International Journal of Approximate Reasoning 103 (2018) 227–247

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

GPS trajectory data segmentation based on probabilistic logic ☆

Sini Guo [a], Xiang Li [b,c,*], Wai-Ki Ching [a,b,d], Ralescu Dan [e], Wai-Keung Li [f], Zhiwen Zhang [a]

[a] *Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong*
[b] *School of Economics and Management, Beijing University of Chemical Technology, Beijing, China*
[c] *Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China*
[d] *Hughes Hall, Wollaston Road, Cambridge, UK*
[e] *Department of Mathematical Sciences, University of Cincinnati, Cincinnati, USA*
[f] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

ARTICLE INFO

Article history:
Received 4 May 2018
Received in revised form 20 September 2018
Accepted 25 September 2018
Available online 9 October 2018

Keywords:
GPS trajectory data segmentation
Probabilistic logic
Newton's method

ABSTRACT

With the rapid development of internet economy, transparent logistics is stepping into a prosperity period with massive transportation data generated and collected every day. In this paper, we focus on the segmentation of GPS trajectory data generated in logistics transportation to analyze the vehicle behaviors and extract business affair information according to the vehicle behavior characteristics, which is challenging due to the complexity of trajectory data and unavailability of road information. We extract the stopping points from the trajectory data sequence based on the duration of nonmovement, and construct business time window and electronic fence by analyzing the driving habits of vehicles. Furthermore, we propose a probabilistic logic based data segmentation method (PLDSM) which not only helps finding all the business points but also assists in inferring the business affair categories. An efficient numerical algorithm integrating duality theory and Newton's method is proposed to obtain the optimal solution. Finally, a practical example is presented to validate the effectiveness of PLDSM. The results greatly enrich the data segmentation technique and promote the practicability of probabilistic logic.

1. Introduction

Massive data are generated every second with the rapid development of information technology. Although the processing capability of computer has achieved great improvement, data storage and retrieval are still challenged by the increasingly accumulated complicated data. Therefore, how to filter the redundant data and extract valuable information from raw data are a major concern in real applications. Data segmentation attracts people's attention for its efficiency in data summarization and information mining. As a data-based information technique, data segmentation is intended for segmenting data sequence into a series of disjoint segments based on some predetermined criteria. It has been applied successfully in various fields, such as image processing [42] [17], DNA sequence segmentation [5] and vehicle trajectory analysis [2]. Data

☆ This paper is part of the Virtual special issue on New Trend in Aggregation and Fusion of Information, edited by Irene Díaz and Yusuke Nojima.

* Corresponding author.
E-mail address: lixiang@mail.buct.edu.cn (X. Li).

https://doi.org/10.1016/j.ijar.2018.09.008
0888-613X/

segmentation is able to partition a digital image into regions of pixels by similarity in color and intensity. In vehicle trajectory analysis, data segmentation can be applied to simplify the vehicle movement to compact the continually evolving spatio-temporal database, achieving a fast and efficient support of applications such as recognition of trajectory patterns, prediction of congested areas [2]. In this paper, we mainly focus on GPS trajectory data segmentation.

Various trajectory data segmentation methods have been proposed in the past few years, which can be divided into two classes. The first is attribute-driven segmentation method. It is capable of partitioning the spatial trajectory into several number of segments such that the movement inside each segment is homogeneous with respect to some movement attributes, such as speed, residence time and heading. Yoon and Shahabi [41] partitioned a trajectory into a small number of spatially and temporally homogeneous segments. In each segment, the vehicle approximately moved at a constant speed. Buchin et al. [6] considered location, heading, speed, curvature, sinuosity and curviness to segment the trajectory into a minimum number of segments. Krumm and Horvitz [25] segmented the trajectory of a car by finding the data sequence with speed less than two miles per hour. Panagiotakis et al. [34] described the representativeness of each segment based on local density and trajectory similarity, and then identified the segment borders and the number of segments by the novel segmentation algorithm. Aronov et al. [3] designed the outlier-tolerant criterion and the standard deviation criterion to segment the trajectory by proposing a univariate attribute function. More attribute-driven data segmentation can be found in [7,14,15,40,43].

The second kind of trajectory data segmentation refers to the pattern-driven segmentation method, which mostly focuses on pattern detection based on individual or group of moving entities. Patterson et al. [36] used trajectory data to infer an individual's transportation mode into bus, walk and car to predict the most likely route. In view of that people had to walk through the transition between two different transportation modes, Zheng et al. [44] proposed a walk-based trajectory data segmentation method to identify the transportation modes. Laube and Imfeld [26] developed a relative motion framework to study the similar behaviors of different entities, where a collection of spatio-temporal patterns were defined based on the moving direction. As an extension of the relative motion framework, Laube et al. [27] extracted four movement patterns including flock, leadership, convergence and encounter from geospatial data. Benker et al. [4] redefined the flock as a set of objects which moved along paths close to each other for a predefined time. Jeung et al. [22] proposed a convoy query and outlined preliminary technique to detect the convoy of vehicles from massive spatio-temporal data. Zheng et al. [45] developed a change point-based segmentation method to partition the trajectory into segments with different transportation modes. Other pattern-based segmentation studies can be found in [1,16,23,30,35,46].

Most of the current GPS trajectory data segmentation are performed by means of cluster algorithm. Cluster algorithm is a well-known generic unsupervised learning technique, which can be applied to determine relevant similar groups of trajectory data according to some common characteristics. For example, Giannotti et al. [18] employed a cluster algorithm to extract the regions that vehicles usually visited based on the density of locations, and calculated the travel time of each segment between two regions to infer the vehicle behaviors. Pelekis et al. [37] proposed a cluster algorithm to segment trajectory data using various distance functions based on location, speed, acceleration and direction. There are two popular cluster algorithms commonly applied in the trajectory data segmentation: K-means (KM) and density-based spatial clustering of applications with noise (DBSCAN). The KM-based segmentation pursues to optimize the defined clustering quality function with a certain number of clusters [28,29]. The prerequisite of this method is knowing the number of clusters in advance. In contrast, DBSCAN clusters the trajectories based on density criteria without knowing the number of clusters [33]. It merely depends on two inputs: ε and $Minpts$, where ε refers to the radius of a neighborhood and $Minpts$ denotes the minimum number of data points in each neighborhood. For example, Gong et al. [20] proposed an improved DBSCAN algorithm to identify the stopping points of a GPS trajectory. Chen et al. [10] designed a T-DBSCAN algorithm by considering the spatio-temporal characteristics of the GPS trajectory. In addition to the above cluster algorithms, some other machine learning techniques are applied in data segmentation, such as K-nearest neighbor (KNN) [24], neural networks (NN) [9], convolutional neural network (CNN) [13] and deep neural network (DNN) [39].

Although the above machine learning algorithms are able to solve the complicated trajectory data segmentation problem effectively, there are still some restrictions. The first is that most of the current machine learning based data segmentation methods rely on the road information such as map matching or GIS components. In real applications, it may be expensive or even impossible to derive the GIS data. The second is the harsh demand on the accuracy of training data. The machine learning algorithms perform well in case that the training data is accurate. However, the data errors and information loss are inevitable in the data retrieval and collection process, which may cause the machine learning methods to obtain a wrong conclusion violating the common sense and logic. To tackle with the above deficiencies, we incorporate probabilistic logic into data segmentation to reduce the negative influence of erroneous data, and propose the PLDSM, which not only segments the data sequence accurately, but also helps on recuperating information loss. As an application, we apply it to logistics transportation field.

In recent years, the rapid development of internet economy in China greatly promotes the prosperity of logistics, which permeates every aspect of our daily life. For example, on 11th November 2017, the total trading volume in Alibaba turned to 168.2 billion RMB, generating 812 million logistics orders. As an important part of logistics, the fourth party logistics is playing an increasingly important role for its efficiency in collaborating information resources and information sharing. Instead of providing transportation service for real products, the fourth party logistics is intended for providing logistics programming, consultation and logistics information system for the first, second and third party logistics. To achieve a transparent transportation and have a better control of the logistics information, the transportation vehicles in the first, second and

Fig. 1. Before segmentation.

Fig. 2. After segmentation.

third party logistics are usually required to install electronic sensors including GPS receivers, electronic locks, and electronic fueling tank caps, which collect massive transportation data for the fourth party logistics. However, the fourth party logistics has no access to obtaining the business affair information including business time and business locations. As for this application scene, in this paper, we investigate the application of data segmentation to mine the business affair information from GPS trajectory data. The key issue is to segment the series of GPS data into segments of which each segment corresponds to a particular business affair. For example, based on trajectory data generated by an individual vehicle, Fig. 1 only provides a travel trajectory. After data segmentation, the trajectory is segmented into five segments, of which each contains the origin and destination of a business affair. Four business points A, B, C and D where business affairs take place are detected (see Fig. 2). In the meanwhile, we need to extract more business affair information by analyzing the segmented trajectory data characteristics. For example, according to the location and time of each business point and the daily driving habits of each vehicle, we can deduce the category of products this vehicle is transporting. This paper focuses on identifying the time and destinations of business affairs, and business affair categories without travel reports. Compared with previous data segmentation methods, our proposed PLDSM is able to extract the hidden logistics transportation information when GIS data is unavailable and performs better in recuperating information loss. The contribution of this paper can be stated as follows:

- Probabilistic logic is firstly applied in segmenting GPS trajectory data without referring to any road information, which is able to overcome the deficiency of high dependence on GIS data in previous studies.
- The PLDSM is proposed and an efficient numerical algorithm integrating dual theory and Newton's method is designed, achieving a better performance than traditional machine learning technique when the data is inadequate and not accurate enough.

The remainder of the paper is structured as follows. In Section 2, we introduce the transportation data for segmentation. Section 3 presents the PLDSM and formulates a maximum entropy model. In Section 4, Newton's method together with duality theory are employed to solve the optimal solution of the maximum entropy model. A practical example is then presented in Section 5 to validate the effectiveness of PLDSM. Finally, Section 6 concludes the paper.

2. Data preparation

In this section, we introduce different types of data collected in the transportation process. Various electronic devices are equipped in the vehicles to track the transportation status timely. These include logistics electronic locks, electronic fueling tank caps and GPS receivers.

