Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 3634942, 15 pages
https://doi.org/10.1155/2019/3634942

Research Article

An Automatic Extraction Method of Coach Operation Information from Historical Trajectory Data

Jun Li[1,2] **Qingqi Li**[1] **Yan Zhu**[1] **Yan Ma**[3] **Yubin Xu**[3] **and Chao Xie**[2]

[1]*College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China*
[2]*National Engineering Laboratory for Transportation Safety and Emergency Informatics, China Transport Telecommunications & Information Center, Beijing, China*
[3]*China Academy of Civil Aviation Science and Technology, Beijing, China*

Correspondence should be addressed to Jun Li; junli_geo@126.com

Received 28 October 2018; Revised 13 January 2019; Accepted 23 January 2019; Published 11 February 2019

Academic Editor: David F. Llorca

Quality of travel service for road transport relies heavily on richness of transport operation data. Currently, most types of data including coach operation data are collected by manual investigation which is time-consuming and labor-intensive, and this significantly hinders the realization of intelligent traffic information service. In view of the above problems, this paper is aimed at introducing a method of automatically extracting coach operation information using historical GPS trajectory data of massive coaches. The method first analyzes trajectory characteristics of coaches within stations and identifies the highly dense point clusters as coach stations using the DBSCAN clustering algorithm. Then the schedule information is obtained by conducting error adjustment on the actual arrival and departure time series of multiple shifts, and the name of coach station is queried from point of interest (POI) and geographical name database provided by online map. Finally, the regular driving route of coaches is extracted by an incremental trajectory merging method. The proposed method is applied in handling historical trajectory data in the Beijing-Tianjin-Hebei region in China, and experimental results show that the extraction accuracy is 84% and verify its effectiveness and feasibility. The proposed method makes use of data mining techniques to extract coach operation information from big trajectory data and saves a lot of labor work, time, and economic cost required by on-site investigation.

1. Introduction

Intelligent traffic information service is to make use of advanced technologies (internet of things, cloud computing, etc.), transport operation data, analysis, and decision algorithms to obtain traffic knowledge for providing decision-making services for government, enterprises, and the public. Typical examples are real-time traffic monitoring [1, 2], transit planning [3, 4], travel time planning [5, 6], and online ticket inquiry and booking system [7]. Since the 20th century, intelligent traffic information services have been widely used in various fields and become an important component and hot research direction in intelligent transportation field [8, 9]. Quality of intelligent traffic information service is affected not only by the accuracy and efficiency of analysis algorithms, but also by the completeness and quality of transport operation data [10]. Scholars have carried out a lot of research work on traffic analysis and decision algorithms [11]; however, few studies are focused on transport operation data collection, especially improving the automation level, and accuracy of data collection process. There are mainly two types of collection methods for transport operation data: one is collected by monitoring equipment, such as a camera or an induction coil to obtain traffic flow, while the other is collected by manual investigation, such as the coach operation information, including station location, name, schedule, and operation route. In real world, most of transport operation data need to be collected through manual work. Taking the coach operation information in China as an example, a large number of passenger transport enterprises are involved in this industry, and the mode of operation in road passenger stations is isolated and low standardized [12]. It would be a very time-consuming and labor-intensive task to manually collect detailed coach operation information in a large area

or even across the country. Therefore, to realize intelligent traffic information service for road passenger transport, there is an urgent need for the methods of collecting detailed coach operation information in a quick, efficient, and low cost way.

With the development of sensing and location technologies, a lot of vehicles are installed GPS (Global Positioning System) receivers and wireless communication equipment. The vehicles are continuously collecting real-time information including locations, motion parameters, and positioning time while moving and then transmit them to the data center. This type of data is called floating car data [13, 14]. Massive vehicle positioning data is increasingly accumulated which creates a new way to solve the above challenges [15]. The emergence of these data makes it possible to make use of big data analysis technologies to mine rich knowledge from trajectory data [16], including operation information of passenger vehicles.

The research work related to this paper is mainly divided into three aspects. One is the extraction of regions of interest based on trajectory data. Palma et al. and Bhattacharya et al. proposed methods of analyzing the movement characteristics (velocity, azimuth, acceleration, etc.) at specific locations and found interesting/important places related to person/object [17, 18]. Many scholars further explored region of interest for various applications, and, among those, Zheng et al. used trajectory mining algorithms to analyze user-generated GPS trajectory data for recommending tourist attractions and personalized tourist attractions [19]. Besides, Li et al. used the DBSCAN algorithm to extract parking lot locations by analyzing the typical characteristics of floating car data within parking lots [20].

Another aspect is route extraction based on trajectory data. Schoredl et al. and Li et al. proposed two methods of extracting high-precision road maps from trajectories of the vehicles equipped with GPS receivers [21, 22], and their methods are suitable for high and low sampling frequency of location data, respectively. Different from theirs, Cao and Krumm proposed a novel gravitational model to convert original GPS trajectory into a road network that can guide the path selection [23]. Aiming at extracting more detailed information, Tang et al. proposed a method for mining lane-level road network information from low-precision vehicle GPS trajectories based on the number of traffic lanes and steering rules [24]. Kuntzsch et al. established an explicit crossover model with a fractional function to extract traffic network from GPS data and proved the feasibility of the method through experiments on GPS datasets of different sizes and data quality [25].

In addition to region of interest and route identification, scholars have also done a lot of research on extraction of spatiotemporal patterns based on trajectory data. For example, Kang and Yong proposed a method of finding spatiotemporal patterns in trajectory data, which first discovered meaningful spatiotemporal regions and then extracts frequent spatiotemporal patterns using a prefix-projection approach [26]. Similarly, Lei et al. [27] and Zhao et al. [28] proposed space-time analytical model/framework to capture movement patterns of objects. Different from those, Lu et al. proposed a visual analysis method to study the behavior of vehicles along a route, focusing on the temporal and spatial distribution of travel time, that is, the time spent on each road segment and the travel time change during peak/off-peak hours [29]. Temporal and spatial patterns of moving objects extracted from trajectory data can also be used for travel planning recommendations; for example, Zheng et al. [30] and Hsieh et al. [31] extracted interesting locations and activities from trajectory data and then recommended travel routes. Some scholars have also carried out work on periodic pattern recognition such as a probabilistic period detection method for moving objects [32]. Although scholars have proposed various methods on trajectory data to identify region of interest, route of interest, and spatiotemporal patterns, few studies have explored the comprehensive use of data mining methods to extract coach operation information.

This paper is aimed at introducing a method of automatically extracting coach operation information using historical GPS trajectory data of massive coaches. After analyzing the typical characteristics of GPS trajectories collected within stations, the DBSCAN spatial clustering algorithm is used to identify station location. Then, schedule information is obtained by conducting error adjustment on actual arrival and departure time series of multiple shifts, and coach station name is identified by point of interest (POI) and geographical name database provided by online map. Finally, the regular driving route of each coach line is extracted by an incremental trajectory merging algorithm. This method was promising to obtain coach operation information in a large area in a fast and low-cost way.

2. Methodology

The development and popularization of internet of vehicles have caused floating cars to generate a large amount of tracking data. Tracking data record the location and motion parameters of vehicles while vehicles are moving and are an indispensable data source for studying behavior of vehicles and for mining hidden information [33]. One among many floating car platforms is the National Commercial Vehicle Monitoring Platform (NCVMP) established by the Ministry of Transport of China in 2010 for monitoring several special types of commercial vehicles including coaches. For coaches, the most important information is coach operation information including the location and name of stations, schedules, and operation routes of coaches. Due to various kinds of reasons, the drivers of coaches may sometimes adjust their driving route according to road conditions or driving habits, so the route of different shifts of even the same coach or the arrival time at the same station is often not exactly the same. Therefore, how to accurately extract station locations, regular schedules and driving routes from historical trajectory data are the key to this research.

