc. In brief, the emergence and decease pattern of urban hotspots also reflects the characteristics and regularity of people's travel activities to a certain extent.

6. Conclusion

The existing methods for studying the detection of urban hotspots using taxi-based GPS trajectory data are mostly based on homogeneous Euclidean space, without considering the constraints of road network configuration. Vehicle movement is a network-constrained mobility process. In this study, we propose a network distance and graph-partitioning-based clustering method for improving the accuracy of hotspot detection using taxi trajectories. The main novelty of the study is to propose the four-step method to detect hotspots based on the linear representation of drop-off

events as sub-trajectory between two adjacent trajectory points with operation status being "occupied" and "empty". This paper first measures the similarity of road segments based on drop-off events using Jaccard distance, then constructs the similarity graph based on road segments, next partitions the similarity graph into clusters, and finally identifies the hotspots using the rule of head/tail breaks. This framework presents the following advantages: (1) it can more precisely identify hotspots; and (2) it does not require determining parameters by prior knowledge in contrast to most existing approaches that are very sensitive to their parameters.

In the experiment, we apply the proposed method to detect urban hotspots using real road network and taxi trajectory data in Wuhan City, China. First, an experiment on taxi trajectory data during a one-hour period is carried out to validate the feasibility of the proposed method. We demonstrate by comparing the results of the proposed method with the results of three classic hotspot detection approaches (i.e., TCDGDF, DBSCAN, and single linkage) that the proposed method can more effectively recognize hotspots. Second, the proposed method is applied to detect urban hotspots and analyse their spatio-temporal variations based on GPS trajectory data. The GPS trajectories of more than 6500 taxis in a workday are selected to conduct the experiment. The dynamic patterns of hotspots during different times of a single day reflect the characteristics and regularity of people's travel activities to a certain extent.

The limitations of our research include: (1) the community detection algorithm used in the step of partitioning the similarity graph can influence the result of detected hotspots. In addition, community detection algorithms normally correspond to various computational costs. Therefore, effective and robust community detection approaches which consider attributes of nodes and edges based on a weighted graph merit further research. (2) This study is focused on refined hotspot detection compared with previous studies. Hence, studies on further improving the accuracy of detected hotspots is inevitable since determining the accurate range of hotspots is also important. Taking a railway station as an example, usually only the main entrance becomes a hotspot, although commonly multiple entrances exist. Hotspots may occur on other entrances on particular days, such as the Spring Festival travel rush. In summary, hotspot regions vary with time, which is significant for analysing the changing process of hotspots including emergence, expansion, shrinkage and decease (Scholz and Lu, 2014).

Acknowledgement

Acknowledges the funding support from several project: an Area of Excellence project (1-ZE24), a startup project (1-ZE6P), and a National Natural Science Foundation of China (No.41622108).

References

Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical analysis, 27(2), pp.93-115.

Chang, H.W., Tai, Y.C. and Hsu, J.Y.J., 2009. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, *5*(1), pp.3-18.

El Mahrsi, M.K. and Rossi, F., 2012, September. Graph-based approaches to clustering network-

constrained trajectory data. In *International Workshop on New Frontiers in Mining Complex Patterns* (pp. 124-137). Springer Berlin Heidelberg.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Ferdous, R., 2009, November. An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In Internet, 2009. AH-ICI 2009. First Asian Himalayas International Conference on (pp. 1-6). IEEE.

Girvan, M. and Newman, M.E., 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), pp.7821-7826.

Gower, J.C. and Ross, G.J.S., 1969. Minimum spanning trees and single linkage cluster analysis. *Applied statistics*, pp.54-64.

Han, B., Liu, L. and Omiecinski, E., 2015. Road-network aware trajectory clustering: Integrating locality, flow, and density. *IEEE Transactions on Mobile Computing*, *14*(2), pp.416-429.

Huttenlocher, D.P., Klanderman, G.A. and Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, *15*(9), pp.850-863.

Jaccard, P., 1901. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Jiang, B., 2013. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer*, *65*(3), pp.482-494.

Jiang, B. and Liu, X., 2012. Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information.*International Journal of Geographical Information Science*, *26*(2), pp.215-229.

Jiang, B., 2016. Head/tail breaks for visualization of city structure and dynamics. *European Handbook of Crowdsourced Geographic Information*, p.169.

Kim, J. and Mahmassani, H.S., 2015. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia*, *9*, pp.164-184.

Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G. and Yang, Q., 2011, March. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on* (pp. 63-68). IEEE.

Liu, X. and Ban, Y., 2013. Uncovering spatio-temporal cluster patterns using massive floating car data. *ISPRS International Journal of Geo-Information*,*2*(2), pp.371-384.

Mohaymany, A.S., Shahri, M. and Mirbagheri, B., 2013. GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. *Geo-spatial Information Science*, *16*(2), pp.113-119.

Moons, E., Brijs, T. and Wets, G., 2009. Improving Moran's index to identify hot spots in traffic safety. Geocomputation and urban planning, pp.117-132.

Nie, K., Wang, Z., Du, Q., Ren, F. and Tian, Q., 2015. A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. *Sustainability*, *7*(3), pp.2662-2677.

Okabe, A., Satoh, T. and Sugihara, K., 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool.International Journal of Geographical Information Science, 23(1), pp.7-32.

Patel, V.R. and Mehta, R.G., 2012. Data clustering: integrating different distance measures with modified k-means algorithm. In *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011* (pp. 691-700). Springer India.

Pei, T., Wang, W., Zhang, H., Ma, T., Du, Y. and Zhou, C., 2015. Density-based clustering for data containing two types of points. *International Journal of Geographical Information Science*, *29*(2), pp.175-193.

Porta, S., Crucitti, P. and Latora, V., 2006a. The network analysis of urban streets: a primal approach. *Environment and Planning B: planning and design*, *33*(5), pp.705-725.

Porta, S., Crucitti, P. and Latora, V., 2006b. The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and its Applications*, *369*(2), pp.853-866.

Rosvall, M. and Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), pp.1118-1123.

Rui, Y., Yang, Z., Qian, T., Khalid, S., Xia, N. and Wang, J., 2016. Network-constrained and category-based point pattern analysis for Suguo retail stores in Nanjing, China. *International Journal of Geographical Information Science*,*30*(2), pp.186-199.

Scholz, R.W. and Lu, Y., 2014. Detection of dynamic activity patterns at a collective level from large-volume trajectory data. *International Journal of Geographical Information Science*, *28*(5), pp.946-963.

Shannon, C.E. and Weaver, W., 2002. The mathematical theory of communication.

Shen, J., Liu, X. and Chen, M., 2017. Discovering spatial and temporal patterns from taxi-based Floating Car Data: a case study from Nanjing. GIScience & Remote Sensing, pp.1-22.

Shen, Y., Zhao, L. and Fan, J., 2015. Analysis and Visualization for Hot Spot Based Route Recommendation Using Short-Dated Taxi GPS Traces.*Information*, *6*(2), pp.134-151.

Steenberghen, T., Dufays, T., Thomas, I. and Flahaut, B., 2004. Intra-urban location and clustering of road accidents using GIS: a Belgian example. International Journal of Geographical Information Science, 18(2), pp.169-181.

Steenberghen, T., Aerts, K. and Thomas, I., 2010. Spatial clustering of events on a network. *Journal of Transport Geography*, *18*(3), pp.411-418.