	Time	Longitude	Latitude	Speed	Heading
L_1:	t_1	x_1	y_1	v_1	h_1
L_2:	t_2	x_2	y_2	v_2	h_2
			……		
L_n:	t_n	x_n	y_n	v_n	h_n

	Time	Longitude	Latitude	Residence time
B_1	t_1^*	x_1^*	y_1^*	δ_1
B_2	t_2^*	x_2^*	y_2^*	δ_2
		……		
B_N	t_N^*	x_N^*	y_N^*	δ_N

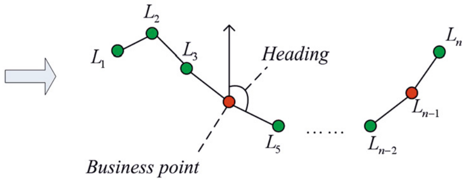

Fig. 3. GPS log and trajectory.

Electronic locks: The electronic locks, by means of mobile GPRS wireless network system, are installed to monitor the whole transportation process of products in the real time, assisting transportation enterprises and the clients to get real-time dynamic information of goods to ensure the effective supervision during transportation. This can reduce the cost of enterprise and enhance competitiveness. Every time the cargo door of the vehicle is opened, there is a record in the electronic lock.

Electronic fueling tank caps: The electronic fueling tank caps are designed to monitor the fueling of vehicles in the transportation process to prevent drivers from stealing the oil. Again every time when the vehicle is fueled, there is an electronic record in the tank cap.

GPS receivers: GPS (Global Positioning System) receivers are widely applied in transportation fields, which are intended for recording and uploading real-time data including position coordinates, speed, heading and some other geographical information.

The GPS receivers renew the positioning data every several seconds, generating a sequence of GPS points, $L = \{L_1, L_2, \dots, L_n\}$. Each GPS point L_i includes multiple vehicle attributes (see Fig. 3), $L_i = \{t_i, x_i, y_i, v_i, h_i\}$, $i = 1, 2, \dots, n$, where t_i refers to the timestamp the GPS data is uploaded, x_i and y_i stand for the longitude and latitude of the vehicle at t_i, respectively, v_i denotes the instant speed, and h_i represents the heading ranging from 0° to 360°. In particular, $h_i = 0°$ means the heading direction is North and $h_i = 90°$ refers to the East.

In addition to the transportation data L, some limited historical business data $B = \{B_1, B_2, \dots, B_N\}$ is also available for the purpose of analyzing the regular patterns of business affair distributions (see Fig. 3), where $B_j = \{t_j^*, x_j^*, y_j^*, \delta_j\}$ denotes the j-th business data, t_j^* represents the time when the vehicle arrives at the j-th business point, x_j^* and y_j^* refer to the longitude and latitude, and δ_j is the residence time of vehicle at the j-th business point, $j = 1, 2, \dots, N$.

3. Data segmentation

In this section, we present the procedure of segmenting the GPS trajectory data to find the business points. As shown in Fig. 4, firstly, the outliers are detected and removed in the data processing. Secondly, the stopping points of the vehicle are identified by analyzing the movement status of the vehicle at each GPS point. Finally, probabilistic logic is introduced to filter the business points from the stopping points by comprehensively considering the electronic lock status, electronic fueling tank cap status, residence time, residence location and trajectory.

3.1. Data processing

GPS receiver records and uploads data frequently, and this will inevitably generate some outliers. Outliers refer to the data that are significantly distant from others in terms of a certain similarity metric. There are two kinds of outliers commonly emerged in the GPS data stream. The first is speed outlier. If the speed at a GPS point is zero, while the vehicle keeps moving for a consecutive time period before and after this point, then the speed at this instant is an outlier. Similarly, if the speed is nonzero while the vehicle stays still for a consecutive time period before and after this point, then it is another kind of speed outlier (see Fig. 5). To deal with these outliers, a popular method is to replace the outlier speed with the average of the speed before and after this GPS point. For example, if $v_i = 0$, $v_{i-1} > 0$ and $v_{i+1} > 0$, then $v_i = 0$ is replaced by $v_i = (v_{i-1} + v_{i+1})/2$. The second is location outlier. It is generally assumed that the moving distance of a vehicle in a short time interval is less than a certain threshold due to the limited speed. If the distance of two consecutive GPS point is far beyond the distance the vehicle travels at the maximum speed, then it is deemed as a location outlier, which exhibits an abrupt jump in the trajectory (see Fig. 6). Usually we delete this outlier to achieve a smooth trajectory.

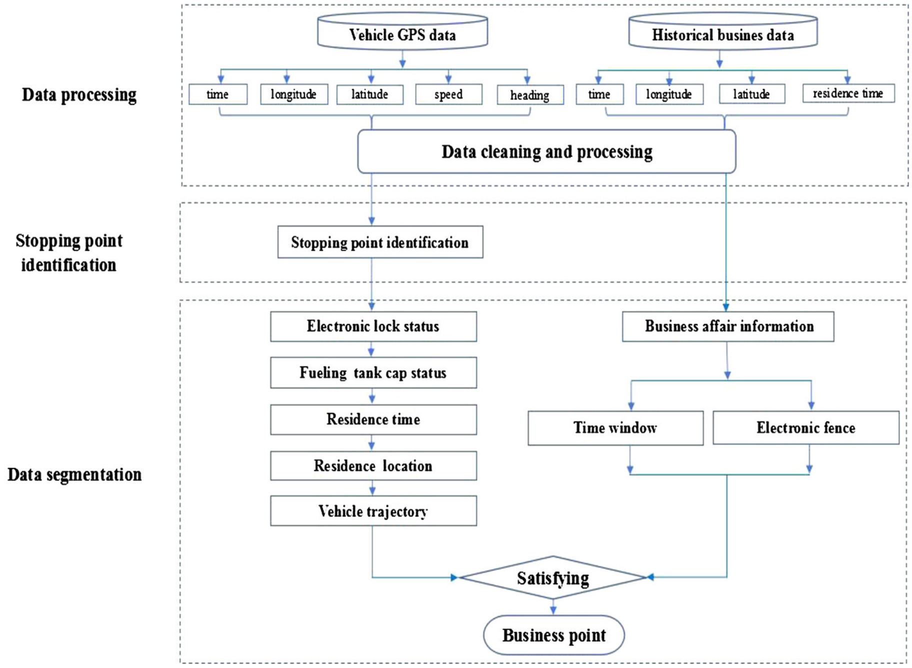

Fig. 4. Framework for GPS trajectory data segmentation.

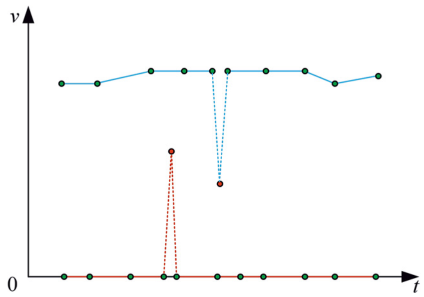

Fig. 5. Speed outlier. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

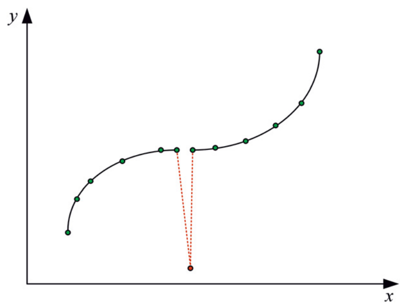

Fig. 6. Location outlier.

3.2. Stopping points detection

The stopping point detection is designed for identifying potential trip ends within the data stream by searching for time periods of nonmovement. On a two-dimensional plane, we first sequentially connect the discrete points to form a trajectory, and then divide the trajectory into trips if the time interval of nonmovement exceeds a certain threshold. A GPS point is deemed as a trip end point if the vehicle keeps moving for a consecutive time period before this point and then stays nonmovement for a consecutive time period after this point. For example, Table 1 provides a GPS trajectory data sequence. It is seen that the vehicle stops at 2014/5/18 2:26:19, stays nonmovement for 438 s and then starts a new trip.

Table 1
GPS trajectory data.

t	x	y	h	v
2014/5/18 2:25:14	121.368100	31.250116	121	31
2014/5/18 2:25:26	121.368100	31.249650	127	20
2014/5/18 2:25:43	121.368966	31.249166	120	25
2014/5/18 2:25:57	121.369833	31.248700	120	22
2014/5/18 2:26:19	121.370750	31.248383	0	0
2014/5/18 2:29:12	121.371233	31.248383	0	0
2014/5/18 2:32:12	121.371266	31.248366	0	0
2014/5/18 2:32:19	121.371283	31.248366	0	0
2014/5/18 2:33:37	121.371283	31.248366	0	0
2014/5/18 2:34:04	121.371600	31.248050	171	12
2014/5/18 2:34:21	121.371400	31.247233	193	25
2014/5/18 2:34:22	121.371383	31.247166	193	25
2014/5/18 2:34:35	121.371116	31.246283	196	27
2014/5/18 2:34:47	121.370883	31.245383	193	29

This step produces a trip-end file that contains all the stopping points coordinate information, including time, longitude and latitude. In addition, the residence time of each stopping point can also be obtained by computing the time of nonmovement.

3.3. Business point identification based on probabilistic logic

All the stopping points are detected after the above procedures. To further filter the business points from these stopping points, we sequentially analyze the electronic lock status, fueling tank cap status, residence time, residence location and trajectory of each stopping point.

Electronic lock and fueling tank cap

For each stopping point, we first check the status of the electronic lock and fueling tank cap. If there is an unlocking record in the electronic lock, then it is definitely a business point as the lock can only be unlocked when cargo discharging takes place. In case that no unlocking record is available, we turn to identify the fueling tank cap status. If the fueling tank cap is unlocked, it can be inferred that the vehicle lies in a gas station where the vehicle is refueling.

