International Cartographic Association
Association Cartographique Internationale

Data Cleansing Method for Sparse Trajectory Data: A Case Study of Shared Electric Bicycles in Tengzhou

Zhaoxin Dai [a, *], Weixiang Peng [b], Chengcheng Zhang [a, *]

[a] *Chinese Academy of Surveying and mapping, daizx@lreis.ac.cn; 94100594@qq.com*
[b] *China University of Geosciences (Wuhan Campus), weixiang_peng@163.com*

* Corresponding author

Abstract: Location based service (LBS) technologies provides a new perspective for the spatiotemporal dynamics analysis of urban systems. Previous studies have been performed by using data of mobile communications, public transport vehicles (taxis and buses), wireless hotspots and shared bicycles. However, the analysis based on shared electric bicycles (e-bike) has yet to be studied in the literature. Data cleansing and the extraction of origin-destination (O-D) are prerequisites for the study of urban systems spatiotemporal patterns. In this study, based on a dataset that contains a week of shared e-bike GPS data in Tengzhou City (Shandong Province), sparse characteristics of discontinuities and non-uniformities of trajectory GPS and a lack of riding status are captured. Based on the characteristics and combining with the actual road, we proposed a method for the extraction of O-D pairs for every trajectory segments from continuous and stateless trajectory GPS data. This method cleans the incomplete and invalid trajectory records, which is suitable for sparse trajectory data. Finally, a week-long shared e-bike GPS data in Tengzhou City is scrubbed, and by sampling method, the extraction accuracy of 91% is verified. In summary, we provide a preliminary cleansing rules for the sparse trajectory data of shared e-bikes at the first time, which is highly reliable, and is suitable for data mining from other forms of sparse GPS trajectory data.

Keywords: Shared e-bike; GPS trajectory data; recognition of O-D pairs; sparse data; Tengzhou

1. Introduction

The acquisition of a large number of individual spatio-temporal data is steadily becoming realized with the development and application of location-based services (LBS) like global positioning satellite (GPS) technology, social networks and wireless communications (Long et al., 2017). These large-scale individual datasets that contain spatiotemporal characteristics provide new ways for the scientific study of human mobility patterns, the spatial structures of urban residence and employment, and urban planning.

Data cleaning and O-D pair extraction are prerequisites for analysis of urban structures and human mobility patterns based on spatiotemporal LBS data. Previous cleaning methods were mainly intended for data that contained complete attribute or uniform GPS information, such as phone signaling data (Sørensen et al., 2018; Janković et al., 2016), taxi GPS data (Zhou et al., 2018; Cui et al., 2016), and smart card data of bus or subway stations (Zhong et al., 2014; Long and Thill, 2015; Gao et al., 2018), to extract origin-destination (O-D) information and analyse the human mobility patterns, as well as the spatial and distance distributions of urban employment and residence. Under the initiative of low-carbon transportation, cheap and convenient public bicycle rental systems rise, which have become one of the most popular modes of travel for urban residents. These systems have effectively resolved the "last mile problem" in urban transportation, especially in the first tier cities like Beijing, Shanghai and Guangzhou. The study on shared bicycles has been favoured by more city and transportation planning researchers. Shared bicycles and shared electric bicycles (e-bikes) rental system are two most common public bicycle rental system (Kou and Cai, 2019; Zhang and Mi, 2018). Compared to shared bicycles, shared e-bikes is relatively insensitive to long-distance travel, poor air quality or weather (Campbell et al., 2016). Shared e-bikes are highly adept in navigating large roads and small alleys, which makes them an excellent solution for short- and medium-distance trips, especially in second- and third-tier cities. For shared bicycle, a few studies have extracted O-D pair of each ride directly by using the GPS information and riding status. However, for shared e-bicycle, the corresponding research has not yet been reported in recent studies. Additionally, unlike taxi GPS data or shared bicycle data, shared e-bikes travel at relatively high speeds and have limited battery usage, the GPS trajectory points tend to exhibit discontinuities and nonuniformities, and it is also difficult to obtain their riding status information. When shared e-bike GPS trajectory data cleaning is performed using the current method, the resulted O-D pairs may tend to become trivial and incorrect, which subsequently results in the incorrect analysis of human mobility.