The overall workflow of extracting coach operation information from coach trajectory data is shown in Figure 1. It is mainly composed of four parts: location extraction of station, name identification of station, schedule extraction, and route generation. Firstly, the abnormal positioning points in the original dataset that are irregular or illogical are removed, and the trajectory points belonging to coach routes are separated

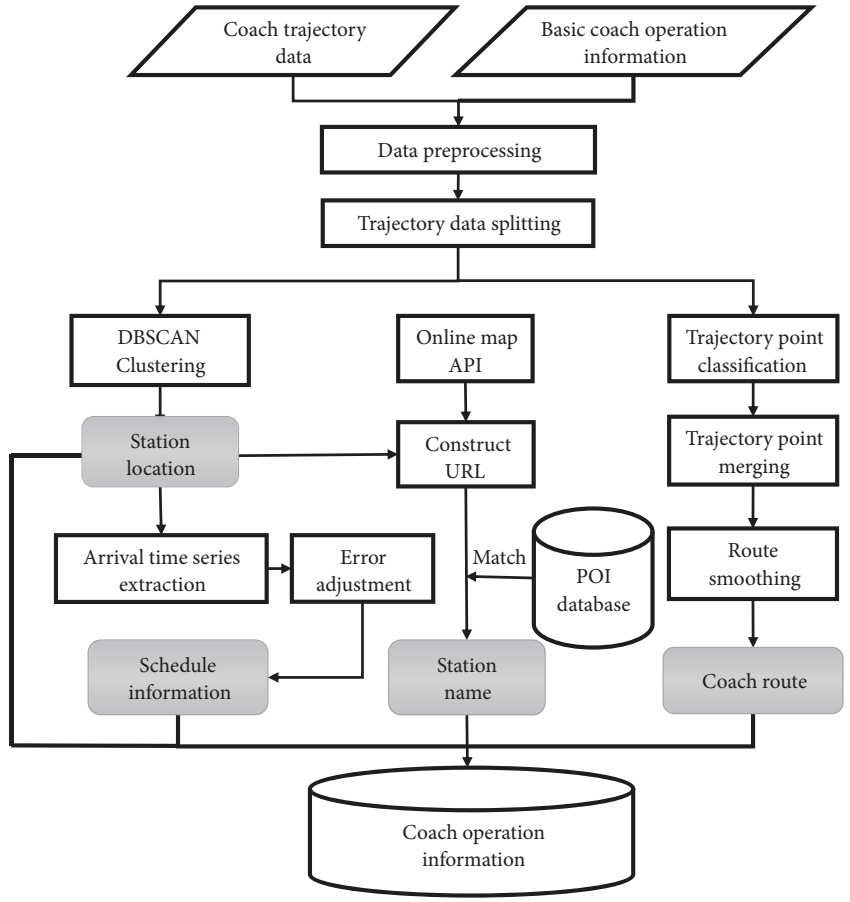


Figure 1: Overall workflow of extracting coach operation information based on trajectory data.

from the original dataset based on basic coach operation information. Secondly, this paper uses the DBSCAN algorithm to extract the location of stations where coaches park for passengers to get on and off and meanwhile calculates the regular arrival and departure time for each station using error adjustment theory. Thirdly, based on the place name and POI database of online maps, station name is obtained through the extracted station locations. In the procedure of driving route extraction, any one trajectory of a coach is first taken as a candidate route, and other trajectories are continuously updated to the candidate route by merging process including trajectory point classification and merging and confidence filtering, and finally the regular driving route of each coach is obtained.

For convenience, the historical trajectory point set of all coach routes is denoted as $CDL = \{L_1, L_2, \ldots, L_i, \ldots, L_m\}$, where L_i is the trajectory data of the i-th coach route. A trajectory is composed of a series of trajectory points and it is represented by $L_i = \{P_1, P_2, \cdots, P_j, \cdots, P_{|L_i|}\}$. The basic coach operation information set is denoted as $TSL = \{G_1, G_2, \ldots, G_j, \ldots, G_m\}$, where $G_i = \{S_i(x, y), E_i(x, y), t_i\}$ is the basic operation data of the i-th coach route, where $S_i(x, y)$ is the location of origin station, $E_i(x, y)$ is the location of destination station, and t_i is departure time of the i-th coach route.

2.1. Data Preprocessing. Due to the variety of in-vehicle positioning and communication equipment and the complicated and diverse driving environment of vehicles, the quality of trajectory data of coaches is uneven such as position drift, abnormal attribute values, or data irregularity. In addition, there are inevitable erroneous data and redundant data, and they affect the knowledge extraction process. Based on research needs, the original trajectory dataset is preprocessed from two aspects: data cleaning and coordinate transformation. First, the trajectory points with abnormal coordinate values are removed, including the cases where the coordinate value is 0 or missing. Second, the trajectory point whose speed is negative or direction exceeds 360 degrees is filtered out. Finally, the coordinates of original trajectory data are stored in the form of latitude and longitude, which is not convenient for distance calculation. Therefore, the coordinate transformation is converted from the WGS-84 coordinate system to the UTM projection coordinate system in order to facilitate extraction of coach stations and routes.

2.2. Trajectory Data Splitting. The GPS receiver mounted on the coach starts to collect data once turned on, and this causes that the original dataset contains not only the tracking points on a coach route, but also the tracking points not collected on any coach route, such as temporary vehicle

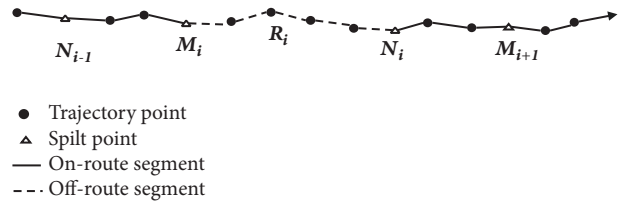


Figure 2: Trajectory data splitting.

scheduling or operating for other routes. The existence of the points on nontarget routes not only increases the workload of data processing, but also is not useful for the extraction of operation information. Therefore, the trajectory data set needs to be split before the operation information extraction. A coach route is determined by three factors: origin station, destination station, and departure time. Trajectory data splitting means separating the tracking points belonging to a route from the original vehicle trajectory data according to three factors of the coach route. This paper proposes a trajectory data splitting algorithm based on the linear sequence of start and end points. The workflow of the algorithm is as follows (Figure 2).

(1) *Split Point Identification*. First, for any coach route L_i, the threshold of matching distance for both origin station $S_i(x, y)$ and destination station $E_i(x, y)$ is set to d_s, and the threshold for the difference between the actual and planned departure time is set to t_s. Second, all trajectory points P_r in L_i are traversed. If a point P_r satisfies $dis(P_r, S_i(x, y)) \leq d_s$ and $timedif(P_r, t_i) \leq t_s$, then P_r is identified as a split point M_i of this coach route; otherwise if a point P_r satisfies $dis(P_r, E_i(x, y)) \leq d_s$ and $dis(P_r, E_i(x, y)) = min$, then P_r is identified as a split point N_i of the coach route.

(2) *Split Point Sequence Generation*. The above split point identification process is repeated until all split points are found, either M_i type or N_i type. According to the fact that the coach first passes through origin station and then arrives at destination station, and every two adjacent split points with a point of M_i type followed by a point of N_i type form a point pair(M_i, N_i). Finally, the split point sequence $W = \{(M_1, N_1), \ldots, (M_j, N_j), \ldots, (M_n, N_n)\}$ is generated.

(3) *On-Route Segment Splitting*. The trajectory points between any pair of split points in the point sequence W are regarded as on-route trajectory segment, while the rest of the points are regarded as off-route segment.

2.3. Coach Station Extraction. Coaches have two types of states throughout the entire operation: moving and docking. Analyzing and identifying the spatial characteristics of trajectory points in these two states are the key to determining location of coach stations. As shown in Figure 3, the trajectory points of coaches collected under moving status (Region A) are generally evenly spaced along the road, while those under docking status (Region B) are more aggregated, and a point cluster tends to appear since the coach would park here for a while. By comparing the spatial distribution and attribute characteristics of GPS trajectory points collected under different operating conditions, it can be concluded

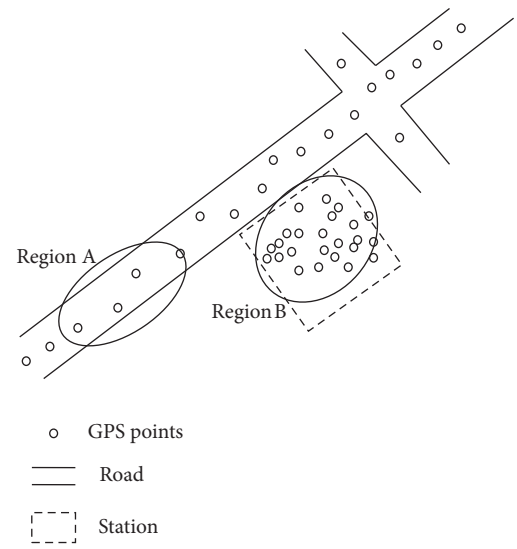


Figure 3: Spatial distribution characteristics of trajectory data under different conditions.

that the trajectory points at parking stations has different characteristics from the points in other places: the number of trajectory points per unit area is relatively large, the number of vehicles is large, the trajectory points tend to form obvious point clusters, and the speed value of most points is zero. In view of these features, this paper uses the DBSCAN clustering algorithm (density-based spatial clustering of applications with noise) proposed by Ester et al. to extract highly dense point clusters in trajectory data of coaches and calculates the coordinates of cluster center as the location of coach station [35].