Business time window

In case that no record in both of the electronic lock and fueling tank cap is available, we construct business time window based on historical business data to analyze the relationship between vehicle residence time and business affair. Suppose that the residence time is subject to a certain kind of probability distribution. We study the statistics of the residence time of historical business point and formulate the frequency distribution function. Denote the residence time of vehicle in the i-th business point as δ_i. The steps are as follows: Firstly, reorder δ_i in an increasing order as $\delta_{(1)}, \delta_{(2)}, \ldots, \delta_{(N)}$, where $\delta_{(1)} = \min\{\delta_1, \delta_2, \ldots, \delta_N\}$ and $\delta_{(N)} = \max\{\delta_1, \delta_2, \ldots, \delta_N\}$. Secondly, divide $[\delta_{(1)}, \delta_{(N)}]$ into m identical intervals: $[\delta_{(1)}, \delta_{(1)}+\sigma)$, $[\delta_{(1)}+\sigma, \delta_{(1)}+2\sigma)$, ..., $[\delta_{(1)}+(m-1)\sigma, \delta_{(N)}]$, where $\sigma = (\delta_{(N)} - \delta_{(1)})/m$. Finally, we construct the statistics of the frequency of stopping points whose residence time lies in $[\delta_{(1)}+i\sigma, \delta_{(1)}+(i+1)\sigma)$, $i = 0, 1, \ldots, m-1$, and denote it as f_i. Then study the statistics of the frequency of business points whose residence time is in $[\delta_{(1)}+i\sigma, \delta_{(1)}+(i+1)\sigma)$ and denote it as f_i^*. Figs. 7 and 8 give the histogram of frequency and distribution function, where the red histogram represents the business points frequency and the blue one denotes all the stopping points frequency. The distribution function is

$$F(t) = \begin{cases} \dfrac{f_i^*}{f_i}, & \text{if } \delta_{(1)}+i\sigma \le t < \delta_{(1)}+(i+1)\sigma,\ i=0,1,\ldots,m-1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For each point, its probability of residence time satisfying the business demand (the criterion we adopt for classifying a business point) is $F(t)$ if the residence time is t.

Electronic fence

Electronic fence is constructed to analyze the relationship between the residence location and business affair. By plotting the historical business data on the two-dimensional plane, it is found that the business points are centered around several fixed positions, of which the density is much higher compared with other positions on the plane. Therefore, we employ DBSCAN to cluster these historical business points. For each cluster, an electronic fence is built to locate the area where business affair usually takes place (see Fig. 9). We can identify whether a stopping point is a business point by electronic fence. If the stopping point lies in the electronic fence, then it is probably a business point. However, due to the existence of data perturbation, the longitude and latitude of a stopping point may vary slightly during the nonmovement period. For example, it follows from Table 1 that the vehicle keeps nonmovement from 2:26:19 to 2:33:37, but its locations are slightly different. To overcome this deficiency, we divide the varying locations of each stopping point into two groups: (1) locations in the

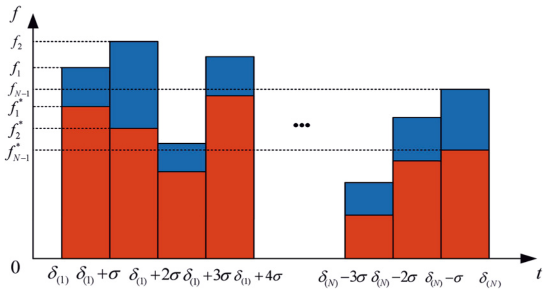

Fig. 7. Histogram of frequency.

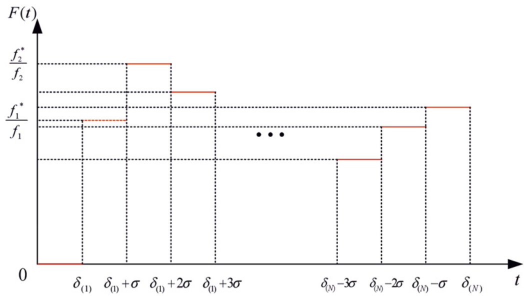

Fig. 8. Frequency distribution function.

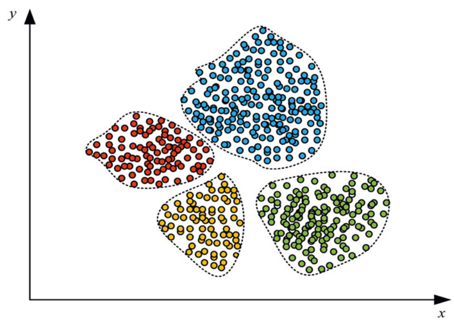

Fig. 9. Electronic fences.

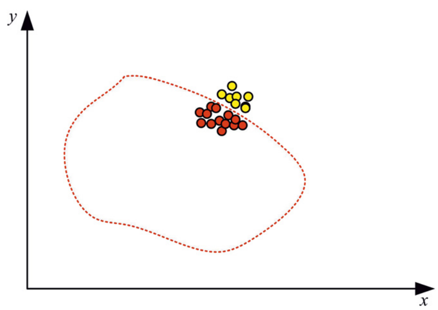

Fig. 10. Stopping point data perturbation.

electronic fence and (2) locations outside the electronic fence (see Fig. 10). Assume that the frequencies of those two groups are f_{in} and f_{out}, respectively. Then the probability for the location satisfying the business demand can be estimated by

$$F = \frac{f_{in}}{(f_{in} + f_{out})}. \tag{2}$$

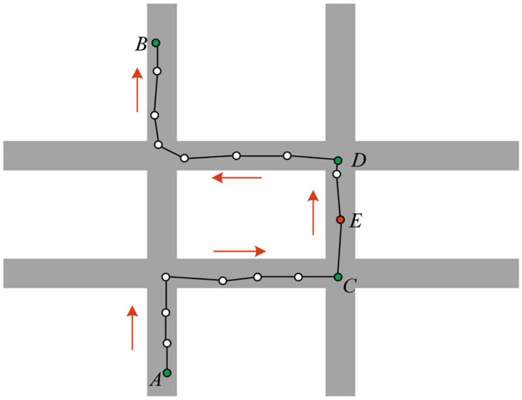

Fig. 11. Special trajectory 1.

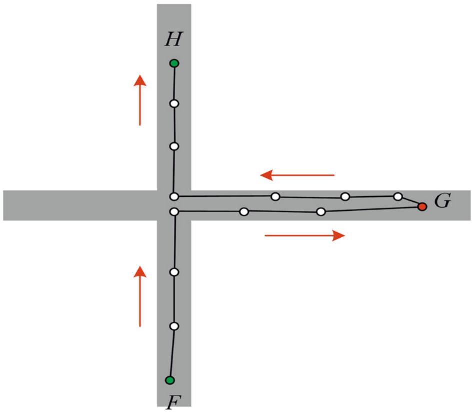

Fig. 12. Special trajectory 2.

Trajectory identification

Some special trajectories can be utilized to determine whether a stopping point is a business point. There are two special trajectories considered. The first is illustrated in Fig. 11 which provides a trajectory from A to B. The vehicle takes the path $A \to C \to D \to B$ rather than $A \to B$ though there is a direct path. Therefore, the stopping point E in path $C \to D$ is likely a business point. Another typical trajectory is shown in Fig. 12, where a vehicle is driving straightly along the road from F to H. At the crossroad, it turns to the right, stops at G for a while, and returns by the same way it came. Then stopping point G is likely to be a business point. Therefore, we can analyze the heading changes of the vehicle before and after the stopping point to identify whether its trajectory satisfies the business demand. If the trajectory matches one of the two special trajectories, then it is probably a business point.

To find the business points accurately and estimate the corresponding probabilities, we introduce probabilistic logic into data segmentation process. Probabilistic logic was first proposed by Nilsson who presented a procedure for computing probabilistic entailment [32], which inspires many subsequent probabilistic logic studies [8,12,19,21,31,38]. Given a set of propositions and their associated probabilities, this procedure computes a range of probabilities within which the probability of some given target proposition lies. The procedure operates by first finding all consistent assignments of truth values to the given propositions and then using these assignments to set up a system of linear equations that can be solved.

In view of the above five attributes: electronic lock status, fueling tank cap status, residence time, location and trajectory, six propositions are set and seven logic formulae are proposed to determine whether the stopping points are business points.

Propositions

- P_1: There is an unlocking record in the electronic lock
- P_2: There is an unlocking record in the electronic fueling tank cap
- P_3: The residence time satisfies the business demand
- P_4: The location satisfies the business demand
- P_5: The trajectory satisfies the business demand
- P_6: The stopping point is a business point

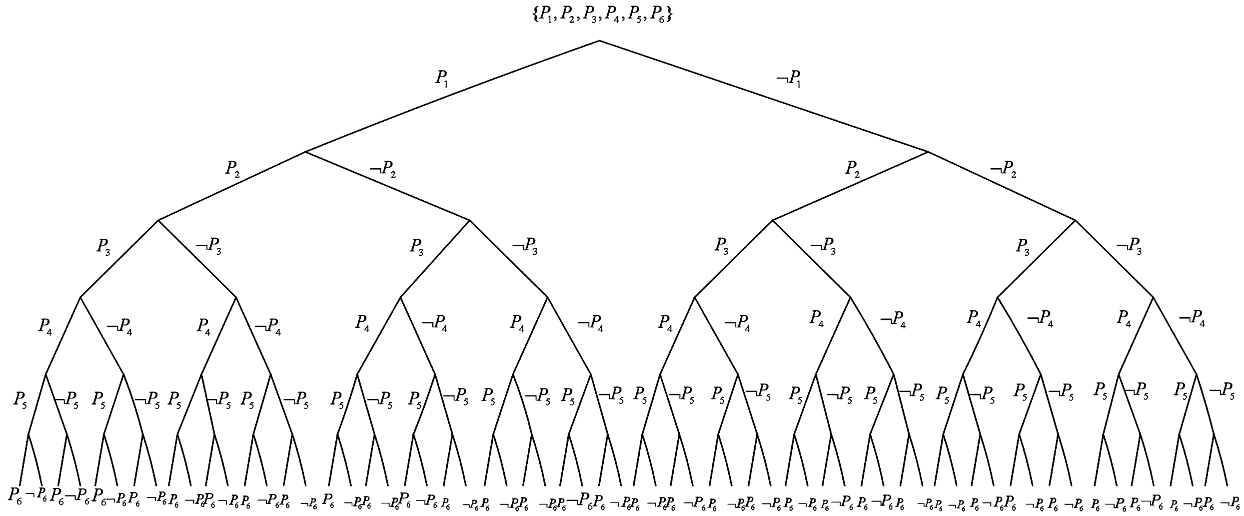

Fig. 13. Semantic tree.