In this paper, we first propose a novel cleaning method for shared e-bike GPS trajectories, which feature nonuniform GPS information and a lack of ride-status attribute information. Furthermore, actual road networks are used to continuously correct and validate the results of the trajectory extraction algorithm. The results of this study are significant for reallocation of e-bikes and provide scientific data for human mobility pattern analysis, urban

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-22-2019

functional zones sectorization, and spatial distribution of occupation and residence.

2. Literature review

At present, there are numerous studies about the analysis of resident mobility characteristics using LBS data. This includes studies using taxi GPS data, smart phone signaling data, smart card data, and the GPS trajectory data of shared bicycles (Nassir et al., 2011; Gordon et al., 2013; Huang et al., 2019).

In terms of taxi GPS trajectory data, Liu et al. (2012) traced a week-long passenger pick-up and drop-off locations of taxis in Shanghai, and analyzed the relationship between the daily movements of urban residents and land-use. Tang et al. (2015) collected 1100 taxis GPS data of 2012 in Harbin city, which has a sampling rate of 30 seconds per point, and extracted the O-D pairs of these taxis using their occupancy status data and location data. Zheng et al. (2014) proposed a set of criteria for assessing the accuracy of GPS taxi data, based on the identification of inconsistencies between movement, speed, and trip length. Based on GPS travel data, Wolf performed the classification of travel trajectories by setting a time threshold for the identification of vehicle parking events, which minimized the misrecognition of traffic jams or other delay-related time segments as vehicle parking events, thus enabling the identification of origin (O) and destination (D) points.

By using mobile phone signaling data and smart card data, Alexander et al. (2015) proposed a method for inferring the average daily O-D pairs of each user from the mobile phone records from anonymized users. Zou et al. (2011) proposed a method for the extraction of O-D pairs from bus trips based on mobile phone positioning. Alsger et al. (2016) implemented, validated and improved currently available algorithms for the estimation of O-D pairs. Xu et al. (2017) constructed a network for the population fluxes that occur during the Spring Festival in China based on Tencent Location Big Data. They then analyzed the relationship between these population fluxes and the level of development of several cities in China. Kim et al. (2017) used the O-D pair data of smart cards to investigate the habitual route selection patterns of bus passengers. Long et al. (2015) constructed a travel model based on smart card records to analyse the places of employment and residence of Beijing and their commuting behaviors, thus providing new ideas about the commuting patterns of large cities. Munizaga et al. (2014) proposed a method for the validation of public transport O-D matrices that were estimated from smart card and GPS data.

For the trajectory data of shared bicycles, Jensen et al. (2010) analyzed the riding behaviors of residents by using the data of the shared bicycle system in Lyon. The extracted data records contain the location and time of the beginning and end of each trip, and the precise trip distances measured by a distance counter on each bicycle. Since 2015, shared bicycle systems have developed at an extremely rapid pace in China. Many researchers have used Python data analysis packages and ArcGIS to study the spatiotemporal features of urban riding patterns by using currently available operational data. Some examples include the study of Yang et al. (2018) on the impact of public bike-sharing systems on public transport systems, and the studies of Shaheen et al. (2011) and Tang et al. (2017) on the bicycle sharing schemes in Hangzhou and Shanghai, respectively.

At present, smart transportation systems have provided new sources of data for the study of urban systems and mobility modes, e.g., the data gathered by smart card and shared bicycle systems. However, studies based on the trajectory data of shared e-bikes have not yet to be reported. Since shared e-bikes are rental goods that are used by a wide range of users, the data of each shared e-bike consists of multiple trajectory segments, each made by a different user. As the trips recorded by the GPS data are simply the spatiotemporal trajectory point coordinates of the vehicle, it is not possible to directly infer the activities of the vehicle's riders from this information. In addition, the GPS trajectory data of shared e-bikes do not contain attributes like riding status. The current data cleansing methods for the extraction of O-D pairs are therefore inapplicable for these data. Therefore, it is necessary to formulate new rules for determining the parking and traveling states of these e-bikes, thus restoring the information of each trajectory segments. In this paper, we proposed a method for data cleansing and O-D pair extraction that is suitable for sparse trajectory data. The findings of this study will provide a scientific basis for future studies about urban structures and the spatiotemporal characteristics of urban resident mobility.