Tang, L., Kan, Z., Zhang, X., Sun, F., Yang, X. and Li, Q., 2016. A network Kernel Density Estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science*, *30*(9), pp.1717-1737.

Xie, Z. and Yan, J., 2008. Kernel density estimation of traffic accidents in a network

space. *Computers, Environment and Urban Systems*, *32*(5), pp.396-406.

Xie, Z. and Yan, J., 2013. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of Transport Geography*, *31*, pp.64-71.

Yamada, I. and Thill, J.C., 2010. Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. Annals of the Association of American Geographers, 100(2), pp.269-285.

Yu, W., Ai, T., Liu, P., and He, Y., 2015. Network Kernel Density Estimation for the Analysis of Facility POI Hotspots. *Acta Geodaetica et Cartographica Sinica*, 44(12), pp.1378-1383.

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

Zhao, P., Qin, K., Ye, X., Wang, Y., and Chen, Y., 2017. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science*, 31:6, 1101-1127.

Figure 1. Road network and a typical 2-hour trajectory of one taxi in Wuhan city, China.

Figure 2. The extraction of drop-off events. (a) Representation of drop-off events in road network space, (b) extracted drop-off events at 8:00-9:00, 5 May in Wuhan city, China.

Figure 3. Preprocessing of abnormal drop-off events. (a) Histogram of distances of all drop-off events, (b) preprocessed drop-off events data.

Figure 4. Proposed framework for detecting hotspots using taxi trajectory data.

Figure 5. Constructing similarity graph based on an artificial network. (a) A sketch map of road network, (b) the corresponding similarity graph.

Figure 6. Constructing similarity graph based on actual road network. (a) Actual road network map, (b) the corresponding similarity graph.

Figure 7. The results of clusters using Info map algorithm considering weights of nodes and edges.

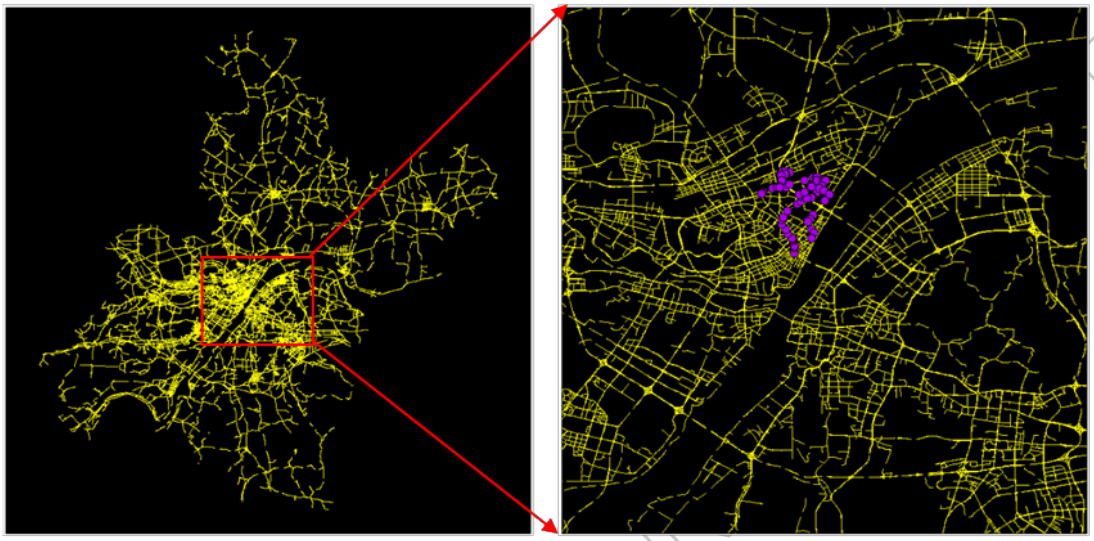
Figure 8. Distribution of clusters' intensity. (a) Rank-intensity distribution; (b) cumulative frequency distribution.

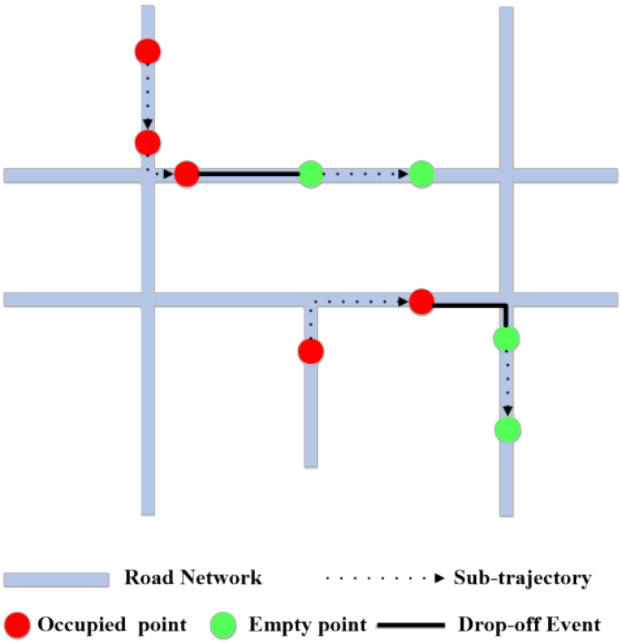
Figure 9. The detected hotspots based on head/tail breaks.

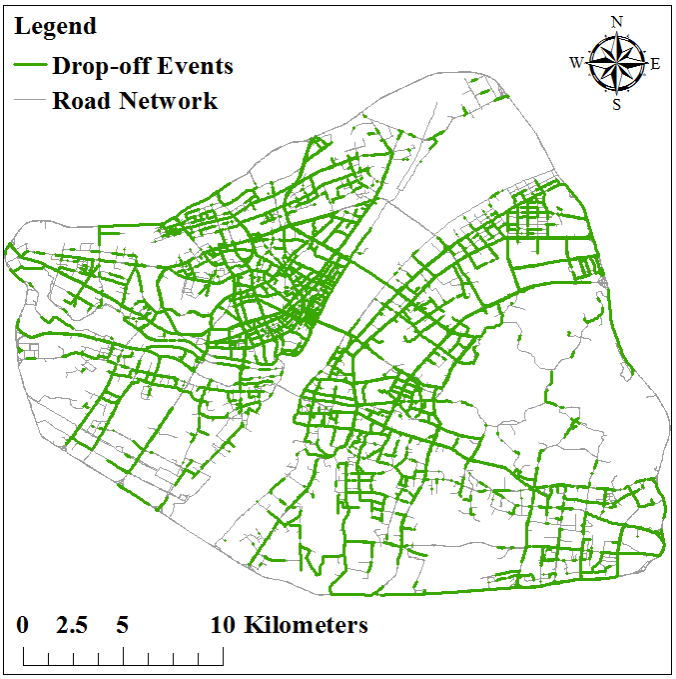
Figure 10. Original activity data and its clustering results. (a) Drop-off events, (b) drop-off points, (c) the proposed method, (d) TCDGDF (σ=0.124km), (e)DBSCAN (Eps= 0.4km , $MinPts$= 50), (f) Single-link (d= 0.14km).