Logic formulae

- $P_1 \to P_6$: The stopping point is certainly a business point if there is an unlocking record in the electronic lock
- $P_2 \to \neg P_6$: The stopping point is not a business point if there is a record in the electronic fueling tank cap
- $\neg P_1 \wedge \neg P_2 \wedge \neg P_3 \to \neg P_6$: The stopping point is not a business point if the corresponding residence time does not satisfy the business demand
- $\neg P_1 \wedge \neg P_2 \wedge \neg P_4 \to \neg P_6$: The stopping point is not a business point if its location does not satisfy the business demand
- $\neg P_1 \wedge \neg P_2 \wedge P_3 \wedge P_4 \to P_6$: The stopping point is a business point if its residence time and location both satisfy the business demand
- $\neg P_1 \wedge \neg P_2 \wedge P_3 \wedge P_5 \to P_6$: The stopping point is a business point if the residence time and trajectory both satisfy the business demand
- $\neg P_1 \wedge \neg P_2 \wedge P_4 \wedge P_5 \to P_6$: The stopping point is a business point if its location and trajectory both satisfy the business demand

For any proposition P_i, $i = 1, 2, \ldots, 6$, there are two possible situations called atoms: W_{i1} where P_i is true and W_{i2} where P_i is false. The reality world must be in one of them, but we might not know which one. Suppose that $Tr(.)$ is an assignment function from propositions set S to $\{0, 1\}$, where $S = \{P_1, P_2, P_3, P_4, P_5, P_6\}$. For any $P_i \in S$,

$$Tr(P_i) = \begin{cases} 1, & \text{if } P_i \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

Based on the above six propositions, a binary semantic tree can be constructed to describe all the 2^6 sets of atoms (see Fig. 13). At each node of the tree, we branch left if the proposition is true and branch right if the proposition is false. The truth values of the six propositions are shown in Table 2, of which each column represents a possible atom with a unique set of truth values for all the six propositions.

Although Table 2 provides 64 sets of atoms, there are actually fewer possible atoms due to some of the true values for our 6 propositions are logically inconsistent. Several different kinds of inconsistencies are detected in Table 2. The first is the inconsistency between $P_2 \to \neg P_6$ and $P_1 \to P_6$, which is marked with a "×". Take $[1, 1, 1, 1, 1, 0]^T$ as an example, it means that the electronic lock and the fueling tank cap are unlocked simultaneously, which is impossible in reality. The second is the inconsistency with $P_1 \to P_6$ which is marked with a "×". For example, the truth values of the six sentences $[1, 0, 1, 1, 1, 0]^T$ are inconsistent since the stopping point is definitely a business point when there is an unlocking record in the electronic lock. The third inconsistency is marked with a "×", which contradicts with $P_2 \to \neg P_6$. Some other inconsistencies due to the contradiction with the proposed logic formulae are marked with different colors. Finally, 24 sets of consistent atoms are marked with a "$\surd$", of which the truth value matrix is expressed as follows:

$$\mathbf{V} = \begin{pmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0 \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0 \\ 1&1&1&1&0&0&0&0&1&1&1&0&1&1&1&1&0&0&0&0&1&0&0&0 \end{pmatrix},$$

where V_{ij} represents the consistent truth value of P_i in j-th possible atom, $i = 1, 2, \ldots, 6$, $j = 1, 2, \ldots, 24$.

Table 2
Truth value assignment in different possible atoms.

P_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P_2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
P_3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
P_4	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
P_5	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
P_6	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
	×	×	×	×	×	×	×	×	√	×	√	×	√	×	√	×	×	√	×	√	×	√	×	√	√	×	√	×	√	×	×	√
P_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P_2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
P_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P_4	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
P_5	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
P_6	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
	×	×	×	×	×	×	×	×	√	×	√	×	√	×	√	×	×	√	×	√	×	√	×	√	√	×	×	√	×	√	×	√

The possible atoms comprise a sample space where a probability distribution is defined. Denote vectors $\mathbf{p} = [p_1, p_2, \ldots, p_6]^T$ and $\mathbf{q} = [q_1, q_2, \ldots, q_{24}]^T$, where p_i is the probability that proposition P_i is true and q_j is the probability that the reality world lies in W_j, $i = 1, 2, \ldots, 6$, $j = 1, 2, \ldots, 24$. Due to the exclusivity and exhaustivity of all the possible atoms, the sum of q_i equals to one. Then the probabilities of the propositions satisfy the following matrix equation:

$$\mathbf{p} = \mathbf{V}\mathbf{q}. \tag{3}$$

Following from Eq. (3), the probabilities of the six propositions are expressed as

$$p_i = \sum_{j=1}^{24} V_{ij} q_j, \quad i = 1, 2, \ldots, 6, \tag{4}$$

which implies that the probability of a proposition is the sum of the probabilities for the possible atoms in which that proposition is true.

For each stopping point, the probabilities of the first five propositions p_1, p_2, p_3, p_4 and p_5 can be estimated by analyzing its GPS trajectory data and historical business data. Here p_1, p_2 and p_5 are binary variables taking values in $\{0, 1\}$. Denote $p_1 = 1$ if there is an unlocking record in the electronic lock, otherwise, $p_1 = 0$. Similarly, $p_2 = 1$ if there is an unlocking record in the fueling tank cap, otherwise, $p_2 = 0$. $p_5 = 1$ if the trajectory satisfies the business demand, otherwise, $p_5 = 0$. Furthermore, p_3 and p_4 take values in interval $[0, 1]$, which can be calculated according to Eqs. (1) and (2). Our objective is to compute p_6 to determine whether this stopping point is a business point. Following from Eq. (4), it is known that p_6 can be uniquely determined once a vector of possible atoms $\mathbf{q} = [q_1, q_2, \ldots, q_{24}]^T$ is selected. However, the vector $\mathbf{q}$ is not unique. Therefore, a maximum entropy model is formulated to select the optimal vector $\hat{\mathbf{q}}$.

$$\begin{cases} \max & -\sum_{j=1}^{24} q_j \ln q_j \\ & \sum_{j=1}^{24} V_{ij} q_j = p_i, \ i = 1, 2, \ldots, 5, \\ & q_1 + q_2 + \ldots + q_{24} = 1, \\ & 0 \le q_j \le 1, \ j = 1, 2, \ldots, 24, \end{cases} \tag{5}$$

where the first constraint ensures that vector $\mathbf{q}$ satisfies $\mathbf{V}\mathbf{q} = \mathbf{p}$, the second constraint implies that the sum of probabilities for all the possible atoms equals to 1, and the third constraint gives the lower and upper bounds of q_j, $j = 1, 2, \ldots, 24$.

4. Model solving based on Newton's method

In this section, we first transform the maximum entropy problem into its dual problem, and then apply Newton's method to solving the problem. There are two major reasons for conducting the transformation. First, the original problem is a maximization problem with constraints while the dual problem is a minimization problem *without* constraints. Second, the number of variables in the dual problem is significantly less than the original problem, therefore the computational cost can be substantially saved. The above maximum entropy model is expressed as

$$\begin{cases} \max & -\sum_{j=1}^{24} q_j \ln q_j \\ & \bar{\mathbf{V}}\mathbf{q} = \bar{\mathbf{p}}, \\ & 0 \le q_j, \ j = 1, 2, \ldots, 24, \end{cases} \tag{6}$$

where $\bar{\mathbf{p}} = [p_1, p_2, p_3, p_4, p_5, 1]^T$ and $\bar{\mathbf{V}}$ composes of the first five rows in $\mathbf{V}$ and a vector of all ones. Remark that the constraint $q_j \le 1$ can be discarded as we require that $q_1 + q_2 + \cdots + q_{24} = 1$, and $0 \le q_j$, $j = 1, 2, \ldots, 24$.

$$\bar{\mathbf{V}} = \begin{pmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0 \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \end{pmatrix}.$$

As shown in [11], the dual problem of (6) is

$$\min_{\boldsymbol{\lambda}} \max_{\mathbf{q}} \left\{ -\sum_{j=1}^{24} q_j \ln q_j + \boldsymbol{\lambda}^T (\bar{\mathbf{p}} - \bar{\mathbf{V}}\mathbf{q}) \right\}, \tag{7}$$

Algorithm 4.1 Newton's Method.

Step 1. Initialize an initial solution $\boldsymbol{\lambda}_0$, and a small enough positive threshold ε. Set $k=0$.
Step 2. Compute the gradient $\nabla f(\boldsymbol{\lambda}_{(k)})$ and Hessian $\nabla^2 f(\boldsymbol{\lambda}_{(k)})$.
Step 3. Derive $\boldsymbol{\lambda}_{(k+1)}$ according to iteration formula:
$\boldsymbol{\lambda}_{(k+1)} = \boldsymbol{\lambda}_{(k)} - \left[\nabla^2 f(\boldsymbol{\lambda}_{(k)})\right]^{-1} \nabla f(\boldsymbol{\lambda}_{(k)})$.
Step 4. If $\| \nabla f(\boldsymbol{\lambda}_{(k+1)}) \| < \epsilon$, then $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_{(k+1)}$;
otherwise, set $k=k+1$ and return to Step 2.
Step 5. Output the optimal solution $\boldsymbol{\lambda}^*$.