3. Data Characteristics

3.1 Data Sources

The data used in this study is the GPS trajectory points of shared e-bikes in Tengzhou city that were acquired between 19 May 2018 and 26 May 2018. The integrated module with GPS and communication is installed in each e-bike, and the GPS information is sent to the specified internet address every minute. Based on the data acquisition interface of HTTP protocol provided by shared e-bike operator, GPS trajectory points data can be acquired from the specified internet address by a high frequency timer system, which is developed by Java language.

Tengzhou city is located in the west-southern of Shandong province, eastern in China. Tengzhou city is China's most beautiful eco-tourism demonstration city, has been awarded 2018 'happiness hundred counties' and 'industrial hundred counties'. The shared e-bikes have been widely used in Tengzhou city, especially in downtown area. As shown in Figure 1, the derived shared e-bike data is a set of un-sorted and continuous GPS points. The information contained by each GPS point includes: the vehicle's ID (stationid), data acquisition time (timestamp) and location (indicated as latitude/longitude coordinates), predicted mileage (anticipated mileage) and margin (margin). The dataset in this study includes the data of 516 shared e-bikes and 98795 GPS trajectory data points.

Data cleansing pertains to the extraction of O-D pairs (which form the basis of trajectory data) from the unsorted GPS points, as well as their corresponding trip times, to form the travel trajectories of each user. Since the source

data does not contain any riding status-related information, the key to O-D extraction lies in the identification of endpoints that belong to two adjacent trips, from the continuous and stateless initial data. The trips made by each user may then be determined on this basis.

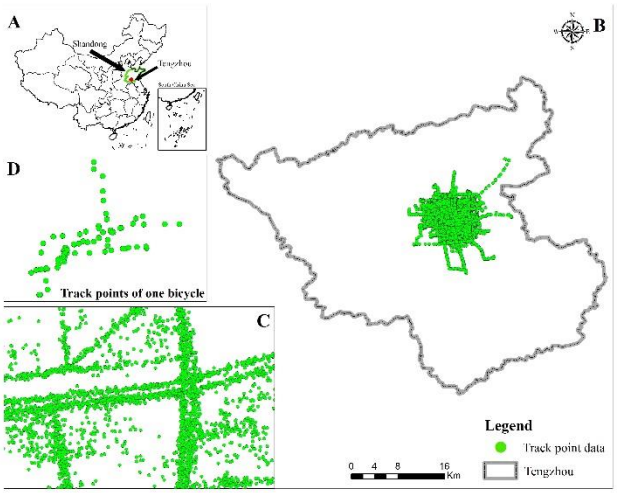

Figure 1. Trajectory GPS points. (Panel A shows the location of Shandong Province and Tengzhou city in China, Panel B is the spatial patterns of a week-long trajectory GPS data in Tengzhou city, Panel C is the zoom window of the stateless GPS data, and Panel D illustrates the trajectory GPS points for one e-bike.)

3.2 Data Characteristics

The characteristics of the shared e-bike dataset that used in this study were analyzed from a temporal perspective, based on the time attributes of the data. In this analysis, t_i ($1 \leqslant i \leqslant n$) is defined as the timestamp of n trajectory points, and $|t_{i+1} - t_i|$ is the sampling interval of the trajectory data.

(1) Data coverage

The number of days covered by the trajectory points of the shared e-bikes was analyzed. It was found that 34 shared e-bikes were used for five days, while 406 shared e-bikes were used between three to four days. Only 27 shared e-bikes were just used for one day. This indicates the shared e-bikes have high utilization rates, and the data can reflect on the travel patterns of partial urban residents. However, as not all rides cover the entirety of the week, the GPS trajectory data is still characterized by discontinuous and incomplete, which makes the data, sparse, to a certain degree.