2.3.1. DBSCAN Algorithm. The DBSCAN algorithm is a classic density-based clustering method and defines a point cluster as the largest set of points connected by density (Han et al.) [36]. The method identifies a region with sufficiently high density as a cluster and can find clusters of arbitrary shapes in a spatial database with noises. Assuming a dataset is $D = (X_1, X_2, \ldots, X_i, \ldots, X_n)$, where X_i is a trajectory point, the algorithm of DBSCAN is conducted as follows.

Step 1. Search for the unprocessed point X_i in the original dataset D. If X_i is neither classified into a cluster nor marked as a noise, check the points in its neighborhood with the radius of *Eps* represented by $N_{Eps}(X_i)$; if the number of points in $N_{Eps}(X_i)$ is not less than *MinPts*, a new point cluster C is created, all the points in $N_{Eps}(X_i)$ are added to the candidate point cluster C, and X_i is marked as processed.

Step 2. Traverse all unprocessed points q in the candidate cluster C one by one. Check the neighborhood $N_{Eps}(q)$, and if $N_{Eps}(q)$ contains at least *MinPts* points, then all points in $N_{Eps}(q)$ are added into the cluster$C$.

Step 3. Repeat Step 2 to continue searching for unprocessed points in C until all points in the candidate cluster C have already been processed.

Step 4. Repeat Step 1 to Step 3 until all points in D are either classified into a cluster or marked as a noise.

2.3.2. Stop Point Extraction. Due to driving safety or traffic rules, the coach would stop at some specific locations on the operation route, such as gas station, service area, station, and traffic lights. In this paper, the above locations are referred to as stop points. The trajectory data P_r of each route is processed using the DBSCAN algorithm to extract all highly dense point cluster C_i, and the coordinates of each cluster center $C_i^c(X_i^c, Y_i^c)$ are calculated by $X_i^c = (1/\|C_i\|) \sum_{P_r \in C_i}^{\|C_i\|} X_{P_r}$, $Y_i^c = (1/\|C_i\|) \sum_{P_r \in C_i}^{\|C_i\|} Y_{P_r}$, where (X_{P_r}, Y_{P_r}) is the coordinates of trajectory point P_r belonging to the cluster C_i, and $\|C_i\|$ is the number of trajectory points within C_i. These cluster centers C_i^c are the stop points for the i-th coach route, and each route has its corresponding stop points.

2.3.3. Coach Station Extraction. Coach station extraction is to identify station locations from the stop points generated by various reasons. Different from the stop points generated by gas filling, temporary stop, etc., the stop points within coach stations have the following characteristics: the density is higher, the same coach has multiple stop points in the same place, and the number of involved coaches is relatively larger. Therefore, this paper adopts the DBSCAN algorithm again on the stop points in the dataset $C_i^c (i = 1, 2, \cdots, k)$, where k is the number of coach trajectories and identifies highly dense clusters O_j. Similarly, the coordinates of cluster center $O_j^c(X_j^c, Y_j^c)$ are calculated by $X_j^c = (1/\|O_j\|) \sum_{C_i \in O_j}^{\|O_j\|} X_{c_i^c}$ and $Y_j^c = (1/\|O_j\|) \sum_{C_i \in O_j}^{\|O_j\|} Y_{C_i^c}$, where $(X_{C_i^c}, Y_{C_i^c})$ is the coordinates of stop point C_i^c belonging to the cluster O_j and $\|O_j\|$ is the number of stop points of the cluster O_j. The cluster centers O_j^c are obtained as the locations of coach stations.

2.4. Schedule Information Extraction. The schedule information includes arrival time, dwelling time, and departure time of a coach to each station on its route. Accurate schedule information is very important for both passengers and drivers, which can save passengers' waiting time and ensure normal operation of coaches. In the previous subsection, the trajectory points collected while the coach moves into a station each time are obtained, and they form a point set $U_i = \{P_i \mid \forall_{j,k}, P_i \in C_j \& C_j^c \in O_k\}$. When all the trajectory points constituting U_i are sorted by time, the location time series of U_i is T_i expressed as follows:

$$T_i = \left(t_1, t_2, \ldots, t_{\|U_i\|-1}, t_{\|U_i\|}\right) \quad (1)$$

where t_1 represents the earliest location time in the point set U_i and $t_{\|U_i\|}$ represents the latest location time. Therefore, t_1 and $t_{\|U_i\|}$ are the arrival and departure time of the station for one shift. The arrival time of a coach to a station is influenced by many factors including weather and traffic condition, and it is not always the same for different shifts. According to the error theory, the arrival timestamps to the same station can be taken as a random variable according to the normal distribution. As such, the arrival time t is represented as $\sim N(\mu, \sigma^2)$, where μ is the expected value of the arrival time or in other words the planned arrival time, and σ is the standard deviation of the arrival time series caused by various kinds of reasons.

Coaches may experience a serious traffic jam or accident on the road, which causes that the arrival times to stations are much different from the planned times, and we call them abnormal arrival time. The Grubbs' test is used to eliminate abnormal arrival timestamps, and the detailed procedures are as follows:

(1) The arrival time series of a station is arranged from small to large (usually the smallest or largest value are first doubted whether they are abnormal or not).

(2) Determine the danger ratio α.

(3) Calculate T (T is the distribution of the order statistics $X(r)$). Suppose $X_{(1)}$ is suspicious, let $T = (\overline{X} - X_{(1)})/\sigma$, and suppose $X_{(n)}$ is suspicious; let $T = (X_{(n)} - \overline{X})/\sigma$, where $\overline{X} = \sum_{t=1}^{n} X_t/n$ and $\sigma = \sqrt{\sum_{t=1}^{n}(X_t - \overline{X})^2/(n-1)}$.

(4) Check the value of $T(n, \alpha)$ corresponding to n and α.

(5) If $T \geq T(n, \alpha)$, the suspicious data is abnormal and should be eliminated and the above procedures are repeated until there are no abnormal data in the time series.

After all abnormal data are eliminated, the calculated average arrival time is taken as the planned arrival time represented by t_a, and the same method is used to calculate the planned departure time represented by t_b.

2.5. Station Name Extraction. After obtaining the geographical location of coach stations and its associated schedule information, we also need to identify the name of stations to compose complete information. In this paper, the Uniform Resource Locator (URL) search function of AutoNavi Map is used to identify station name with the geographic coordinates of stations obtained as an input parameter and the search type set to traffic facility. If the result returned by the function contains the words related to passenger station, it is taken as a station name; otherwise the place name (or the name of administrative unit) near the input coordinates is identified as the station name. For example, to identify the station name at the place with the coordinates of 118.1142E and 39.63N, the point type is set to traffic facility, and the following URL is constructed:

http://restapi.amap.com/v3/geocode/regeo?output=xml&location=118.1142,39.63&key=84a3ffa9cbd674eb2a8170bd1d9-24528&radius=1000&poitype=150400&roadlevel=1&extensions=all

The returned result is shown in Algorithm 1 , and then Tangshan Passenger Transport West Station is taken as the station name in the place of 118.1142E and 39.63N.

2.6. Coach Route Extraction. Another important coach operation information is the planned route of each coach. Our task is to extract the planned route of a coach from multishift

```
<pois type="list">
<poi>
    <id>B013B01ANW
    <name> Tangshan Passenger Transport West Station </name>
    <type> Transportation facilities; long-distance bus station; long-distance bus station </type>
    <tel>0315-2333339</tel>
    <direction>East</direction>
    <distance>525.363</distance>
    <location>118.120142,39.631174</location>
    <address> Station road </adress>
    <poiweight>0.590102</poiweight>
    <businessarea/>
</poi>
</pois>
```

Algorithm 1: Station name search result.

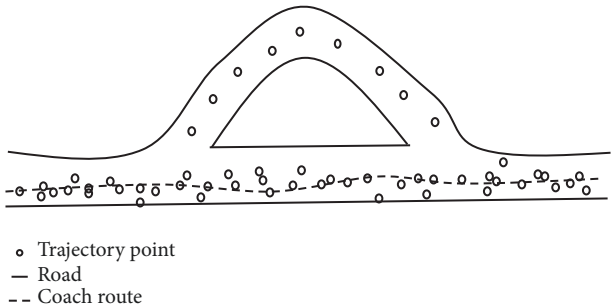


Figure 4: Coach route extraction.

trajectories with possible variances as shown in Figure 4. This paper uses the incremental trajectory merging algorithm composed of four major steps: candidate route generation, trajectory point classification, trajectory point merging, confidence filtering, and smoothing [34]. Firstly, the moving segment of any one trajectory is taken as the candidate route, and then the tracking points in the rest of trajectories are classified to different types according to their spatial and semantic relationship with the candidate route and further merged into the candidate route, and finally the planned route is obtained after confidence filtering and smoothing.