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6]^T$ is the multiplier. Then the optimal solution $\mathbf{q}^*$ can be obtained by solving the equations:

$$\nabla_{q_j} L(\mathbf{q}, \boldsymbol{\lambda}) = -\ln q_j - 1 - \boldsymbol{\lambda}^T \bar{\mathbf{V}}_j = 0, \quad j = 1, 2, \ldots, 24, \tag{8}$$

where $\bar{\mathbf{V}}_j$ represents the j-th column of $\bar{\mathbf{V}}$. Thus, the optimal solution is

$$q_j^* = e^{-1-\boldsymbol{\lambda}^T \bar{\mathbf{V}}_j} > 0, \quad j = 1, 2, \ldots, 24. \tag{9}$$

Substitute Eq. (9) into Eq. (7), we have

$$\min_{\boldsymbol{\lambda}} \left\{ \sum_{j=1}^{24} e^{-1-\boldsymbol{\lambda}^T \bar{\mathbf{V}}_j} + \boldsymbol{\lambda}^T \bar{\mathbf{p}} \right\}. \tag{10}$$

Then Model (6) is transformed into Model (10), which can be solved by the Newton's method by iterating the multipliers $\boldsymbol{\lambda}$. Note that Problem (10) is a minimization problem without constraints and the number of variables is reduced from 24 (original problem) to 6. Denote

$$f(\boldsymbol{\lambda}) = \sum_{j=1}^{24} e^{-1-\boldsymbol{\lambda}^T \bar{\mathbf{V}}_j} + \boldsymbol{\lambda}^T \bar{\mathbf{p}}.$$

Then the gradient of $f(\boldsymbol{\lambda})$ can be expressed as

$$\nabla f(\boldsymbol{\lambda}) = -\bar{\mathbf{V}}\mathbf{q}^* + \bar{\mathbf{p}},$$

and the Hessian of $f(\boldsymbol{\lambda})$ is expressed as

$$\nabla^2 f(\boldsymbol{\lambda}) = \bar{\mathbf{V}} \mathrm{diag}(\mathbf{q}^*) \bar{\mathbf{V}}^{\mathrm{T}},$$

where diag($\mathbf{q}^*$) is a diagonal matrix with diagonal entries $\mathbf{q}^*$. It is clear that $\nabla^2 f(\boldsymbol{\lambda})$ is a positive definite matrix because all the entries of $\mathbf{q}^*$ are positive and $\bar{\mathbf{V}}$ is a full rank matrix. Suppose that $\boldsymbol{\lambda}_{(k)}$ is the k-th iteration solution of (10), $k = 0, 1, 2, \ldots$, where $\boldsymbol{\lambda}_{(0)}$ is the initial solution. The Newton's iteration formula is given as follows:

$$\boldsymbol{\lambda}_{(k+1)} = \boldsymbol{\lambda}_{(k)} - \left[\nabla^2 f(\boldsymbol{\lambda}_{(k)})\right]^{-1} \nabla f(\boldsymbol{\lambda}_{(k)}). \tag{11}$$

Based on the above iteration formula, a sequence of points $\boldsymbol{\lambda}_k$ are generated until the norm of gradient $\|\nabla f(\boldsymbol{\lambda}_k)\| = 0$ or less than a given small enough threshold ϵ. In our numerical experiment, we set $\epsilon = 0.0001$. Then the optimal solution $\boldsymbol{\lambda}^*$ is obtained. The steps are shown in Algorithm 4.1.

5. Numerical example

This section presents a real GPS trajectory data segmentation problem to validate the effectiveness of PLDSM. By comprehensively analyzing the GPS trajectory data and historical business affair records, all the business points are filtered out and the business affair categories are inferred by PLDSM. As a comparison, a widely applied machine learning technique KNN is used to tackle with the same trajectory data segmentation problem. The results reveal that PLDSM performs better in effectiveness and practicability.

Example 5.1. Suppose that the fourth party logistics is greatly concerned about the business affair information of the cooperative third party logistics vehicle with ID 238245, which transports various cargoes to different retailing depots every day in Shanghai. Every time there is a cargo discharging, a business affair is completed, and then the vehicle driver would record the business affair information including the arrival time, departure time and location. The fourth party logistics has no access to all the business affair information except some limited historical business data from 2014/5/10 00:00:00 to 2014/5/17 24:00:00 derived by surveying the driver of this vehicle (see Appendix). In the meanwhile, all the GPS trajectory data is available by the GPS receiver installed in this vehicle. To verify the effectiveness of PLDSM, we take the GPS trajectory data and historical business data from 2014/5/10 00:00:00 to 2014/5/17 24:00:00 as the training dataset to construct a

Table 3
Frequency distribution function of residence time.

δ (s)	f^*	f	$F(t)$
$t < 600$	0	79	0.0000
$600 \le t < 900$	10	15	0.6667
$900 \le t < 1200$	8	11	0.7273
$1200 \le t < 1500$	3	5	0.6000
$1500 \le t < 1800$	8	15	0.5333
$1800 \le t < 2100$	2	5	0.4000
$2100 \le t < 2400$	1	2	0.5000
$2400 \le t < 2700$	2	2	1.0000
$2700 \le t < 3000$	0	0	0.0000
$3000 \le t < 3300$	1	1	1.0000
$3300 \le t < 3600$	1	1	1.0000
$t \ge 3600$	0	7	0.0000

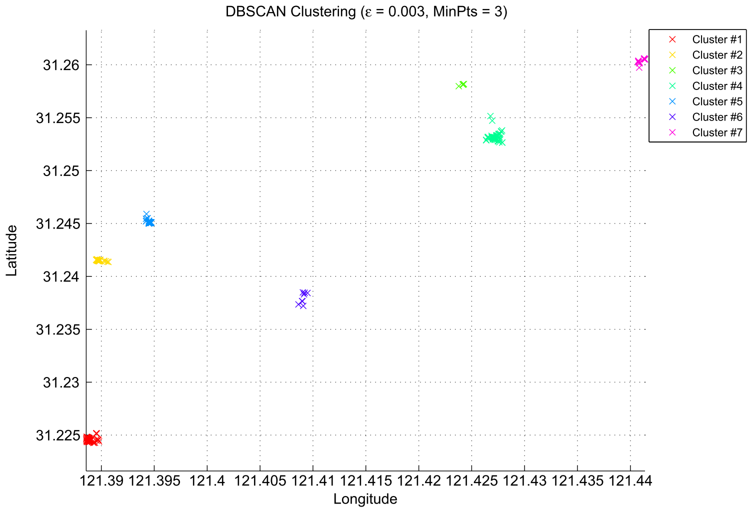

Fig. 14. Clusters of historical business data.

trajectory data segmentation mechanism, then apply it to segment a new GPS data sequence of this vehicle from 2014/5/18 00:00:00 to 2014/5/19 24:00:00 and compare the results with the surveyed business affair information.

In the training stage, we firstly detect all the stopping points from 2014/5/10 00:00:00 to 2014/5/17 24:00:00, and statistic the frequency of stopping points and business points. According to Eq. (1), the distribution function of residence time is obtained (see Table 3 above).

Secondly, the electronic fence is built by using DBSCAN to cluster the historical business data. Remark that there are 36 available business points (see Appendix). However, due to the data perturbation caused by the inexact positioning, each business point has multiple slightly different locations. To fully use the business affair information and achieve a better efficiency of clustering, we take all the varying locations into consideration to cluster the business points. Set the parameters in DBSCAN as follows: $\varepsilon = 0.003$ and $Minpts = 3$. Then the business points are divided into seven clusters: $C_1 = \{1, 7, 13, 16, 19, 25, 30, 34\}$, $C_2 = \{2, 8, 20, 26, 31, 35\}$, $C_3 = \{3\}$, $C_4 = \{4, 10, 14, 17, 22, 27, 32, 36\}$, $C_5 = \{6, 11, 18, 23, 28, 33\}$, $C_6 = \{12, 24, 28\}$ and $C_7 = \{5, 9, 15, 21\}$ (see Fig. 14). For each cluster, the electronic fence is built (see Fig. 15).

After the above training stage where the residence time frequency distribution and electronic fence are built, we start to segment the data sequence from 2014/5/18 00:00:00 to 2014/5/19 24:00:00. According to the data segmentation procedure in Fig. 4, we firstly detect all the stopping points in this data stream. As shown in Appendix, the minimum residence time of all the business points is 605 s. Therefore, the threshold in the stopping point detection process is set as 605 s since the stopping point cannot be a business point if its residence time is less than 605 s. By the stopping point identification process, 14 stopping points are obtained. Then we estimate the probability of residence time satisfying the business demand p_3 (see Table 4). Secondly, check the electronic lock status and fueling tank cap status of each stopping point. For vehicle 238245, no electronic lock data and electronic fueling data is available, then the values of p_1 and p_2 are zeros.

In the next step, for each stopping point, we calculate its probability of residence location satisfying the business demand (see p_4 in Table 4). It can be seen clearly that the locations of seven stopping points lie in the electronic fence (see Fig. 16). Take stopping point 3 as an example, due to data perturbation, it has 18 varying locations, 16 of which lie in the electronic fence, and 2 lie outside the electronic fence. Therefore, $p_4 = 16/18 = 0.8889$.

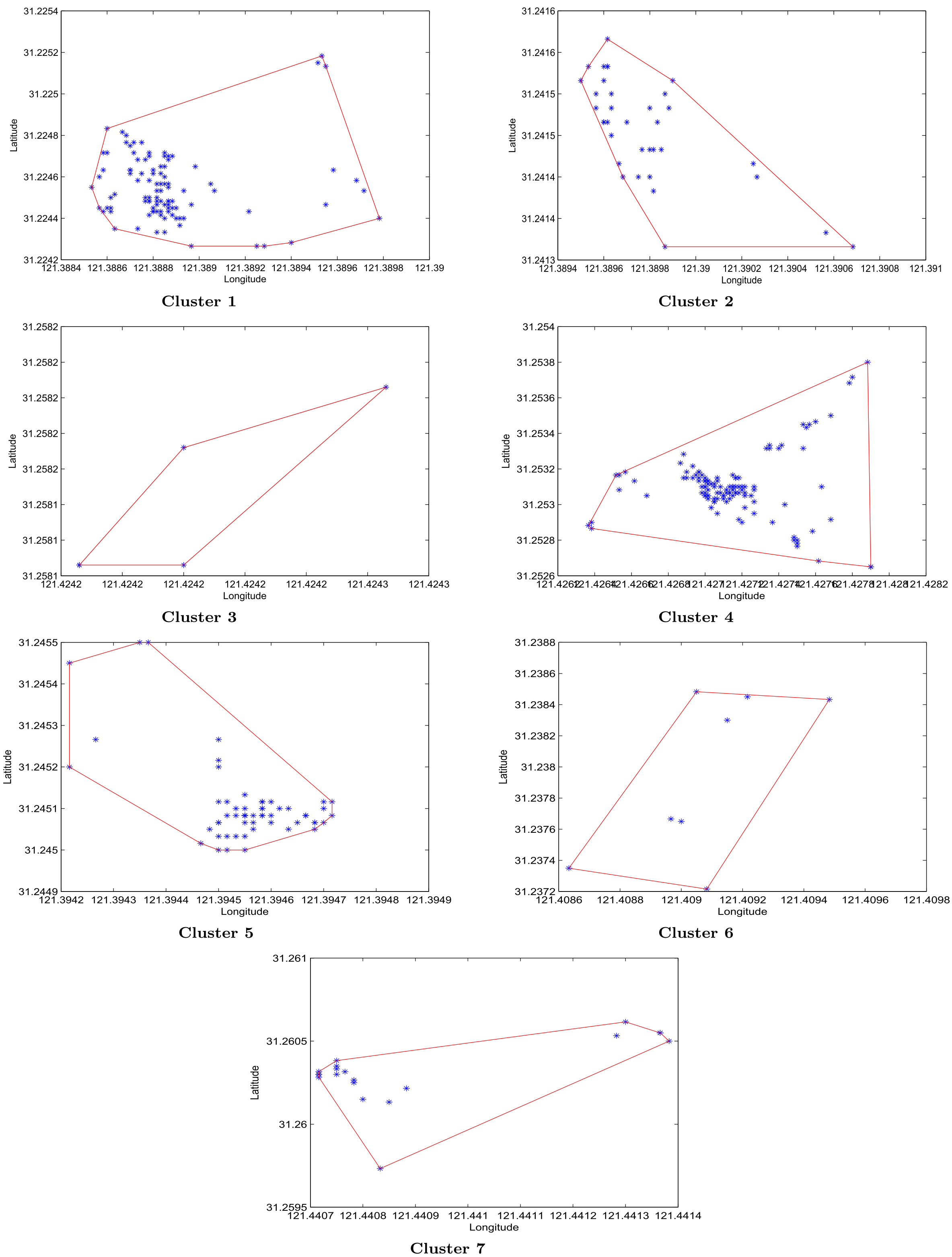

Fig. 15. Electronic fences of different clusters.