(2) Sampling interval

The GPS tracking devices were configured with a sampling interval of one acquisition per minute. By calculating the time intervals of the acquired data, it was found that 61% of the data was acquired with a sampling interval of 1 minute while 89.6% of the data was acquired with a sampling interval within 2 minutes. Since the sampling intervals are relatively uniform, the data is usable and analyzable. However, there has still 10.4% of the data have sampling intervals longer than 2 minutes. It means that sparse trajectories with non-uniform distributions in the time dimension are present in the data, and this results in the loss of some fragmentary data.

The sparse vehicle trajectories of urban traffic are generally characterized by two significant features: (1) The trajectory points in the vehicle trajectory are not distributed uniformly in the time dimension. (2) The time spanned by the trajectories of each vehicle accounts for a very small proportion of the total observation time (Zhong et al., 2014; Xiao, 2014). Based on the above-mentioned definition, the shared e-bike trajectory data obtained in this study is determined to be sparse data. This may be attributed to two reasons: firstly, unlike the shared bikes, shared e-bikes use electric energy as their driving force, which are limited to the battery. The batteries usually need to be replaced manual. When the battery is wearing out, the GPS device may receive a weak or no signal. Secondly, shared e-bikes are highly adept in navigating large roads and small alleys, when passes through small alleys, tunnels or roads covered by trees, the GPS signal might become too weak and the location of the vehicle may not be updated in a timely manner. All of this may lead to data losses over long periods of time. Additionally, certain difficult-to-avoid problems like device operation issues or packet loss may also lead to the loss of (some) trajectory data. Therefore, sparse trajectory data is always present in traffic data.

4. Cleansing and extraction of trajectory data

Since the raw data derived from the GPS devices consist only of the spatiotemporal trajectory points coordinates of the shared e-bikes, they do not directly reflect upon the trajectory and route information of each user. Therefore, our data cleansing and trajectory extraction procedures involve: (1) the classification of the endpoints of two adjacent trips from continuous and stateless raw data, (2) the removal of incomplete and invalid trajectory records, and (3) the extraction of the trajectory segment/O-D pair of each user.

Our trajectory cleansing method, which is suitable for the sparse data of shared e-bikes, satisfies the following requirements of trajectory extraction: efficacy (the ability to restore the O-D pair of the trajectory segment of each user), completeness (the sampling density of each trajectory segment is maintained at a certain level), accuracy (the errors of the information retrieval process are limited to a certain range) and rationality (the trajectories are matched with actual road networks to continuously adjust the algorithm).

4.1 Determination of data cleansing indices

4.1.1 *Selection of Cleaning Indices*

In the absence of riding status-related information, it is necessary to search for the parking points of each moving object, based on the spatiotemporal points coordinates provided by the object's GPS records, to identify the riding status of each trajectory (sample) point. The endpoints of two adjacent trips by the moving object may then be identified. The state of motion of a moving object is adjudged using the time difference, distance, and average

speed between two adjacent sample points and the instantaneous speed of each sample point.

In the following, t_i ($1 \leqslant i \leqslant n$) is defined as the timestamp of n trajectory points, while y_i ($1 \leqslant i \leqslant n$) and x_i ($1 \leqslant i \leqslant n$) are the latitude and longitude coordinates of the n trajectory points, and R is the Earth's radius (6371004 m in this paper). T_i is the time interval (in seconds) between the i-th and i+1-th points in the GPS trajectory data, d_i is the distance between the i-th and i+1-th points in the GPS trajectory data, $\overline{v_i}$ is the average speed (in km/h) of the moving object between the i-th and i+1-th points in the GPS trajectory data, $v_{si}$is the instantaneous speed of the moving object at the i-th point in the trajectory data.