(1) Candidate Route Generation. The candidate route is a temporary result of route extraction process, which stores the moving path of coaches in the form of vector. For the sake of clarity, the vector node point is called the route node point. Each route node point has a weight attribute showing the reliability of this node point on the planned route of the coach. When a node point is newly added, its weight attribute is given an initial value of 1.

(2) Trajectory Point Classification. The trajectory points are classified into three categories based on the relationship between the trajectory points and the candidate route segments. If the distance between a trajectory point and the candidate route is less than or equal to d_1 and the direction difference between them is less than or equal to θ_1, this trajectory point is classified to R_1 type. If the distance between a trajectory point and the candidate route is between d_1 and d_2, and both two adjacent points of the trajectory point meet R_1 type condition with the same candidate route, this point is classified to R_2 type. If a point belongs to neither R_1 type nor R_2 type, it is classified as R_0 type.

(3) Trajectory Point Merging. Different types of trajectory points are merged to the candidate route using different methods. When the trajectory point is R_1 type, the edge of candidate route *match_edge* that is closest to the trajectory point is first found. If the distance of the point from any one node point of *match_edge* is not more than d_3, the weighted mean value of coordinates of the point and the node point is calculated and taken as the new position, and the weight value of the node point is increased by 1. If the point is far from both node points of *match_edge* but the distance to the edge *match_edge* is less than or equal to d_4, then the weight values of both node points of *match_edge* are increased by 1. When the trajectory point is R_2 type, it is inserted to *match_edge*. When the trajectory point is R_0 type, a new candidate route is created starting from this trajectory point.

(4) Confidence Filtering and Smoothing. Confidence filtering is to remove the extracted route with a low accuracy or probability of being a real planned route. It is achieved by removing the route node points, edges, or an entire route with the weight values less than λ_1. Route smoothing is to smooth slight sawteeth existing in the extracted route without changing the major shape of the route. This paper adopts the adaptive smoothing algorithm based on bending angle, which determines the smoothing degree according to the sawtooth angle [34]. As shown in Figure 5, Points A, B, and C form a sawtooth, and D is the foot point of the sawtooth vertex B projected onto the straight line AC. The sawtooth is polished by moving its Vertex Point B towards the Foot Point D, and the moving distance is calculated by $d_M = \eta \cdot d$, where d is the distance between Point B and Point D, $\eta = (\alpha - \theta_2)/(180 - \theta_2)$, and α is the vertex angle. The maximum moving distance is set to 5 m in order to ensure that the maximum amending distance does not exceed the location error.

Table 1: Sample GPS data of passenger vehicles.

License plate number	administrative unit code	longitude	latitude	speed(km/h)	Direction (degree)	Positioning time
PAE1234	110000	116.280670	37.332818	28	158	2017-08-01 12:30:06
FAN1053	110000	117.191765	41.227195	70	79	2017-08-09 00:55:08
PAW5020	120000	117.141736	39.114240	65	223	2017-08-09 06:43:30

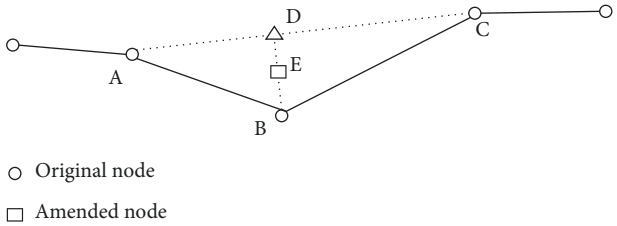


Figure 5: Schematic diagram of smoothing method (from Li et al.)[34].

3. Result and Analysis

3.1. Study Area and Dataset. The Beijing-Tianjin-Hebei region in China is selected as our study area to demonstrate the performance of the proposed method as shown in Figure 6. There are two kinds of study datasets used in this paper. One is the GPS trajectory data of about 300 coach routes collected by the NCVMP from August 1st to 20th, 2017, and the sample data is shown in Table 1. The main fields of the trajectory data include license plate number (encrypted for privacy protection), administrative unit code, positioning time, longitude, latitude, speed, and direction. The spatial distribution of GPS trajectory data is shown in Figure 6. Another dataset is the basic coach route information, including route number, coordinates of origin and destination station, and departure time, and the sample dataset is shown in Table 2.

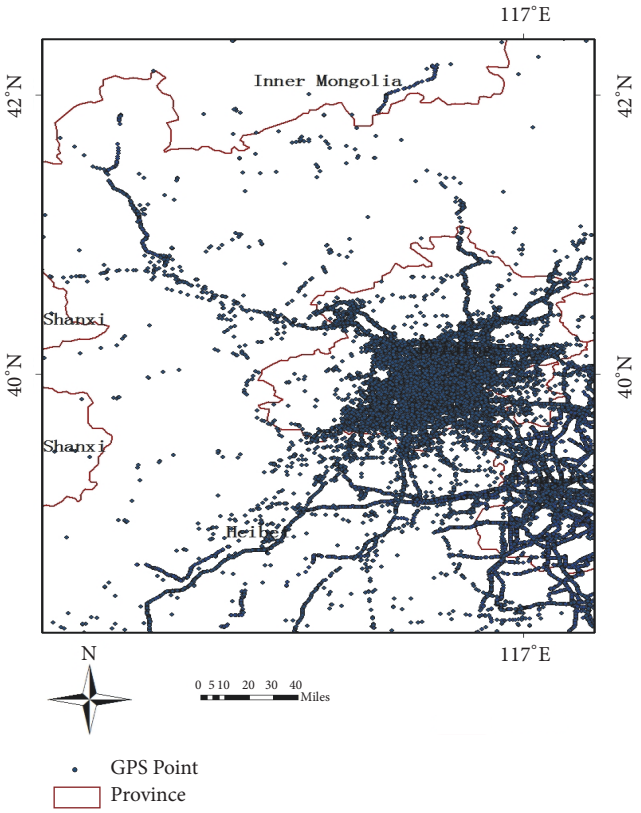


Figure 6: Study area and dataset.

3.2. Results and Accuracy Analysis. According to Section 2, the coordinates of the coach trajectory points are first converted from the WGS-84 to the UTM projection, and the abnormal points are eliminated. Considering that there are inevitably location errors in both trajectory data and basic coach operation information, and the coach station has a certain spatial size, the buffer radius d_s in the trajectory splitting introduced in Section 2.2 is set to 0.2km. Through investigation and analysis, the actual departure times of multiple shifts for the same coach route are very close to the planned departure time, so the time difference threshold t_s is set to 0.5h. The GPS sampling interval of the trajectory data is between 30 seconds and 5 minutes, so the clustering parameters *Eps* and *MinPts* are set to 20m and 5, respectively, to extract stop points. In view of the operation characteristics of the coaches within stations and the regional features of stations, the above two parameters are set to 80m and 5, respectively, in the second DBSCAN clustering. It means a place can be taken as a coach station only if at least 5 coach shifts or coaches stop in this place. In the extraction process of coach route, the parameter value selection is shown in Table 3 according to the similarity of multiple trajectories of the same coach route.

In order to demonstrate the process of extracting coach operation information more clearly, one coach route (Route 212) in the study dataset is taken as an example, which leaves from the Liuliqiao Bus Station in Beijing at 19:00. Figure 7(a) shows the trajectory data of all shifts for Route 212. As we can see from the main figure and enlarged view, the raw dataset contains the trajectory segment before 19:00 (in other words, the coach route operation has not started yet) and also the return trip part belonging to a different coach route. Figure 7(b) shows the result after trajectory splitting, and the remaining data are the trajectories operating for Route 212, thereby reducing the number of trajectory points to be processed. For the coach station extraction, the Qibin Road Station is taken as an example to demonstrate clustering process as shown in Figure 8. Figure 8(a) shows the trajectory points around the Qibin Road Station, while Figure 8(b) shows the five stop points extracted by the DBSCAN algorithm according to point density, and Figure 8(c) shows that

Table 2: Sample of basic coach route information.

Route No.	Origin station		Destination station		Departure time
	Longitude	Latitude	Longitude	Latitude	
2204	116.472997	39.90157	118.173720	39.62909	07:00
152	116.30451	39.88339	125.704918	42.536931	16:20

Table 3: Parameter values in trajectory merging.