Table 4
GPS trajectory data segmentation results by PLDSM.

No.	δ (s)	p_1	p_2	p_3	p_4	p_5	p_6
1	1123	0.0000	0.0000	0.7273	0.0000	1.0000	0.7273
2	1902	0.0000	0.0000	0.4000	0.0000	1.0000	0.4000
3	2486	0.0000	0.0000	1.0000	0.8889	1.0000	1.0000
4	766	0.0000	0.0000	0.6667	0.6000	1.0000	0.8667
5	2181	0.0000	0.0000	0.5000	1.0000	0.0000	0.5000
6	1356	0.0000	0.0000	0.6000	0.8462	0.0000	0.5077
7	60870	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	1184	0.0000	0.0000	0.7273	0.0000	1.0000	0.7273
9	2423	0.0000	0.0000	1.0000	0.0000	1.0000	1.0000
10	1178	0.0000	0.0000	0.7273	0.0000	0.0000	0.0000
11	1777	0.0000	0.0000	0.5333	0.9167	1.0000	0.9616
12	723	0.0000	0.0000	0.6667	1.0000	1.0000	1.0000
13	1472	0.0000	0.0000	0.6000	1.0000	0.0000	0.6000
14	1079	0.0000	0.0000	0.7273	0.0000	0.0000	0.0000

Table 5
The confusion matrix.

True class	Recognized class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Finally, check whether the trajectory at each stopping point satisfies the business demand. According to the heading change of the vehicle, it is found that the trajectories of stopping points 1, 2, 3, 4, 8, 9, 11 and 12 satisfy the business demand (see Fig. 17). Take stopping point 1 as an example, at first, the heading of vehicle 238245 is around 100°. At point A, the vehicle turns left with the heading changing to 350°. After arriving at B, it stops with heading changing to 0° and then stays unmoving for a certain period. From B to C, the heading changes from 0° to 170°, representing that the vehicle completes a u-turn. Finally, the vehicle turns left again at point D and then leaves in the direction of 90°. Therefore, the trajectory of stopping point 1 satisfies the business demand. The trajectories before and after those stopping points can be seen in Fig. 18, where the red and green curves denote the trajectory before and after the stopping point respectively.

After deriving the corresponding probabilities: p_1, p_2, p_3, p_4 and p_5, the value of p_6 is obtained by solving model (6) with Newton's method. The last column of Table 4 lists the final probabilities of these stopping points being business points. Assign the stopping points with $p_6 \geq 0.5$ as business points, then stopping points 1, 3, 4, 5, 6, 8, 9, 11, 12 and 13 are recognized as business points, where the ones marked with red are true business points in reality. There are two commonly accepted criteria measuring the efficiency of data segmentation: precision and recall, which are defined based on the following confusion matrix (see Table 5), where TP denotes true positive business points, representing the number of real business points which are recognized correctly, FN refers to the number of real business points wrongly recognized as non-business points, FP stands for the number of real non-business points wrongly recognized as business points and TN represents the number of real non-business points recognized correctly. The precision (Pre) and recall (Rec) are expressed as follows:

$$Pre = \frac{TP}{TP + FP}, \quad Rec = \frac{TP}{TP + FN}. \tag{12}$$

According to Table 4 and Eq. (12), the precision for PLDSM is 70%, and the recall is 100%. As comparison, we employ the most widely applied machine learning technique KNN to filter out the business points from the 14 stopping points in Table 4. Using the same training dataset including longitude x^*, latitude y^*, residence time δ and trajectory p_5, a KNN classifier is constructed by employing the Euclidean distance and setting parameter $\mathrm{K} = 3$. The segmentation result is shown in the last column of Table 6, where 1 implies that the stopping point is recognized as business point and 0 represents that the stopping point is recognized as non-business point.

The results in Table 6 reveal that stopping points 3, 5, 6, 11, 12, 13 and 14 are recognized as business points by KNN while the real business points are 3, 4, 5, 6, 11, 12 and 13. Therefore, the precision and recall for KNN are both 86%. Although KNN performs better in precision compared with PLDSM, it cannot filter out all the real business points, i.e., business point 4. In contrast, the recall of PLDSM is 100%, representing PLDSM is able to find all the business points from the data sequence. In real GPS trajectory data segmentation application, the fourth party logistics concerns most about the recall of segmentation rather than precision since omitting any potential business point could significantly influence the inference over the whole business affair and sometimes even lead to a wrong judgment in business affair analysis. Therefore, achieving a larger recall is the most important objective to cover all the real business points. In case that all the real business points are covered, we turn to the comparison in precision.

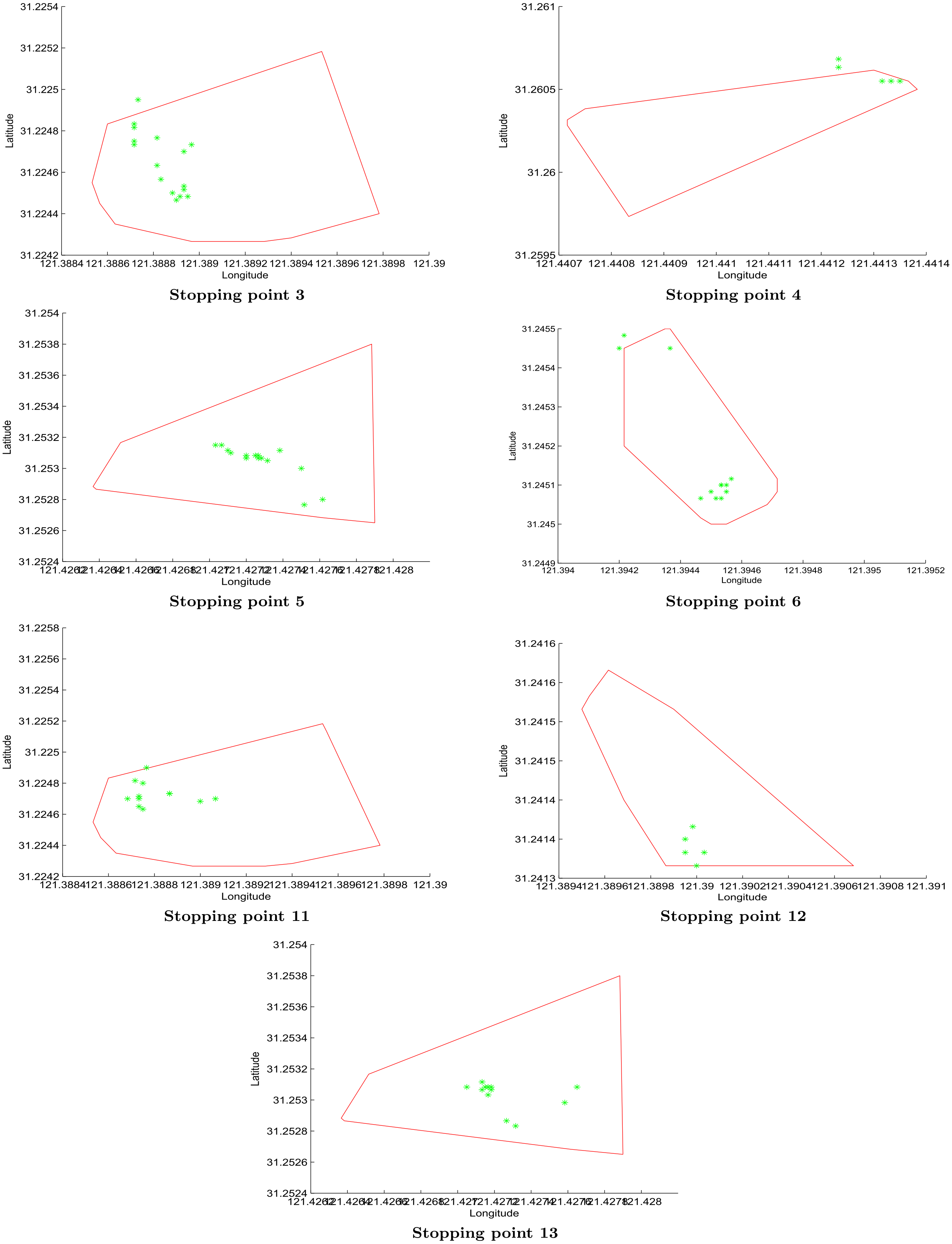

Fig. 16. Stopping points in the electronic fence.