$$T_i = |t_{i+1} - t_i| \quad (1 \le i \le n-1)$$

$$d_i = R \cdot \arccos\left(1 - \left(\begin{array}{l} \left(\sin\frac{(90-gpsx_i)\pi}{180}\cdot\cos\frac{gpsy_i\cdot\pi}{180} - \sin\frac{(90-gpsx_{i+1})\pi}{180}\cdot\cos\frac{gpsy_{i+1}\cdot\pi}{180}\right)^2 \\ +\left(\sin\frac{(90-gpsx_i)\pi}{180}\cdot\sin\frac{gpsy_i\cdot\pi}{180} - \sin\frac{(90-gpsx_{i+1})\pi}{180}\cdot\sin\frac{gpsy_{i+1}\cdot\pi}{180}\right)^2 \\ +\left(\cos\frac{(90-gpsx_i)\pi}{180} - \cos\frac{(90-gpsx_{i+1})\pi}{180}\right)^2 \end{array}\right) \div 2\right)$$

$$(1 \le i \le n-1)$$

$$\overline{v_i} = \frac{d_i}{T_i} \qquad (1 \le i \le n-1)$$

$$v_{si} = \frac{d_{i-1}+d_i}{|t_{i+1}-t_{i-1}|} \quad (2 \le i \le n-1)$$

4.1.2 *Preliminary threshold determination of indices*

To ensure that trajectory O-D points extraction is performed scientifically, the endpoints of a moving object's trips are determined by setting threshold values for the time interval, distance and average speed between two trajectory points and the instantaneous speed of each point. This provides a preliminary set of trajectories for each moving object.

The thresholds for the trajectory cleansing indices were initially defined as follows:

(1) Time interval of the sample points. If the sample interval is greater than 10 minutes (less than 10% of all samples), the trajectory is then adjudged to contain a loss of fragmentary data and is therefore invalid.

(2) Average speed between sample points. The walking speed of a normal person generally fluctuates around 1 m/s (approximately 4 km/h) (Duim et al., 2017). Therefore, if the instantaneous speed of a sample point (i.e., the riding speed of a shared e-bike) is less than that of a walking person, this point may be adjudged as a parking point. However, to avoid the misidentification of traffic-induced delays as parking points (e.g., traffic jams and traffic lights), the criterion for a sample point to be identified as a trip endpoint is riding speeds continuously lower than 4 km/h for two minutes.

4.2 Algorithm anomaly identification and modification based on actual road network

Firstly, the endpoints of the moving object's trips are preliminarily identified using threshold values. Then, combining with actual road networks, the preliminary results in section 4.1 are then corrected through anomaly identification and corrections.

(1) First algorithm modification

If the sampling interval of the trajectory points range between 2 minutes and 10 minutes, some of the intervening trips between the sample points may be lost. In this case, the endpoints of a trip must be judged according to the distance moved and speed of the shared e-bike. Based on a large number of threshold adjustment trials, setting that the moved distance greater than 200 m and average speed lower than 4 km/h, the last GPS point is determined to be a parking point. To restore trajectory segment in limit errors, the middle point between two trajectories is decided to be the destination point of the previous trajectory and the Original point of the next trajectory.

(2) Second algorithm modification

Based on the result of algorithm modification 1, excluding the invalid trajectories. If an identified trajectory segment only contains two sample points (i.e., only Original and Destination points), there may contain three scenarios as follows.

a. If the time difference between two sample points is greater than 2 minutes, one trip may have occurred during this period, and these two points are the original and destination points of this trip.

b. If the two sample points correspond to the original point of a trip and a point within a trip, respectively, other GPS trip data is missing, and it is impossible to determine the D point of the trip.

c. Both two points are GPS points within a trip, and it is impossible to determine the Original and Destination points of the trip.

In view of these scenarios, combining with actual road networks, and threshold adjustment trials, trajectory segments that only contain two sample points are determined to be invalid, which can be eliminated directly.