Parameter	Value	Parameter	Value
d_1	50m	θ_1	30°
d_2	500m	θ_2	140°
d_3	20m	λ_1	3
d_4	5 m		

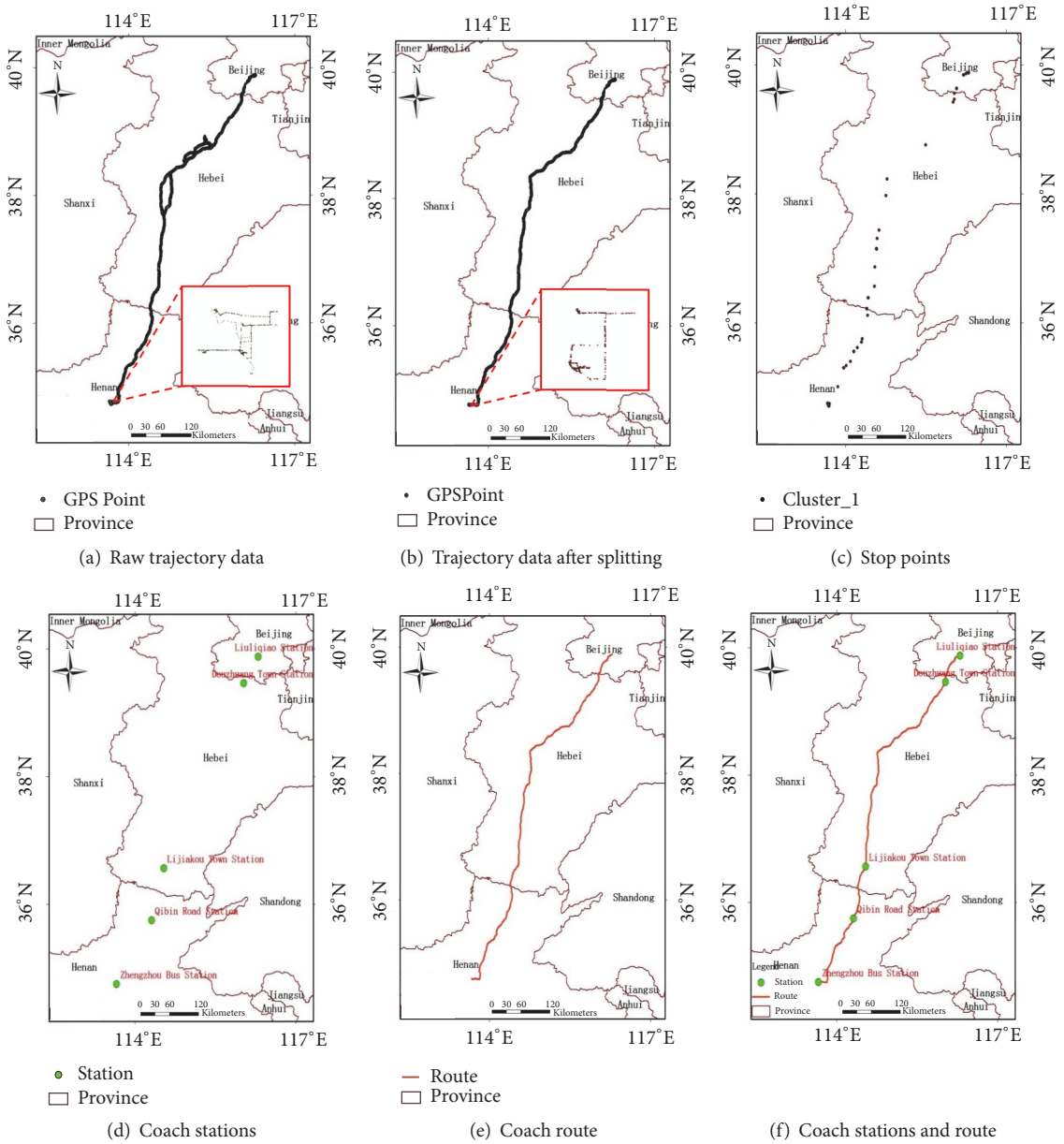


Figure 7: Extraction results at different steps for a coach route (Route 212).

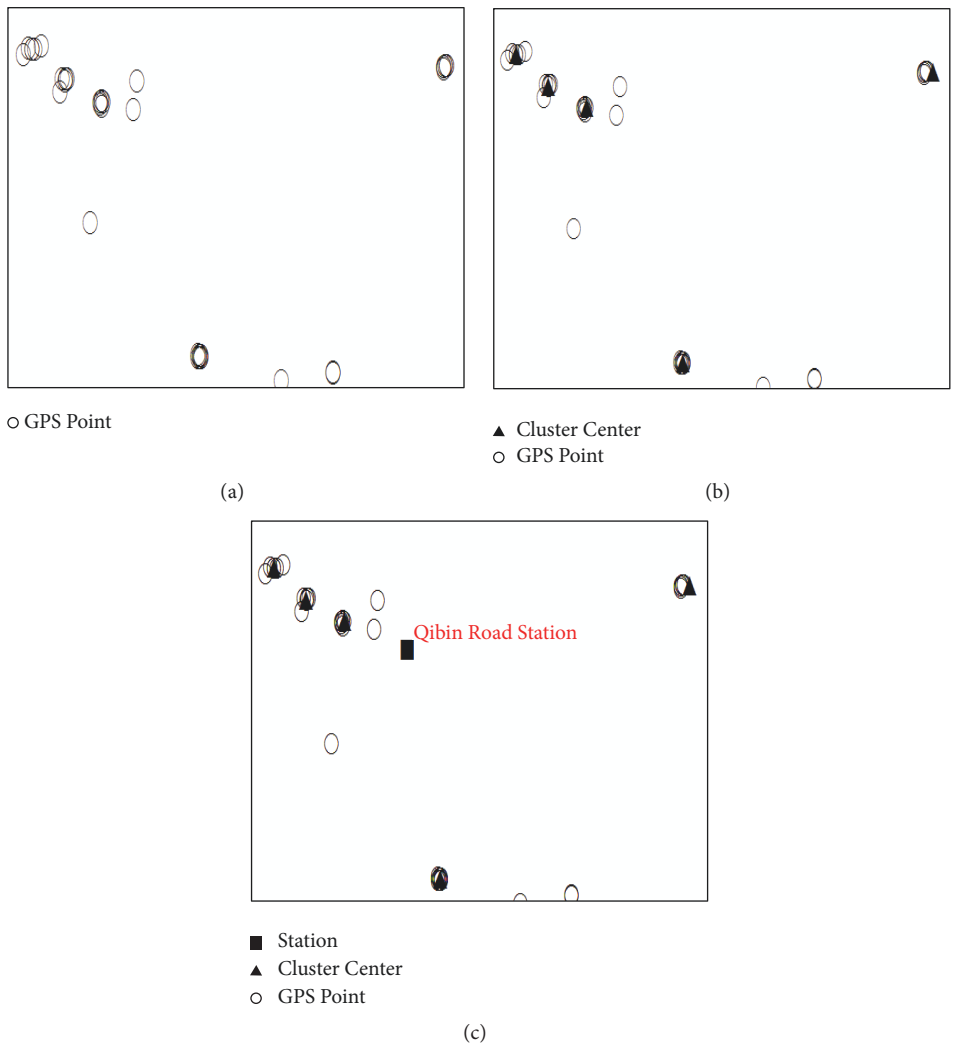


Figure 8: Coach station extraction process.

a coach station is identified from the stop points after another clustering. For Route 212, quite a few stop points are extracted (Figure 7(c)) and only 5 coach stations are further extracted (Figure 7(d)) since some stop points are generated due to traffic jam, gas filling, etc. Figures 7(e) and 7(f) show the final extracted coach route which starts from the Liuliqiao Coach Station in Beijing and ends at the Zhengzhou Coach Station. Table 4 shows the operation schedule of this coach route showing the arrival and departure time for each station.

Figure 9 shows the final coach station and route extraction result of the entire study dataset, and Table 5 demonstrates the schedule information for part of schedule routes, including schedule code, all passing station name, station location, and arrival and departure time.

A total of 1,284 coach stations were extracted from the study data. In order to verify the location extraction accuracy of coach stations, we superimposed the extraction stations on satellite imageries of Google Earth and randomly selected 300 coach stations to verify whether the extracted stations match with satellite imageries or not. By comparison, 253 stations were correctly extracted, while 47 were wrong extraction (e.g., highway service areas, road intersections). The extraction accuracy of coach station is 84%. Figure 10 shows the superimposed effect of stations on remote sensing images for four typical areas, in which the yellow pin point represents raw trajectory point of coaches, while the green pin point represents the extracted coach station. Figures 10(a), 10(b), 10(c), and 10(d) are a large coach station, a roadside parking lot, a road intersection in a village, and a place on highway, respectively. The extracted station in (a) is apparently correct. Through investigation, although the extracted locations in (b) and (c) are informal coach stations, they are two locations where nearby passengers get on and get off coaches every day, and they actually function as two coach stations. The extracted location in (d) is in the middle of a highway. After investigation, we find that at the extracted location is a place where traffic jam occurs very frequently; therefore, the GPS trajectory points collected here are very dense, which causes it to be wrongly extracted since it has similar characteristics to coach stations.