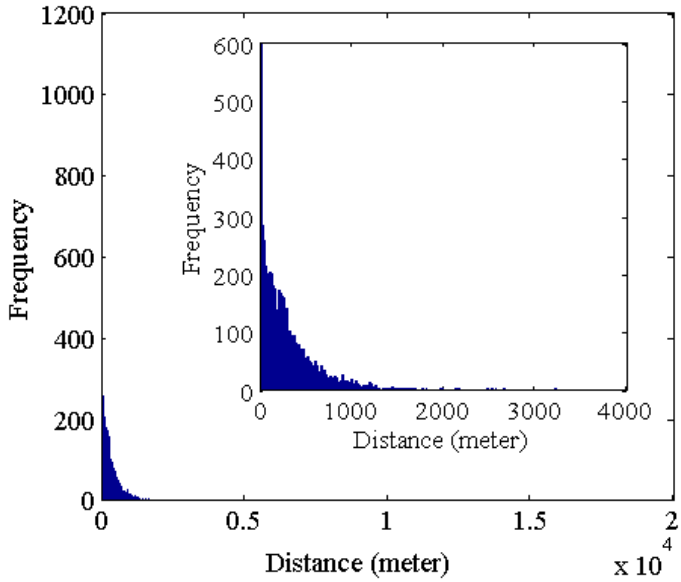
Figure 11. Comparison of the proposed method with other three methods in a local scale. (a) The proposed method, (b) TCDGDF, (c) DBSCAN, and (d) Single-link.

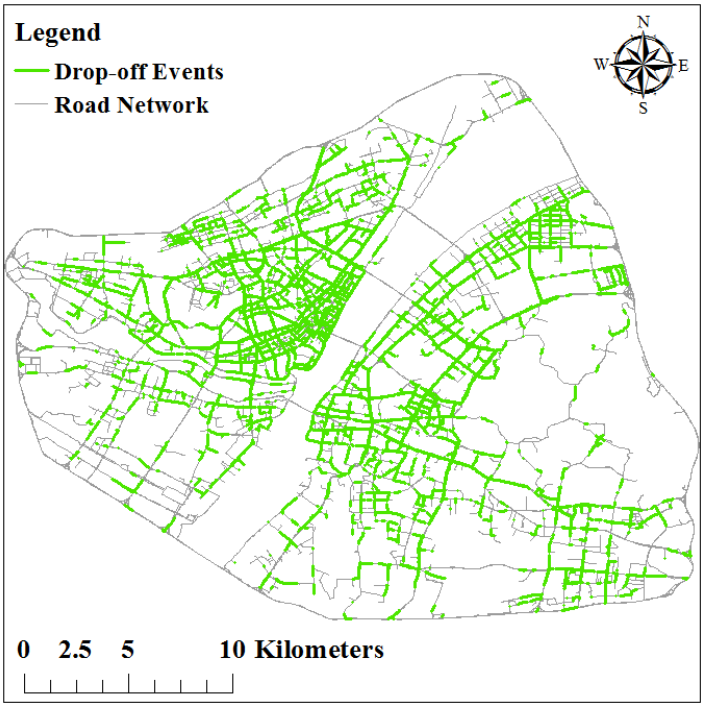
Figure 12. Dynamic patterns of hotspots during a workday (5 May 2014). (a) First half day (0:00–12:00), (b) second half day (12:00–24:00).