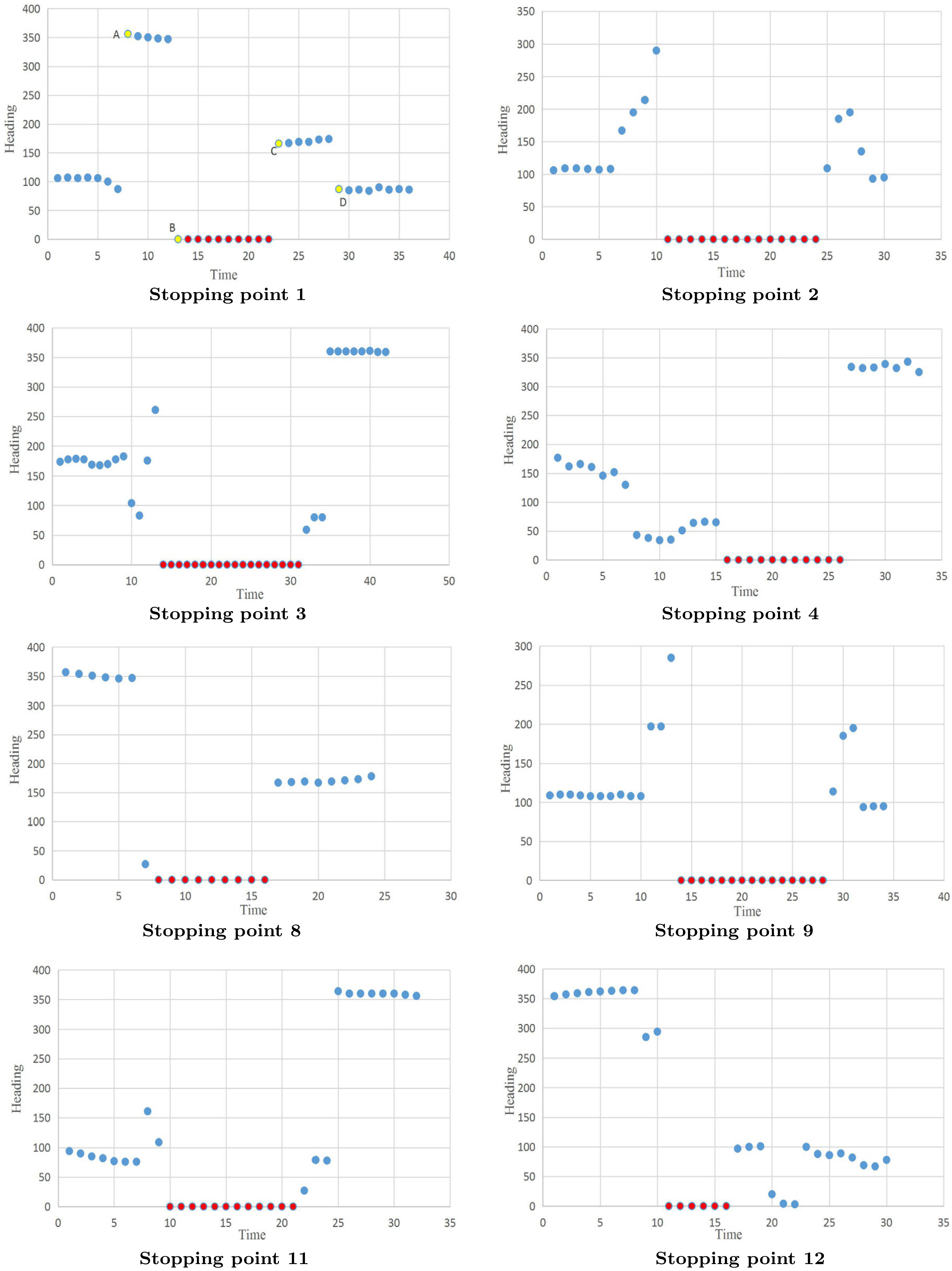

Fig. 17. Heading changes at different stopping points.

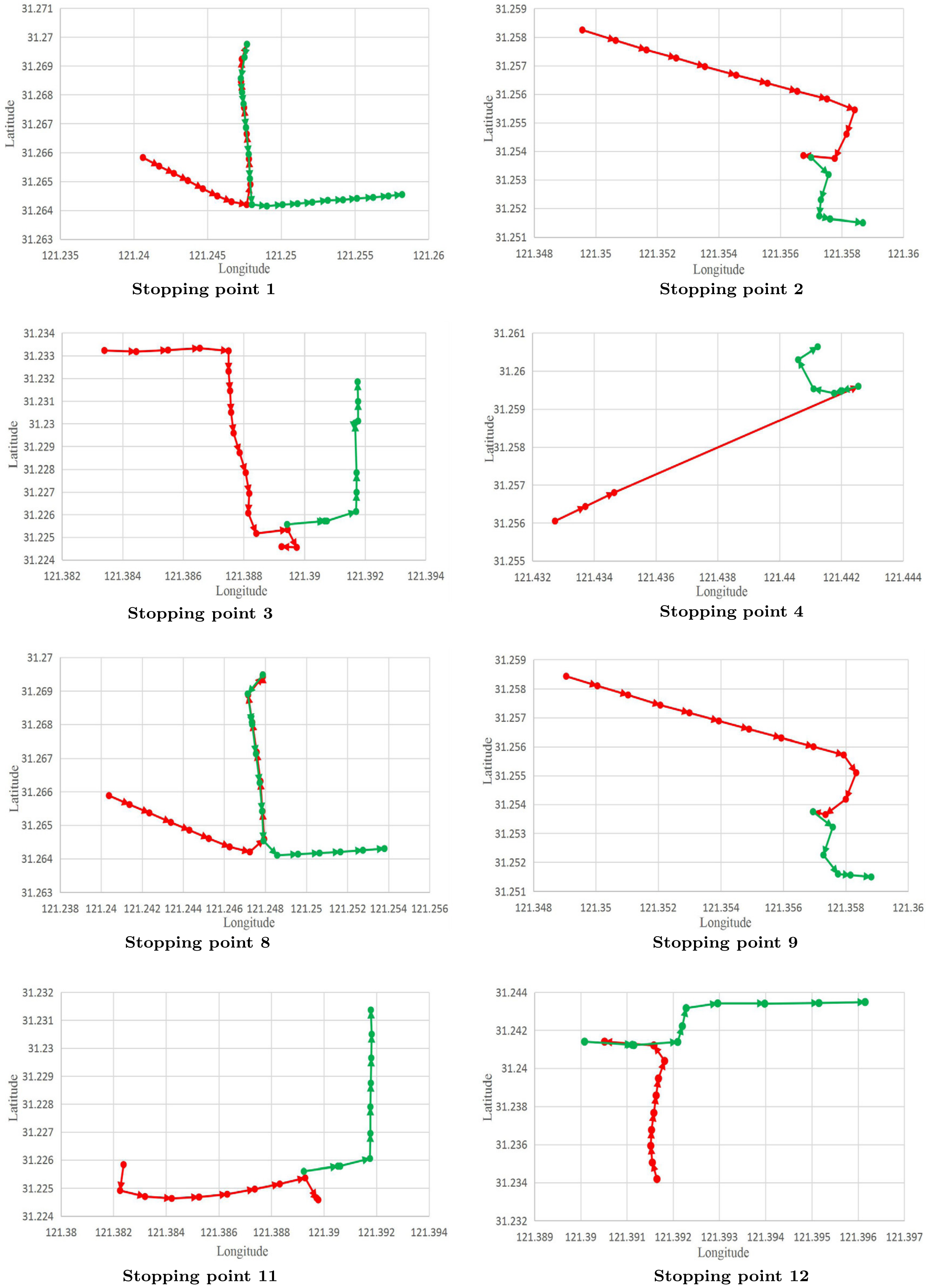

Fig. 18. Trajectories before and after stopping points.

Table 6
GPS trajectory data segmentation results by KNN.

No.	x^*	y^*	δ	p_5	Output
1	121.247683	31.269750	1123	1.0000	0.0000
2	121.356750	31.253850	1902	1.0000	0.0000
3	121.389250	31.224583	2486	1.0000	1.0000
4	121.434650	31.256800	766	1.0000	0.0000
5	121.427400	31.253633	2181	0.0000	1.0000
6	121.394233	31.245966	1356	0.0000	1.0000
7	121.241216	31.242183	60870	0.0000	0.0000
8	121.247883	31.269433	1184	1.0000	0.0000
9	121.357350	31.253650	2423	1.0000	0.0000
10	121.370883	31.248650	1178	0.0000	0.0000
11	121.389716	31.224650	1777	1.0000	1.0000
12	121.390516	31.241400	723	1.0000	1.0000
13	121.427616	31.253100	1472	0.0000	1.0000
14	121.394250	31.245616	1079	0.0000	1.0000

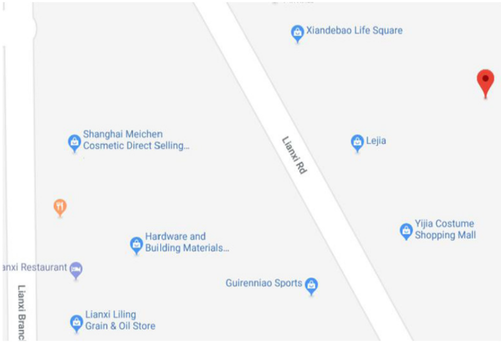

Fig. 19. Location for stopping point 1 and 8.

Fig. 20. Location for stopping point 9.

In addition, it follows from Table 4 that stopping points 1, 8 and 9 are "wrongly" recognized as business points by PLDSM, which are not the known business points according to the surveyed business data. Then we check the residence time and residence locations of these three points, and study their GPS trajectory. Interestingly, it is found that this vehicle stops around these locations for a certain period every day. Stopping points 1 and 8 are in the same location shown in Fig. 19, which is nearby several markets, squares and malls. Every day, the vehicle drives to this location, stays for around 20 minutes, then takes a u-turn with a change of 180 in direction and leaves. As regard to stopping point 9, we search its location (see Fig. 20) by Google map, and find that it is close to a supermarket RT Mart. The GPS trajectory data reveals that this vehicle usually stops at this area at around 2:00:00, stays for 30 minutes and leaves. It is known that the replenishment of supermarkets is conducted during the midnight to avoid the negative impact on the shopping activities of customers in the daytime. And fresh food such as meat, vegetables and fruits are usually pursued and replenished in the early morning to ensure the freshness of food. Therefore, according to the time and location, we can infer that stopping points 1, 8 and 9 are probably business points, and this vehicle is intended for replenishment of malls and supermarkets. The core cause of these three exceptions is the incompleteness of business data due to some particular business affairs may be deliberately concealed by the driver. Then the real precision for PLDSM is actually 100% since all the business points are detected, which reflects the benefits of incorporating common logic and real driving experiences in GPS trajectory data segmentation. Generally speaking, our proposed PLDSM not only can segment the GPS trajectory data and find all the business points accurately, but also assists to infer the business affair categories.