Table 4: Extracted schedule for Route 212.

Time	Station				
	Liuliqiao Coach Station	Douzhuang County	Lijiakou County	Qibin Road	Zhengzhou Coach Station
Arrival time	--	20:23	02:25	06:34	09:13
Departure time	19:00	20:53	05:00	06:36	--

Table 5: Schedules for part of coach routes.

Schedule No.	Station	Longitude	Latitude	Stay time(minutes)	Arrival time	Departure time
2007	Baoding	117.3089	39.71596		--	07:31
	Xiacang	117.3936	39.76028	3	07:51	07:54
	Yutian	117.7402	39.8814	2	08:49	08:51
	Shaliuhe	117.9575	39.87243	6	09:35	09:41
	Tangshan	118.1142	39.630		10:38	--
1265	Liuliqiao	116.2995	39.8778		--	15:44
	Zhuozhou	115.9581	39.4709	2	17:31	17:33
	Gaobeidian	115.8432	39.3126	6	18:00	18:06
	Pinggang	115.6019	39.2663		18:51	--
15963	Liuliqiao	116.2997	39.8781		--	15:29
	Daban	118.6777	43.5117	2	00:07	00:09
	Lindong	119.3796	43.9715		01:14	--

3.3. Efficiency Analysis. In order to test the efficiency of the proposed method in extracting coach operation information, five subdatasets of different sizes including 10, 50, 100, 300, and 500 coach routes were selected from the study data, and the execution times were recorded. The number of extracted stations was 47, 153, 288, 1117, and 1778, respectively, for the five datasets, and the corresponding execution times were 35, 189, 442, 1630, and 2557 seconds as shown in Figure 11. The result indicates that the average execution time to extract coach operation information for each route is less than 5 seconds, and each coach route has an average of four stations. As can be seen from the change trend of the black solid curve, the execution time of the method is nearly linearly related to the number of processed routes. Therefore, the proposed method can be applied in real-world applications since the execution time would not significantly increase as the number of coach routes increases. In addition, the algorithm of this paper is programmed in a batch-processing mode in which a large number of datasets can be processed at the same time.

4. Discussion

In this paper, the massive historical coach trajectory data is applied in extracting coach operation information. The experimental results prove that the proposed method can automatically extract accurate coach operation information in a large area. Note that the extracted coach operation information is much more abundant than the input data of the proposed method: basic coach operation information. The transport agency manages basic coach operation information since enterprises need to get approval from them to operate a coach route; however, they cannot master detailed operation information because the operating enterprises can choose driving routes and set intermediate stations on their own. The extraction result is composed of complete operation information of coaches: station location, station name, coach schedule, and driving route, which provides data basis for intelligent travel information service. The method of this paper has the following characteristics.

(1) The raw trajectory data is split into effective and noneffective segments according to the location of origin and destination station of each coach route. The effective trajectory segment after splitting represents the data collected while the vehicle operates for a specific route, and only this part would be processed in further steps. The splitting process not only reduces the amount of data to be processed, but also prepares for subsequent schedule extraction.

(2) There are various kinds of clustering algorithms that can be taken into account for extracting coach stations. The K-means algorithm [37] and the spectral clustering algorithm [38] have to know the correct number of clusters in advance and cannot identify noise points, so they are not suitable for our research. By contrast, the DBSCAN [35] and HDBSCAN [39, 40] algorithm do not need to specify the number of clusters and are able to detect noise points. We select two sample trajectories to compare the results of DBSCAN and HDBSCAN algorithms using the same parameter values. The parameter $MinPts$ for both DBSCAN and HDBSCAN are set to 5, and the parameter Eps for DBSCAN is set to 20 meters. Figures 12 and 13 show clustering results of two algorithms. It can be seen that most cluster points generated by HDBSCAN appear in the shape of chains (Figures 12(a) and 13(a)), while those generated by DBSCAN are mostly highly dense points

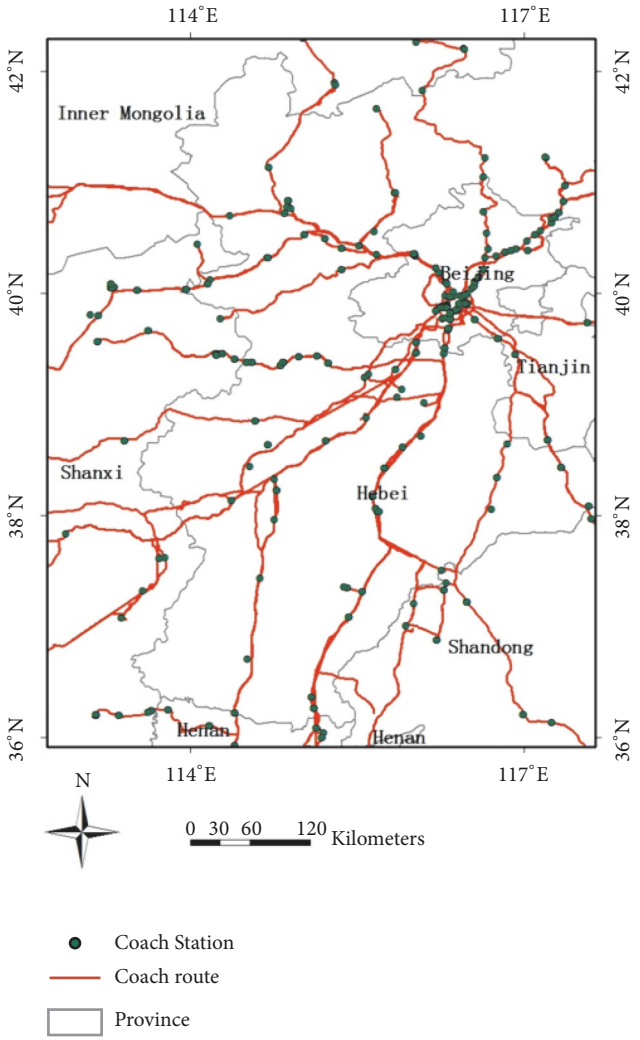


Figure 9: Coach station and route extraction results for the entire dataset.

(Figures 12(b) and 13(b)). Based on algorithm principle and clustering results, we can know that HDBSCAN recognizes a cluster based not on regional point density but on mutual reachability distance, so it would identify point chains as clusters, and this feature would lead to wrong extraction of coach stations. Therefore, we choose DBSCAN that can identify highly dense points.

(3) Two times of clustering is used on the trajectory data to extract coach stations. The clustering for the first time is to identify the highly dense points and calculate the center of point cluster as a stop point of the coach. The clustering based on the obtained stop points for the second time is to further identify the location of coach stations among all stop points caused by various kinds of reasons. Theoretically, one time clustering is adequate for coach station extraction. However, the time complexity of the DBSCAN clustering algorithm is $O(n \times n)$ which is quadratic to the number of processed points. Usually, the number of raw trajectory points is thousands of millions, so the processing time in clustering would be very long if the clustering algorithm is directly conducted on them. Using a two-time clustering strategy is very useful in efficiency promotion.

(4) A few parameters are involved in the proposed method, including distance threshold, time and angle difference threshold, and clustering parameters. Most of these parameters do not change as the trajectory dataset or study area change. For example, the distance threshold, the angle, and time difference threshold used in trajectory merging are determined by considering both vehicle moving characteristics and the geometric features of roads and can be kept unchanged. However, the setting of clustering parameters *Eps* and *MinPts* required in coach station extraction needs to be determined according to the sampling interval of raw trajectory data. If the sampling interval of the trajectory data exceeds 30s, the value of *MinPts* is recommended to be not greater than 5 when extracting stop points. If the sampling interval is less than 30s, its value is recommended to be between 5 and 10.