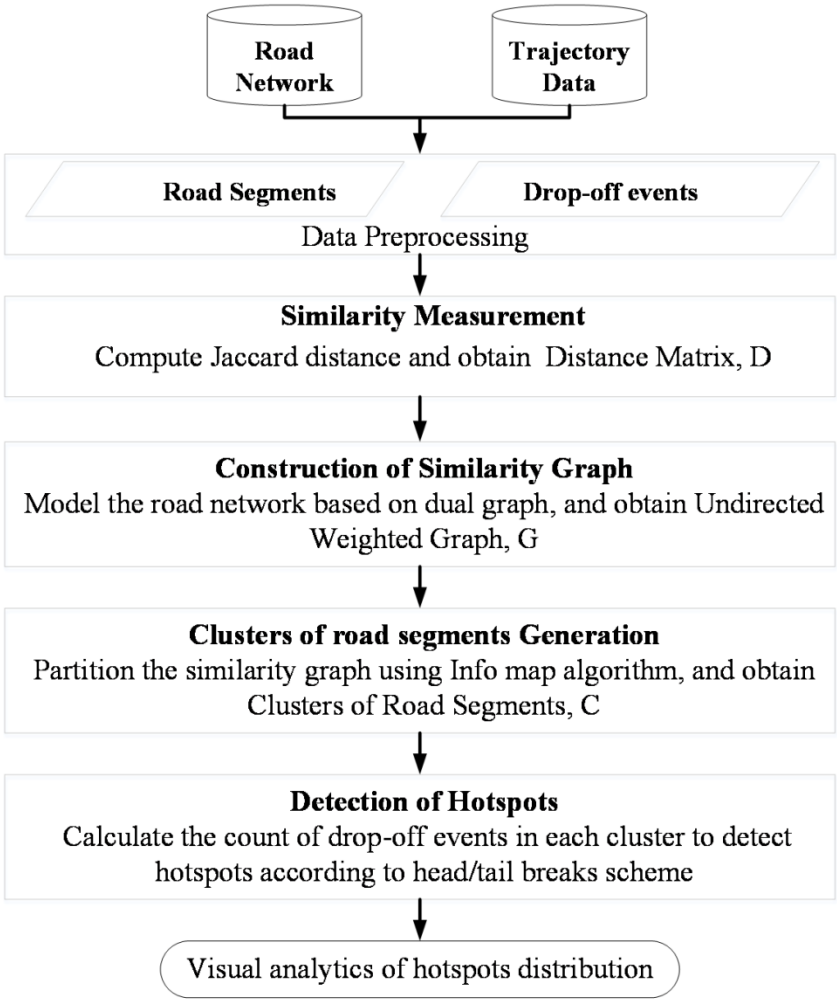


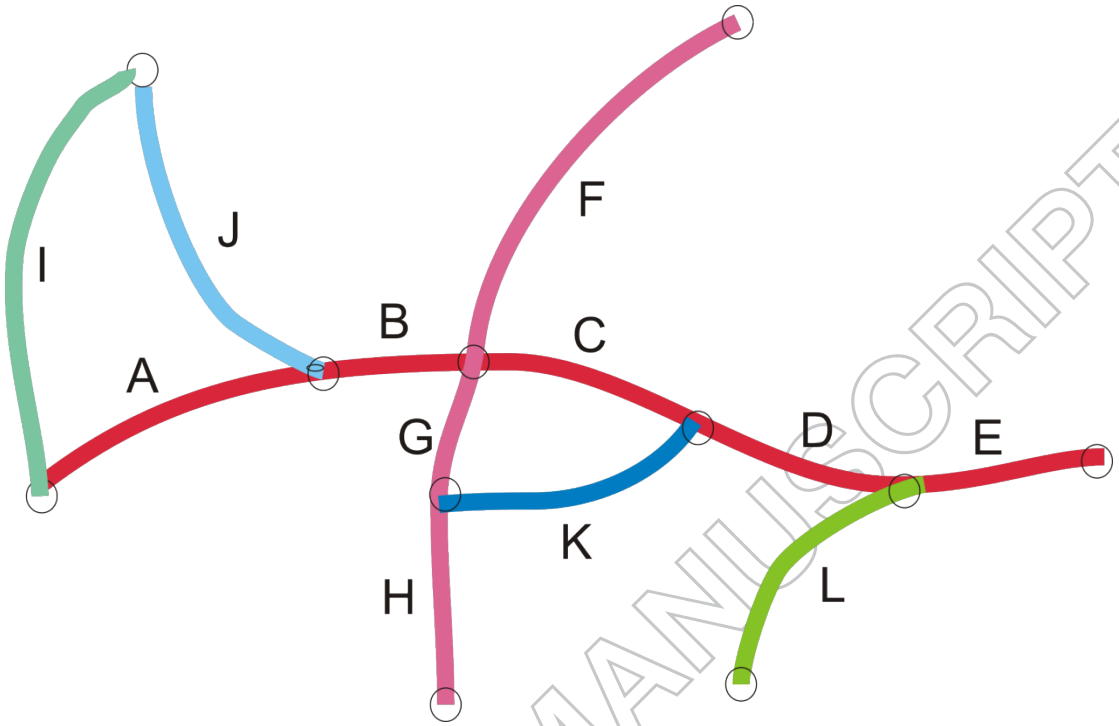


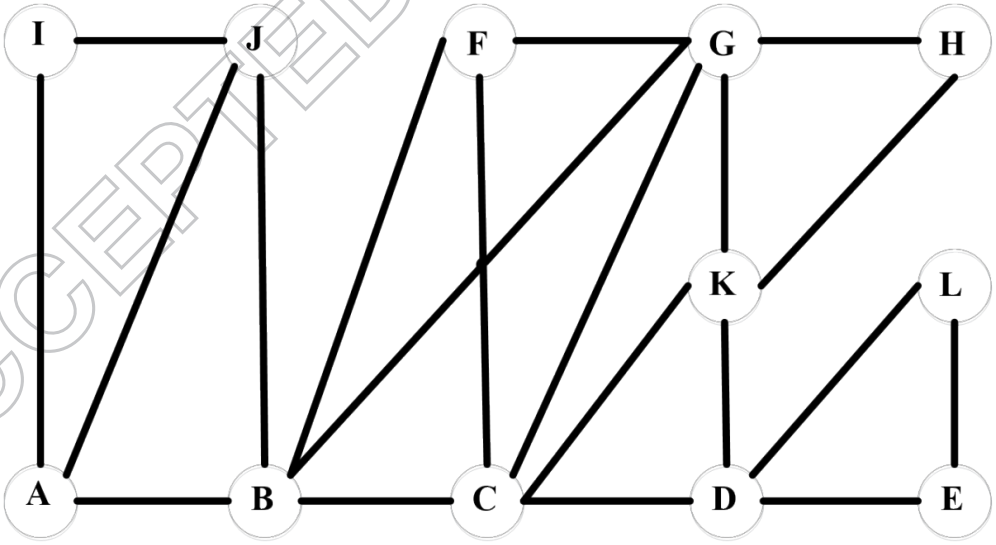


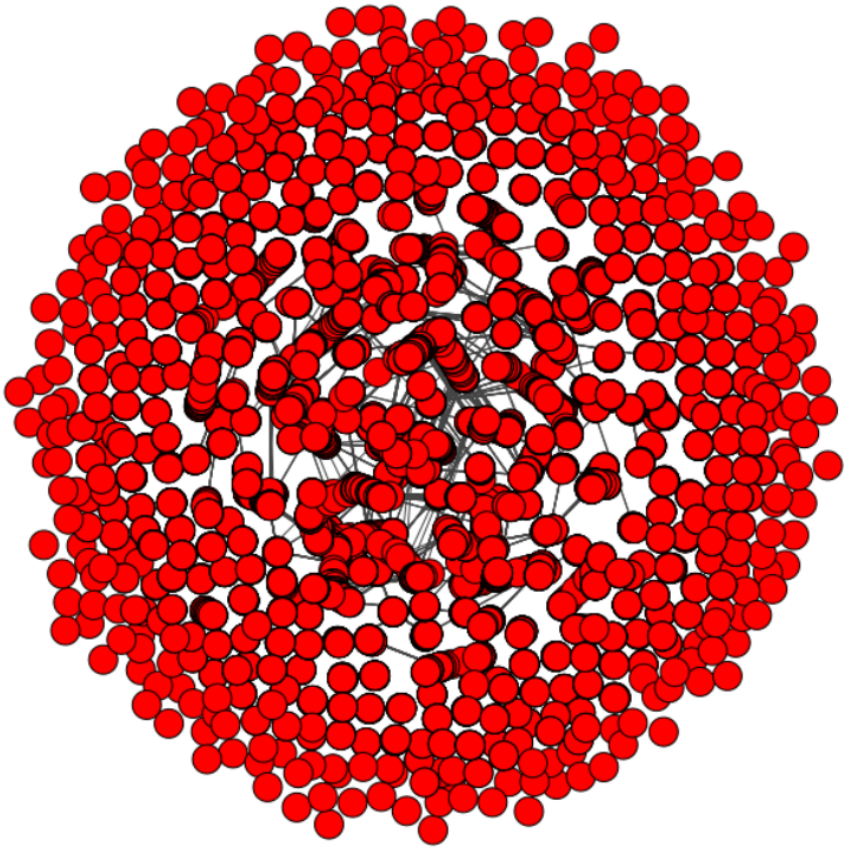


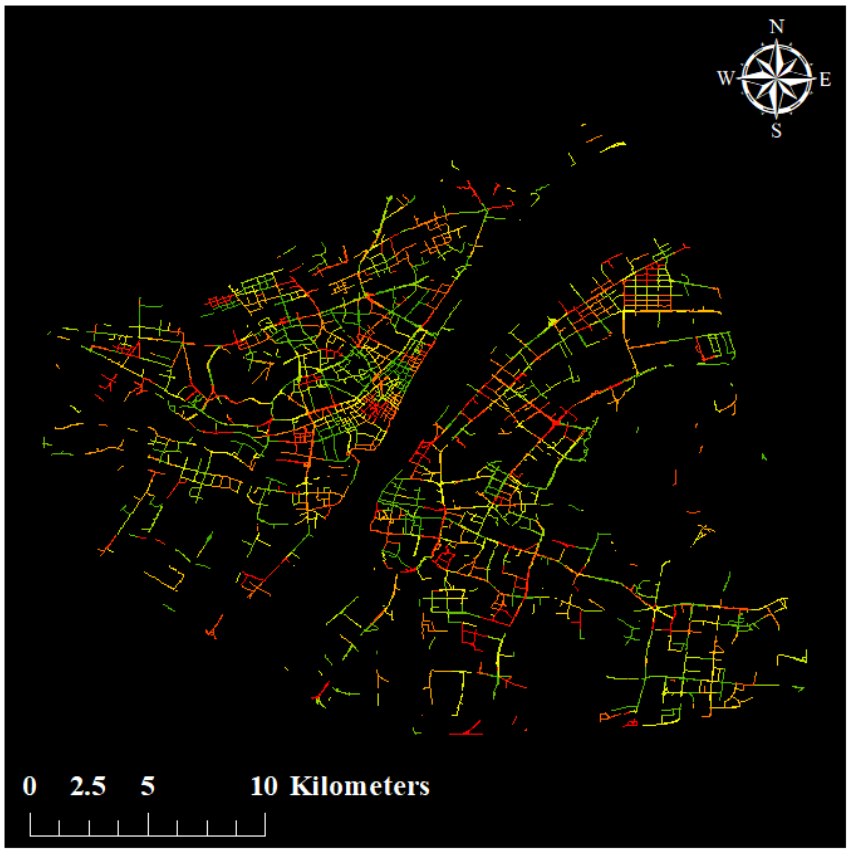


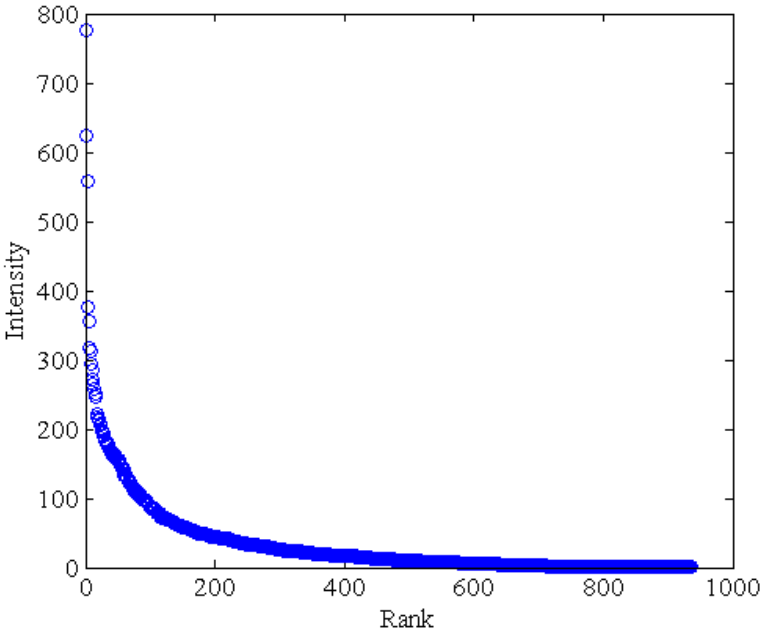


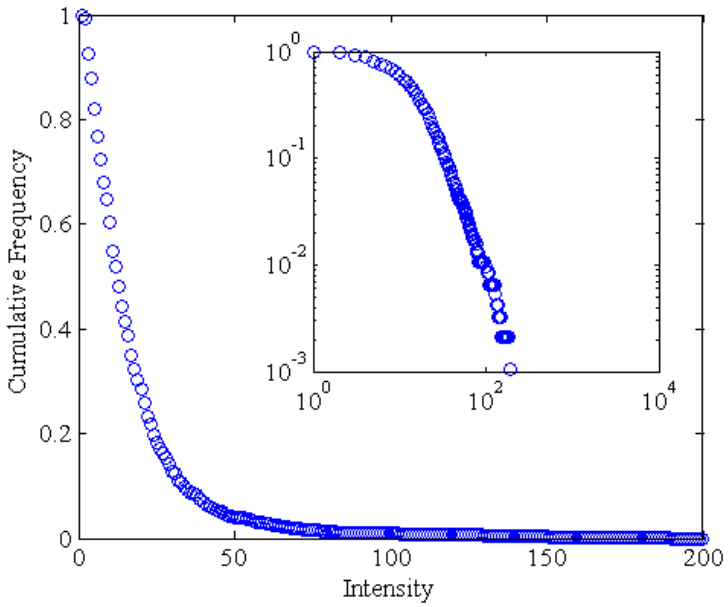


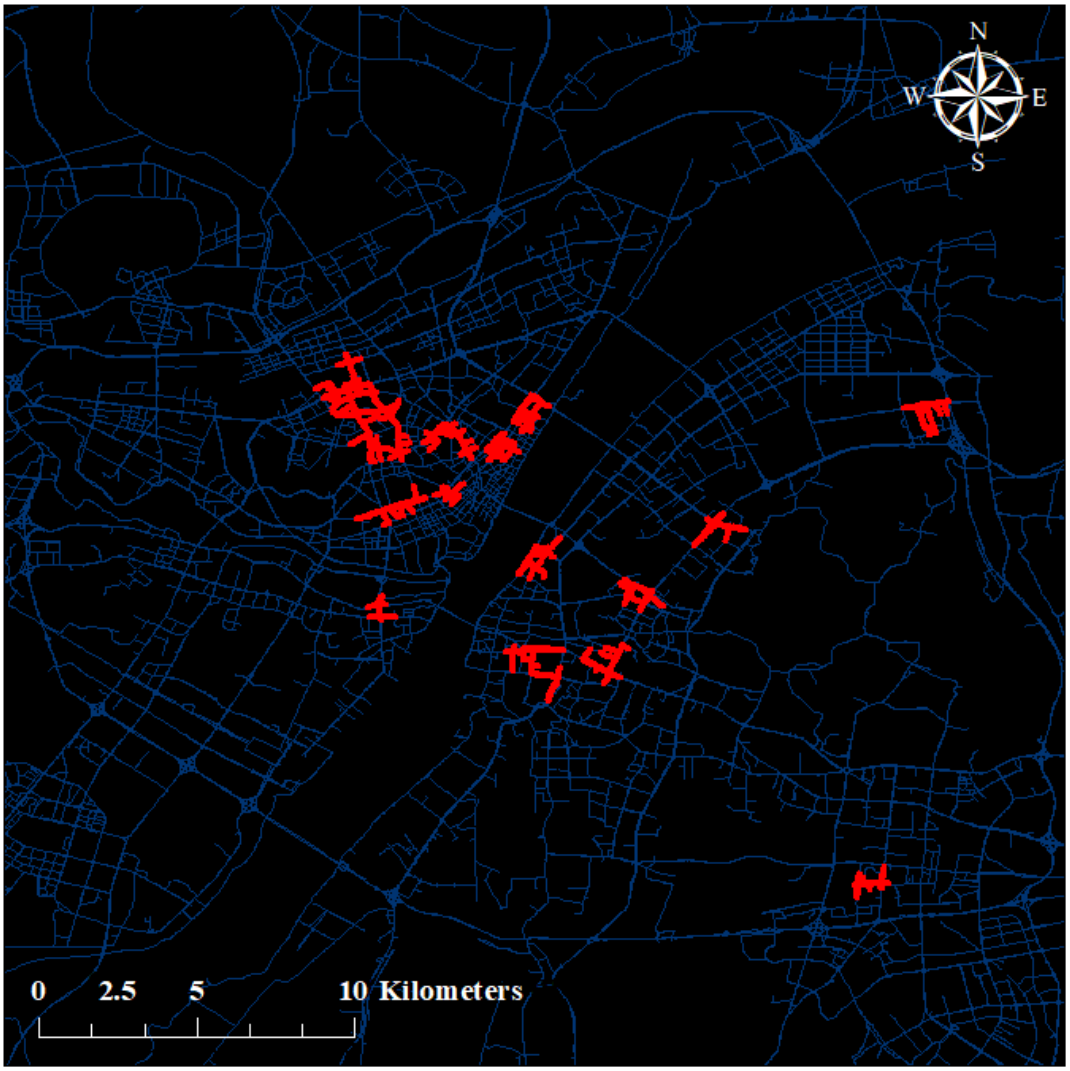