6. Conclusions

GPS trajectory data segmentation is attracting attention of researchers and practitioners for its efficiency and necessity in extracting business affair information from raw data. The unavailability of GIS data and information loss in data collection process significantly increase the difficulty in real trajectory segmentation issues. In this paper, we consider a real application of data segmentation in logistics transportation. Based on limited historical business data, the business time window and electronic fence are built to analyze the driving habits of vehicles. An efficient data segmentation mechanism is formulated by sequentially considering electronic lock data, fueling tank cap data, residence time, residence location and trajectory. More importantly, we firstly incorporate the probabilistic logic into data segmentation and integrate the duality theory and Newton's method to obtain the optimal solution. As application, we solve a real GPS trajectory data segmentation problem by PLDSM and verify its advantages over KNN. The results greatly promote the practicability of probabilistic logic in data segmentation.

Frankly speaking, this study only focuses on the GPS trajectory data segmentation of one vehicle. In real application, the fourth party logistics may be confronted with segmenting the trajectory data of multiple vehicles simultaneously. The challenge is that the correlation of different logistics vehicle trajectories should be considered separately since different vehicles may transport the cargoes to the same depot or only one vehicle is assigned to a certain depot. Therefore, it will be attractive to continue the multi-vehicle GPS trajectory data segmentation study to strengthen the practicability. In the future, we will consider the network based GPS trajectory data segmentation to tackle with the above issue.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 71722007), the Research Grants Council of Hong Kong (No. 15210815), IMR and RAE Research Fund from Faculty of Science, the University of Hong Kong.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijar.2018.09.008.

References

[1] S. An, H.Q. Yang, J. Wang, N. Cui, J.X. Cui, Mining urban recurrent congestion evolution patterns from GPS-equipped vehicle mobility data, Inf. Sci. 373 (2016) 515–526.
[2] A. Anagnostopoulos, M. Vlachos, M. Hadjieleftheriou, E. Keogh, P.S. Yu, Global distance-based segmentation of trajectories, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 34–43.
[3] B. Aronov, A. Driemel, M. Van Kreveld, M. Löffler, F. Staals, Segmentation of trajectories on nonmonotone criteria, ACM Trans. Algorithms 12 (2015) 1–28.
[4] M. Benkert, J. Gudmundsson, F. Hübner, T. Wolle, Reporting flock patterns, Comput. Geom. 41 (3) (2008) 111–125.
[5] J.V. Braun, H.G. Muller, Statistical methods for DNA sequence segmentation, Stat. Sci. 13 (2) (1998) 142–162.
[6] M. Buchin, A. Driemel, M.J. Van Kreveld, V. Sacristan, Segmenting trajectories: a framework and algorithms using spatiotemporal criteria, J. Spat. Inf. Sci. 3 (2011) 33–63.
[7] M. Buchin, H. Kruckenberg, A. Kölzsch, Segmenting trajectories by movement states, in: Advances in Spatial Data Handling, 2012, pp. 15–25.
[8] D. Buchman, D. Poole, Negative probabilities in probabilistic logic programs, Int. J. Approx. Reason. 83 (2017) 43–59.
[9] Y.J. Byon, B. Abdulhai, A. Shalaby, Real-time transportation mode detection via tracking global positioning system mobile devices, J. Intell. Transp. Syst. 13 (4) (2009) 161–170.
[10] W. Chen, M.H. Ji, J.M. Wang, T-DBSCAN: a spatiotemporal density clustering for GPS trajectory segmentation, Int. J. Online Eng. 10 (6) (2014) 19–24.
[11] W. Ching, S. Scholtes, S. Zhang, Numerical algorithms for estimating traffic between zones in a network, Eng. Optim. 36 (2004) 379–400.
[12] S.D. Clercq, S. Schockaert, A. Nowé, M.D. Cock, Modelling incomplete information in Boolean games using possibilistic logic, Int. J. Approx. Reason. 93 (2018) 1–23.
[13] S. Dabiri, K. Heaslip, Inferring transportation modes from GPS trajectories using a convolutional neural network, Transp. Res., Part C 86 (2018) 360–371.
[14] M.L. Damiani, F. Hachem, Segmentation techniques for the summarization of individual mobility data, WIREs Data Min. Knowl. Discov. (2017), https://doi.org/10.1002/widm.1214.
[15] M.L. Damiani, H. Issa, F. Cagnacci, Extracting stay regions with uncertain boundaries from GPS trajectories: a case study in animal ecology, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2014, pp. 253–262.
[16] T. Feng, H.J.P. Timmermans, Transportation mode recognition using GPS and accelerometer data, Transp. Res., Part C 37 (2013) 118–130.
[17] K.S. Fu, J.K. Mui, A survey on image segmentation, Pattern Recognit. 13 (1) (1981) 3–16.
[18] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 330–339.
[19] P.R. Gillett, R.B. Scherl, G. Shafer, A probabilistic logic based on the acceptability of gambles, Int. J. Approx. Reason. 44 (2007) 281–300.
[20] L. Gong, H. Sato, T. Yamamoto, T. Miwa, T. Morikawa, Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines, J. Mod. Transp. 23 (3) (2015) 202–213.
[21] B. Jaumard, A.D. Parreira, An anytime deduction algorithm for the probabilistic logic and entailment problems, Int. J. Approx. Reason. 50 (2009) 92–103.
[22] H. Jeung, H.T. Shen, X. Zhou, Convoy queries in spatio-temporal databases, in: IEEE 24th International Conference on Data Engineering, 2008, pp. 1457–1459.

[23] H. Jeung, M. Yiu, X. Zhou, C. Jensen, H. Shen, Discovery of convoys in trajectory databases, Proc. VLDB Endow. 1 (1) (2008) 1068–1080.
[24] M. Kose, O.D. Incel, C. Ersoy, Online human activity recognition on smart phones, in: Proceedings of Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data, 2012, pp. 11–15.
[25] J. Krumm, E. Horvitz, Predestination: inferring destinations from partial trajectories, in: Proceedings of the 8th International Conference on Ubiquitous Computing, 2006, pp. 243–260.
[26] P. Laube, S. Imfeld, Analyzing relative motion within groups of trackable moving point objects, Geogr. Inf. Sci. 2478 (2002) 132–144.
[27] P. Laube, M. Van Kreveld, S. Imfeld, Finding REMO-detecting relative motion patterns in geospatial lifelines, in: Proceedings of the 11th International Symposium on Spatial Data Handling, 2004, pp. 201–214.
[28] L.A. Leiva, E. Vidal, Revisiting the K-means algorithm for fast trajectory segmentation, in: Proceedings of the 38th International Conference and Exhibition on Computer Graphics and Interactive Techniques, 2011.
[29] L.A. Leiva, E. Vidal, Warped K-Means: an algorithm to cluster sequentially-distributed data, Inf. Sci. 237 (2013) 196–210.
[30] Z. Li, B. Ding, J. Han, R. Kays, Swarm: mining relaxed temporal moving object clusters, Proc. VLDB Endow. 3 (1) (2010) 723–734.
[31] T. Lukasiewicz, Nonmonotonic probabilistic logics under variable-strength inheritance with overriding: complexity, algorithms, and implementation, Int. J. Approx. Reason. 44 (2007) 301–321.
[32] N.J. Nilsson, Probabilistic logic, Artif. Intell. 28 (1) (1986) 71–87.
[33] A.T. Palma, V. Bogorny, B. Kuijpers, L.O. Alvares, A clustering-based approach for discovering interesting places in trajectories, in: Proceedings of the ACM Symposium on Applied Computing, 2008, pp. 863–868.
[34] C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, Y. Theodoridis, Segmentation and sampling of moving object trajectories based on representativeness, IEEE Trans. Knowl. Data Eng. 24 (7) (2012) 1328–1343.
[35] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M.L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, Z.X. Yan, Semantic trajectories modeling and analysis, ACM Comput. Surv. 45 (4) (2013) 1–32.
[36] D.J. Patterson, L. Liao, D. Fox, H. Kautz, Inferring high-level behavior from low-level sensors, in: International Conference on Ubiquitous Computing, vol. 2864, 2003, pp. 73–89.
[37] N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos, Y. Theodoridis, Clustering uncertain trajectories, Knowl. Inf. Syst. 28 (1) (2011) 117–147.
[38] F. Riguzzi, E. Bellodi, R. Zese, G. Cota, E. Lamma, A survey of lifted inference approaches for probabilistic logic programming under the distribution semantics, Int. J. Approx. Reason. 80 (2017) 313–333.
[39] H. Wang, G.J. Liu, J.Y. Duan, L. Zhang, Detecting transportation modes using deep neural network, IEICE Trans. Inf. Syst. 100 (5) (2018) 1132–1135.
[40] Z.X. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, K. Aberer, Semantic trajectories: mobility data computation and annotation, ACM Trans. Intell. Syst. Technol. 4 (3) (2013).
[41] H. Yoon, C. Shahabi, Robust time-referenced segmentation of moving object trajectories, in: Proceedings of the IEEE International Conference on Data Mining, 2008, pp. 1121–1126.
[42] H. Zhang, J.E. Fritts, S.A. Goldman, Image segmentation evaluation: a survey of unsupervised methods, Comput. Vis. Image Underst. 110 (2008) 260–280.
[43] Y. Zheng, Trajectory data mining: an overview, ACM Trans. Intell. Syst. Technol. 6 (3) (2015) 1–29.
[44] Y. Zheng, L. Liu, L. Wang, X. Xie, Learning transportation mode from raw GPS data for geographic application on the Web, in: Proceedings of the 17th International Conference on World Wild Web, ACM, 2008, pp. 247–256.
[45] Y. Zheng, Y. Chen, Q. Li, X. Xie, W.Y. Ma, Understanding transportation modes based on GPS data for web applications, ACM Trans. Web 4 (1) (2010) 1–36.
[46] K. Zheng, Y. Zheng, N.J. Yuan, S. Shang, X.F. Zhou, Online discovery of gathering patterns over trajectories, IEEE Trans. Knowl. Data Eng. 28 (5) (2014) 1974–1988.