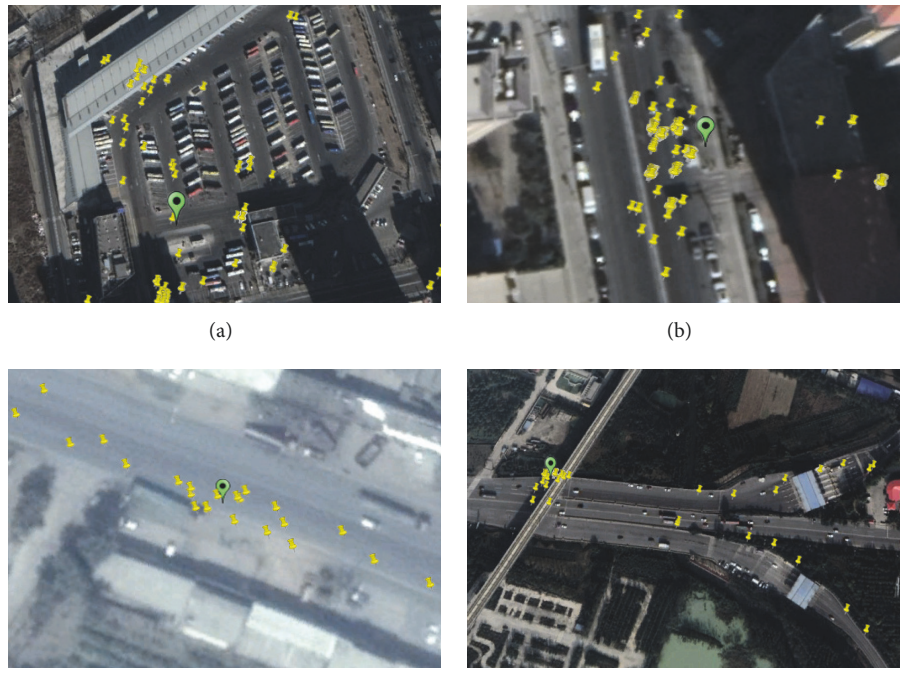
(5) Although the proposed method makes it very convenient to obtain coach operation information in a large area quickly, there are still some problems to be improved. For example, if there is a frequently congested point on the coach route, the trajectory data of vehicles at this place will satisfy the trajectory characteristics within stations, and the congestion point will be identified as coach stations. In addition, for the long-distance trip, the coach driver will often park for a rest in some specific places such as highway service area. These areas actually have similar functions as coach stations, that is, both of them are the place where passenger get on and get off. Therefore, it is very hard to distinguish these places from coach stations based on only trajectory data. In future, the stop time distribution feature or other geographic information will be taken into account to assist in improving the extraction accuracy.

(6) The extraction accuracy of coach stations is influenced by the time span of coach trajectory data. The larger the data volume is, the higher the extraction accuracy is. According to the experimental results, at least 10 days of trajectory data should be used in order to automatically extract coach operation information.

(7) As introduced in Section 2.5, our method is designed to extract not only formal coach stations but also "informal coach stations" where there are no station facilities but function as stations in fact. It means that the word "coach station" throughout the entire paper could mean both formal and informal stations. No distinction is made on this word in order to make the paper simple to follow.

5. Conclusion

Currently, the coach operation information is mainly collected by manual investigation which is time-consuming and labor-intensive. Road transport authorities and enterprises have no efficient way to obtain detailed coach operation information across the country, and this significantly hinders the realization of intelligent traffic information service. In



(a) (b) (c) (d)

Figure 10: Superimposed effect of extracted coach stations on satellite images.

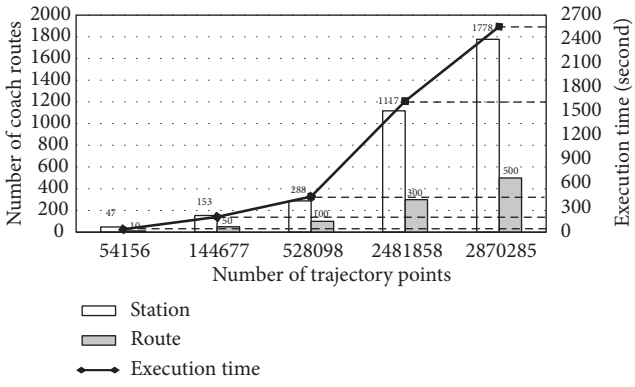


Figure 11: Efficiency analysis of algorithm.

order to extract coach operation information automatically and efficiently, the paper first analyzes the trajectory characteristics of coaches within stations and identifies the highly dense point cluster as coach stations using the DBSCAN clustering algorithm. Then the arrival and departure time for each station is calculated using the error adjustment method. In addition, the name of coach station is obtained through the API of online maps based on the location of extracted stations. Finally, the coach route is extracted by an incremental trajectory merging method. The proposed method is applied in the Beijing-Tianjin-Hebei region to extract coach operation information. Experimental results show that the extraction accuracy is 84% and verify its effectiveness and feasibility. The proposed method makes use of data mining techniques to extract useful information from big trajectory data and saves a lot of labor work, time, and economic cost needed by the on-site investigation. It can provide a data source for establishing a nationwide online ticketing and travel information system for various kinds of road passenger transport.

Data Availability

The historical trajectory data used to support the findings of this study may be released upon application to the China Transport Telecommunications & Information Center, which can be contacted at xiechao@transinfo.com.cn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is financially supported by National Natural Science Foundation of China (no. 41601485), National Key Research and Development Program of China (no. 2017YFB0503801), and the Open Project of National Engineering Laboratory for Transportation Safety and Emergency Informatics (no. YW170301-03). Thanks are due to China Transport Telecommunication & Information Center for providing experiment data and related support.

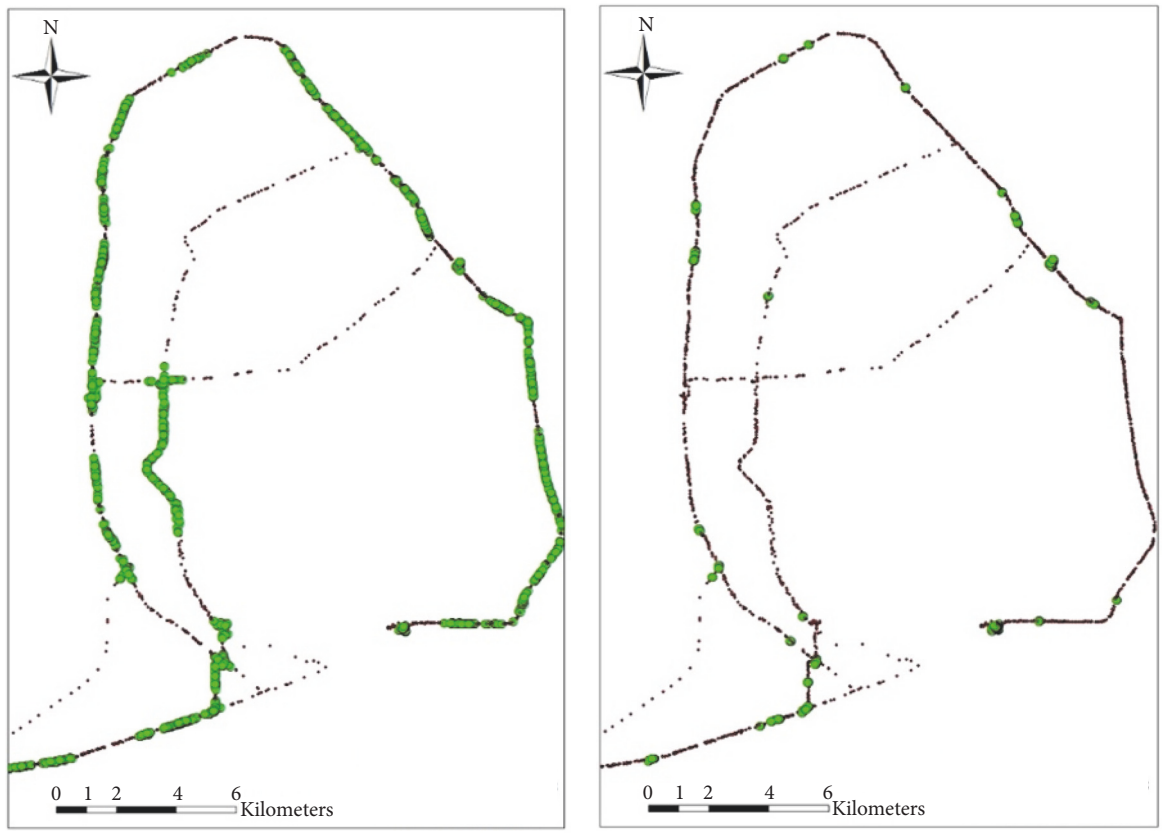


Figure 12: Clustering results of Sample Trajectory A by two algorithms: (a) HDBSCAN and (b) DBSCAN.