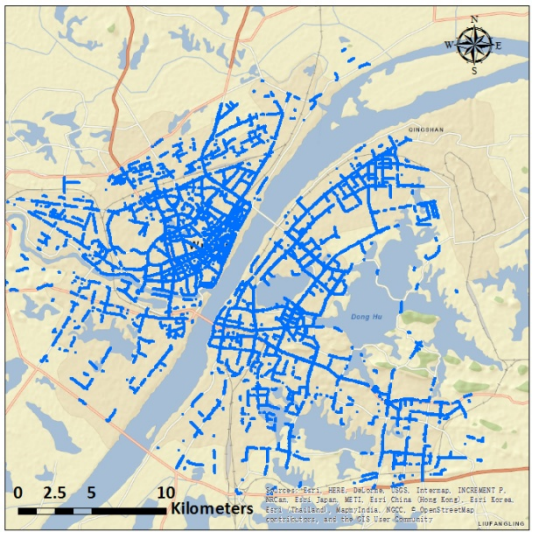


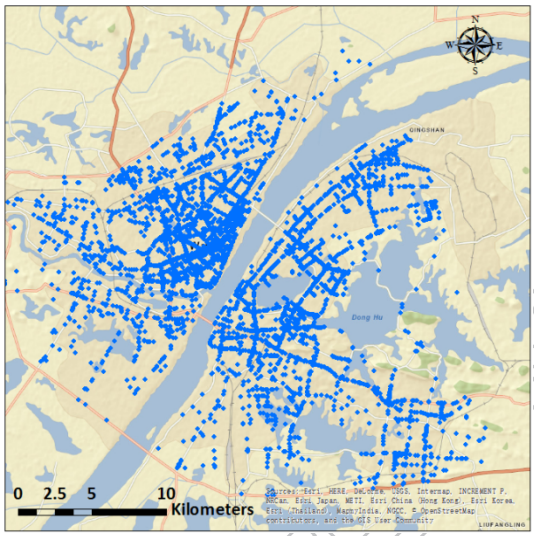


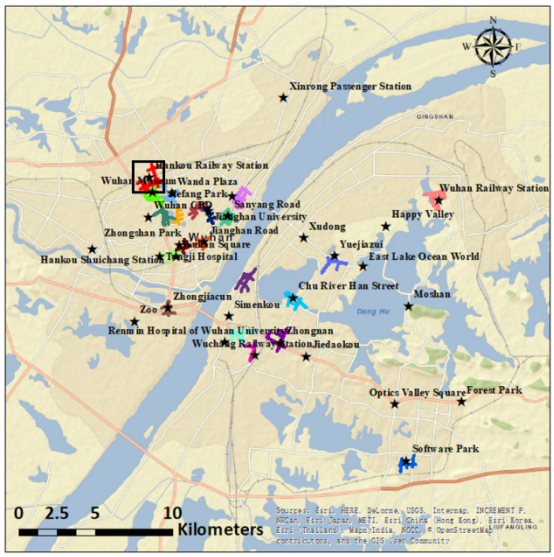


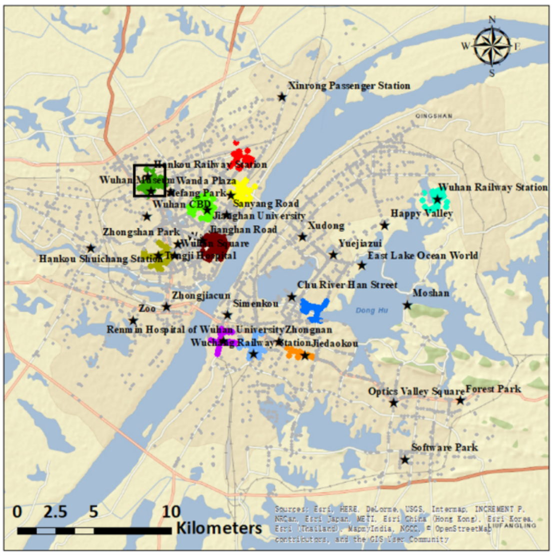


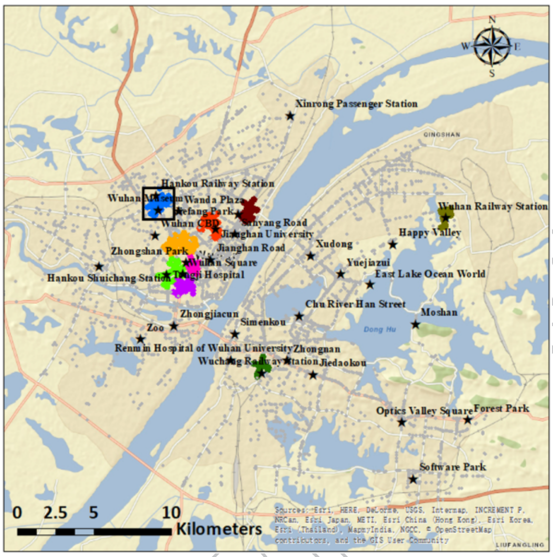


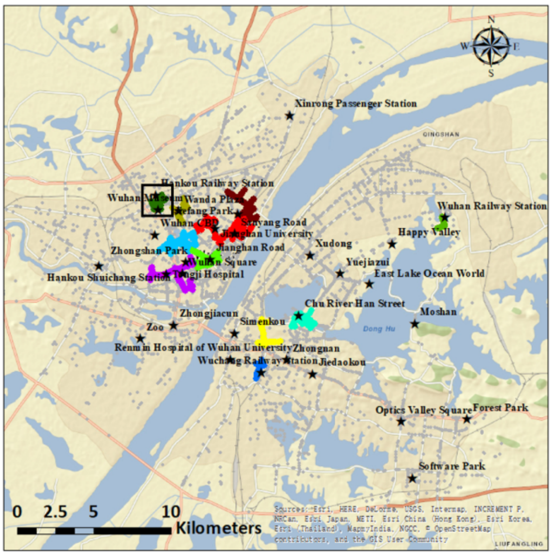


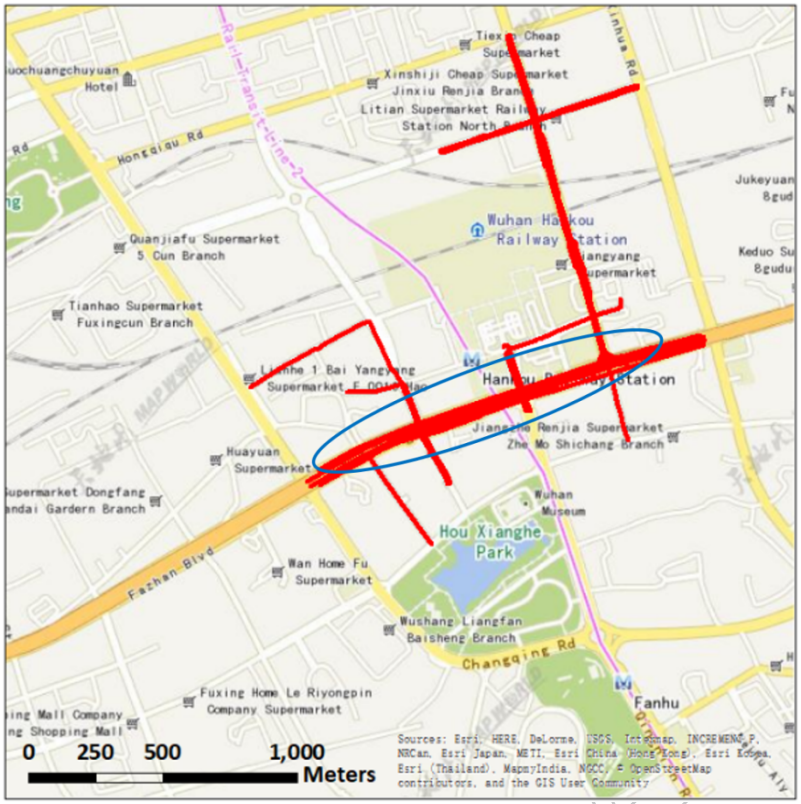


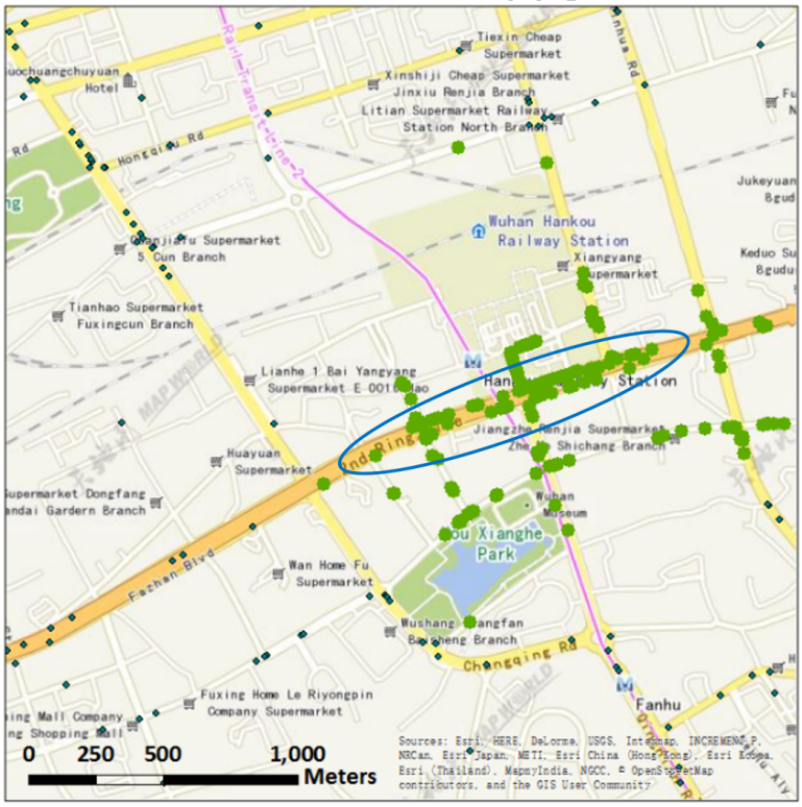