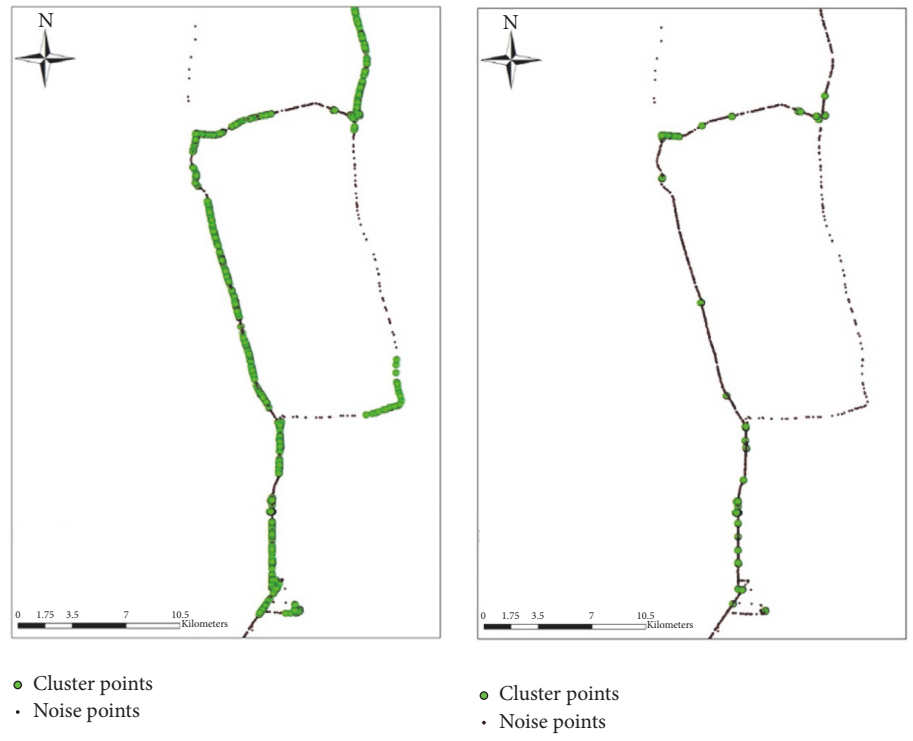


Figure 13: Clustering results of Sample Trajectory B by two algorithms: (a) HDBSCAN and (b) DBSCAN.

References

[1] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transportation Research Part C: Emerging Technologies*, vol. 6, no. 4, pp. 271–288, 1998.

[2] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: a case study in Rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

[3] J. C. Sutton, "GIS applications in transit planning and operations: A review of current practice, effective applications and challenges in the USA," *Transportation Planning and Technology*, vol. 28, no. 4, pp. 237–250, 2005.

[4] K.-H. Chen, C.-R. Dow, and S.-J. Guan, "NimbleTransit: Public transportation transit planning using semantic service composition schemes," in *Proceedings of the 2008 11th International IEEE Conference on Intelligent Transportation Systems*, pp. 723–728, Beijing, China, 2008.

[5] A. Simroth and H. Zähle, "Travel time prediction using floating car data applied to logistics planning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 243–253, 2011.

[6] A. Chepuri, J. Ramakrishnan, S. Arkatkar, G. Joshi, and S. S. Pulugurtha, "Examining travel time reliability-based performance indicators for bus routes using GPS-based bus trajectory data in India," *Journal of Transportation Engineering Part A: Systems*, vol. 144, no. 5, pp. 1–14, 2018.

[7] A. Shingare, A. Pendole, N. Chaudhari, P. Deshpande, and S. Sonavane, "GPS supported city bus tracking & smart ticketing system," in *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)(ICGCIOT)*, pp. 93–98, Greater Noida, Delhi, India, 2015.

[8] A. Carter, "Intelligent transport systems," *Journal of Navigation*, vol. 54, no. 1, pp. 57–64, 2001.

[9] A. Sładkowski and W. Pamuła, Eds., *Intelligent Transportation Systems-Problems and Perspectives*, vol. 303, Springer International Publishing, 2016.

[10] D. Levinson, "The value of advanced traveler information systems for route choice," *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 1, pp. 75–87, 2003.

[11] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[12] T. W. Cao, "Research on problems and status of road passenger stations in china," *Road Transport*, vol. 6, no. 263, pp. 16–18, 2009.

[13] M. Rahmani and H. N. Koutsopoulos, "Path inference from sparse floating car data for urban networks," *Transportation Research Part C: Emerging Technologies*, vol. 30, pp. 41–54, 2013.

[14] B. Y. Chen, H. Yuan, Q. Li, W. H. K. Lam, S.-L. Shaw, and K. Yan, "Map-matching algorithm for large-scale low-frequency floating car data," *International Journal of Geographical Information Science*, vol. 28, no. 1, pp. 22–38, 2014.

[15] G. Draijer, N. Kalfs, and J. Perdok, "Global positioning system as data collection method for travel research," *Transportation Research Record*, vol. 1719, pp. 147–153, 2000.

[16] Y. Liu, X. Liu, S. Gao et al., "Social sensing: a new approach to understanding our socioeconomic environments," *Annals of the Association of American Geographers*, vol. 105, no. 3, pp. 512–530, 2015.

[17] A. T. Palma, V. Bogorny, Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *Proceedings of the 23rd Annual ACM Symposium on Applied Computing*, pp. 863–868, Fortaleza Ceara, Brazil, 2008.

[18] T. Bhattacharya, L. Kulik, and J. Bailey, "Extracting significant places from mobile user GPS trajectories: a bearing change based approach," in *Proceedings of the ACMSIGSPATIALGIS'12*, ACM, Redondo Beach, Calif, USA, 2012.

[19] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated GPS traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, article 2, 2011.

[20] J. Li, Q. M. Qin, L. You, and C. Xie, "Parking lot extraction method based on floating car data," *Journal of Wuhan University (Information Science Edition)*, vol. 38, no. 5, pp. 599–603, 2013.

[21] S. Schroedl, K. Wagstaff, S. Rogers, P. Langley, and C. Wilson, "Mining GPS traces for map refinement," *Data Mining and Knowledge Discovery*, vol. 9, no. 1, pp. 59–87, 2004.

[22] J. Li, Q. Qin, J. Han, L.-A. Tang, and K. H. Lei, "Mining trajectory data and geotagged data in social media for road map inference," *Transactions in GIS*, vol. 19, no. 1, pp. 1–18, 2015.

[23] L. L. Cao and J. Krumm, "From gps traces to a routable road map," in *Proceedings of the Workshop on Advances in Geographic Information Systems*, pp. 3–12, New York, NY, USA, 2009.

[24] L. Tang, X. Yang, Z. Kan, and Q. Li, "Lane-level road information mining from vehicle GPS trajectories based on Naïve Bayesian classification," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2660–2680, 2015.

[25] C. Kuntzsch, M. Sester, and C. Brenner, "Generative models for road network reconstruction," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 1012–1039, 2016.

[26] J. Y. Kang and H. S. Yong, "Mining spatio-temporal patterns in trajectory data," *Journal of Information Processing Systems*, vol. 6, no. 4, pp. 521–536, 2010.

[27] P.-R. Lei, T.-J. Shen, W.-C. Peng, and I.-J. Su, "Exploring spatial-temporal trajectory model for location prediction," in *Proceedings of the 12th IEEE International Conference on Mobile Data Management (MDM '11)*, pp. 58–67, Lulea, Sweden, June 2011.

[28] P. Zhao, X. Liu, W. Shi, T. Jia, W. Li, and M. Chen, "An empirical study on the intra-urban goods movement patterns using logistics big data," *International Journal of Geographical Information Science*, pp. 1–28, 2018.

[29] M. Lu, Z. C. Wang, and X. R. Yuan, "TrajRank: Exploring travel behaviour on a route by trajectory ranking," in *Proceedings of IEEE Pacific Visualization Symposium*, pp. 14–17, China, 2015.

[30] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with GPS history data," in *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, pp. 1029–1038, 2010.

[31] H. P. Hsieh, C. T. Li, and S. D. Lin, "Measuring and Recommending Time-Sensitive Routes from Location-Based Data," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, vol. 5, Article No. 45, 2015, Buenos Aires, Argentina.

[32] J. Li, J. Wang, J. Zhang, Q. Qin, T. Jindal, and J. Han, "A probabilistic approach to detect mixed periodic patterns from moving object data," *GeoInformatica*, vol. 20, no. 4, pp. 715–739, 2016.

[33] M. Gerla, E.-K. Lee, G. Pau, and U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds,"

in *Proceedings of the In Internet of Things (WF-IoT). IEEE World Forum*, pp. 241–246, IEEE, 2014.

[34] J. Li, Q. Qin, C. Xie, and Y. Zhao, "Integrated use of spatial and semantic relationships for extracting road networks from floating car data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, no. 1, pp. 238–247, 2012.

[35] M. Ester, H. P. Kriegel, J. Sander, and X. W. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the The Second International Conference on Knowledge Discovery and Data Ming*, Portland, ORE, USA, 1996.

[36] J. W. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Chapter 10, Morgan Kaufmann, San Francisco, Calif, USA, 3rd edition, 2011.

[37] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, no. 2, pp. 768-769, 1965.

[38] J. Demmel, "CS267: Notes for Lecture 23," April 9, 1999, Graph Partitioning, Part 2.

[39] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, vol. 7819, pp. 160–172, 2013.

[40] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *Journal of Open Source Software, The Open Journal*, vol. 2, no. 11, 2017.