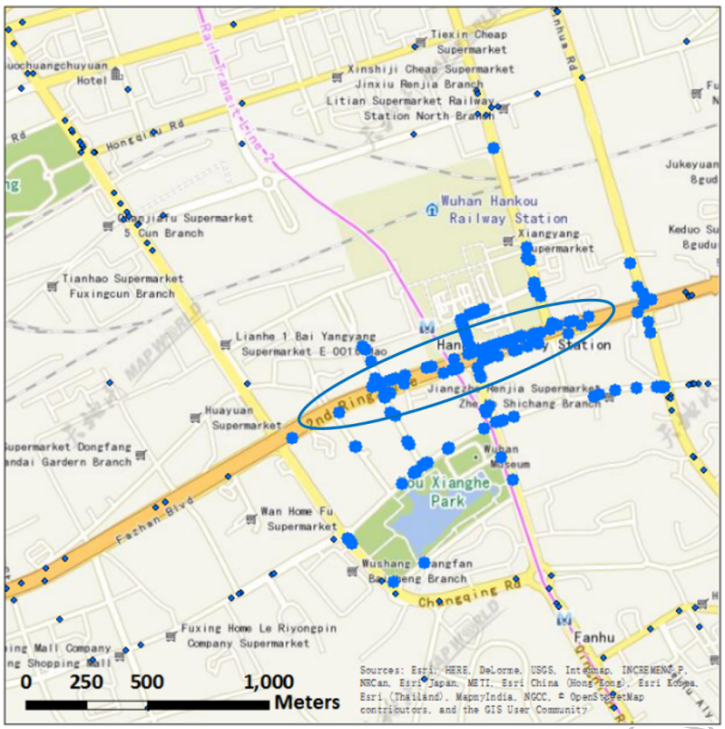


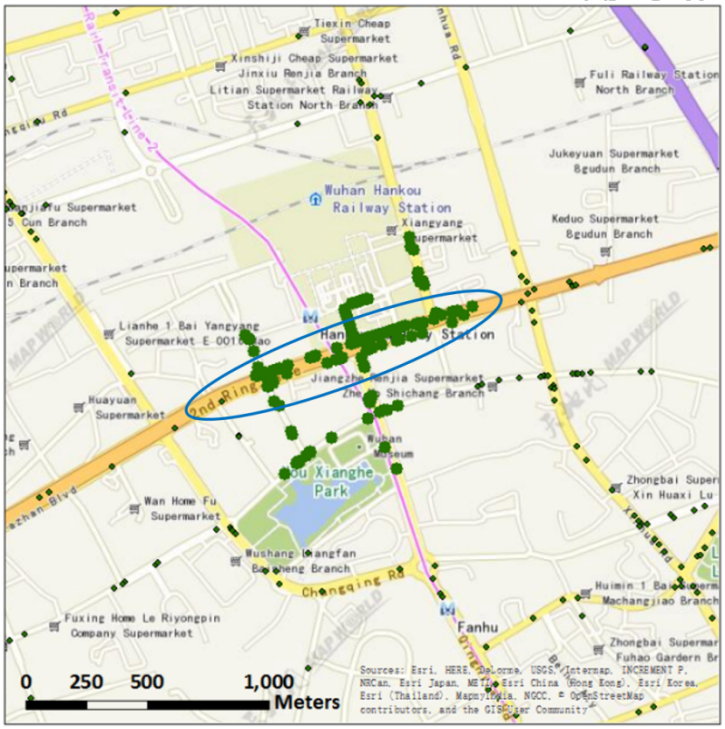


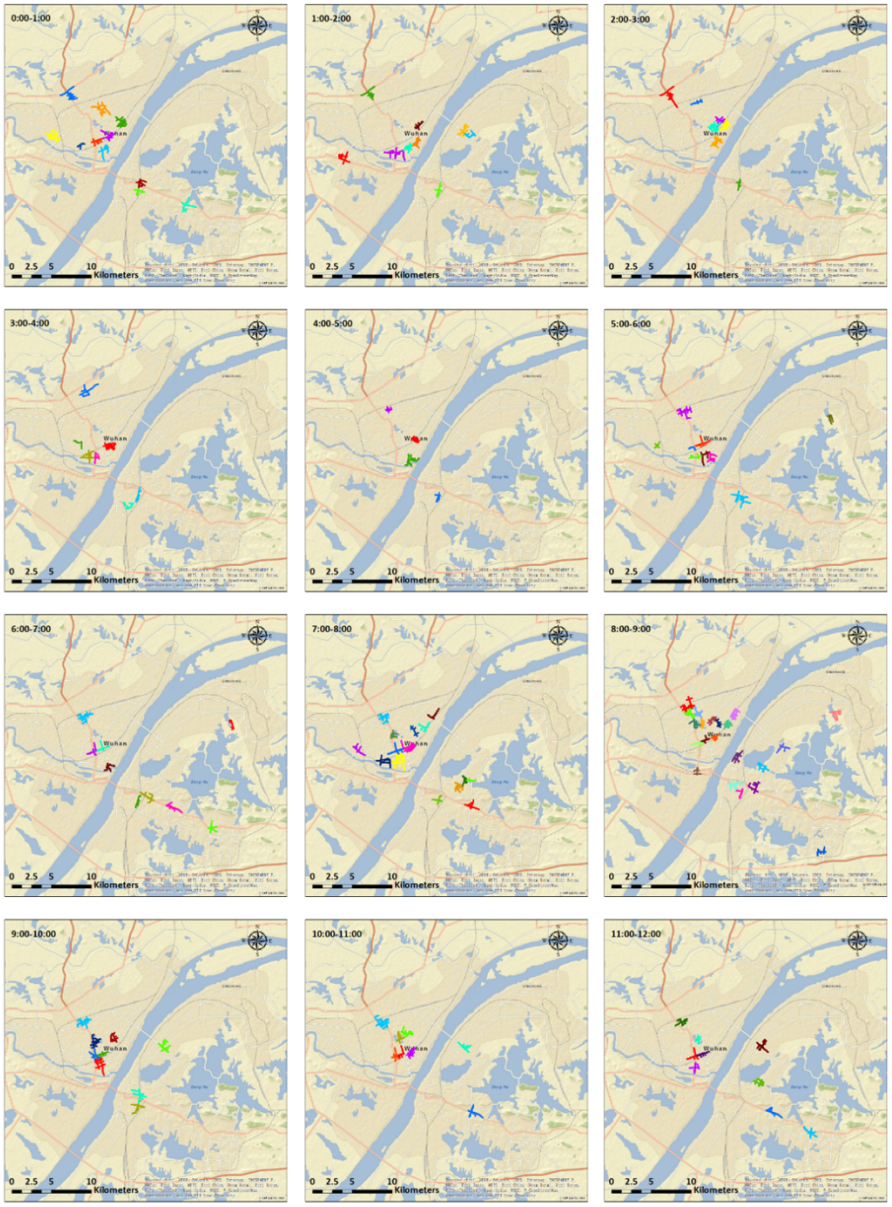












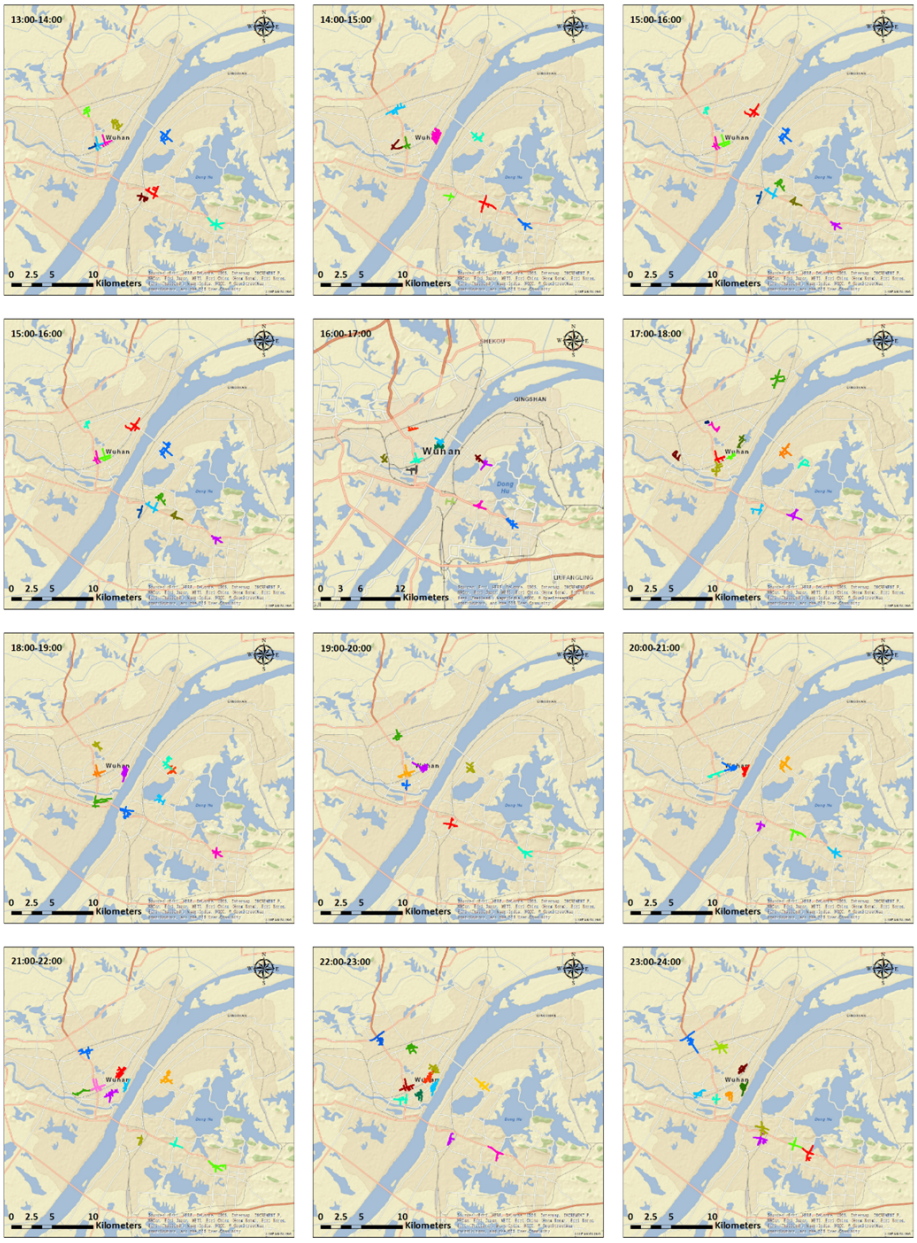


Table 1. The results of three partitions.

Number of clusters	Number in head	Percentage of head	Number in tail	Mean
937	248	26.5%	689	34.4
248	84	33.9%	164	103.9
84	27	32.1%	57	193.5

Table 2. Quantitative comparison of the clustering results.

		Size of Clusters	Number of events	Density of events
Proposed method	Max	15.11	1132	131.28
	Avg	7.26	384	52.49
	Min	3.61	195	24.27
TCDGDF	Max	15.38	647	50.50
	Avg	10.89	339	29.84
	Min	5.85	90	14.33
DBSCAN	Max	16.61	747	50.60
	Avg	11.58	405	35.78
	Min	4.70	180	22.28
Single-link	Max	25.79	1113	60.62
	Avg	12.42	416	33.98
	Min	2.23	106	13.67

Table 3. Quantitative comparison of the clustering results based on single hotspot.

	Size of Cluster	Number of events	Density of events
Proposed method	15.11	1082	71.6
TCDGDF	13.6	593	43.6
DBSCAN	10.9	547	50.0
Single-link	8.0	483	60.6