In addition, if a trip obtained after algorithm modification 1, is too short, the corresponding trajectory segment is then meaningless. The distance moved of a shared e-bike is therefore a criterion for the elimination of invalid trajectories. Based on road network matching tests and threshold adjustment trials, all trajectory segments with Euclidean distances (di) equal or less than 200 m are determined to be invalid trajectories. As indicated in Figure 2, a route that matches an actual road network is shorter than 50 m, it can be basically judged as the displacement error of the e-bike staying or the GPS equipment grasping.

Figure 2. A route shorter than 50 m.

4.3 Algorithm for trajectory cleansing and O-D extraction

On the basis of determination of the cleansing indices and the algorithm anomaly identification and modification by combining with actual road network, algorithm for trajectory cleansing and O-D extraction is identified.

Step 1: Time threshold-based trajectory segmentation

1) Trajectory data that correspond to different days are generally classified as different trajectory segments.

2) If the sampling interval, T_i, is greater than the threshold, T (10 minutes in this paper), no trajectory will be recorded for the shared e-bike in the corresponding period. The earlier sample point (point i) will be marked as the D (destination) point of a trip, and the later sample point (point i + 1) will be marked as the O (original) point of the next trip.

3) If the sampling interval is larger than two minutes and smaller or equal to T, the average speed $\bar{v}$ between the adjacent points in the trajectory segment is calculated (the distance is calculated using the Euclidean distance).

(i) If $\overline{v_i}$ < 4 km/h and Ti > 2min, the shared e-bike is considered to have parked during this period. In the trajectory segment between these points, the earlier point is labeled as the D point of the previous trip and the later point is labeled as the O point of the next trip.

(ii) if T_i < 2min and the average speed $\overline{v_i}$ is less than 4 km/h, the shared e-bike is considered to have stopped temporarily in this period. This trajectory will not be segmented.

Step 2: Speed threshold-based trajectory segmentation

In the trajectory segments that were extracted in Step 1, trajectory segmentation is performed in continuous trajectories with riding speeds lower than 4 km/h for more than two minutes, which also have instantaneous speeds (v_{si} and $v_{s(i+1)}$) less than 4 km/h in two continuous trajectory points (except for the original and destination points). The earlier sample point is marked with D point of the previous trip while the later sample point is marked with O point of the sequential trip.

Step 3: Removal of anomalous trajectory segments

In this step, the trajectory segments that were formed in Step 2 are traversed, and the trajectory segments with fewer than two sample points or d_i <= 200 m are deliminated.

5. Results of the cleansing trajectory in Tengzhou city

11178 valid ride trajectories were scrubbed from the week of disordered GPS trajectory data of shared e-bikes in Tengzhou. To better show the week-long commuting patterns in Tengzhou, the identified commuter travel was spatialized. Each line represents a riding trip (O-D points from origin to destination), riding time, riding distance and the ID of the shared e-bike in the attributes of the GIS layer. We spatialized the trajectory according to the time of the ride, as shown in Figure 4. The statistics show that the rides with a riding time of five minutes or less account for 30.59%, rides from five to ten minutes comprised 36.60% and rides longer than ten minutes accounted for 32.81%. The short distance travels mainly occur in central areas in Tengzhou city, which is similar with the study in Beijing. Through statistical analysis, it is indicated that trip distance for each individual is mostly during from 2km to 10km, which is conform to electric bicycles' service intention—making up for commuting demand of shared bicycle (usually within 2 km). Cases A, B and C shown in the three rectangles are amplifications of the cleaning trajectories of the yellow trips in Panel A, which were matched with Google Maps for validation analysis as described below. The reason for choosing these three cases are their commuting time, which represent three special moments—working time, off-work time and leisure time.

The sampling method was used to visually assess the matching between the experimental results and the road network to validate the accuracy of the cleaned trajectories and O-D points. The rationality analysis was performed by randomly sampling the scrubbed trajectory data and comparing these trajectory segments to the actual road network. A total of 100 travel trajectories were randomly sampled and 91 of these trajectories could be rationally matched to routes that are compatible with the actual road network, indicating that the method proposed for sparse trajectory GPS data is practicable.

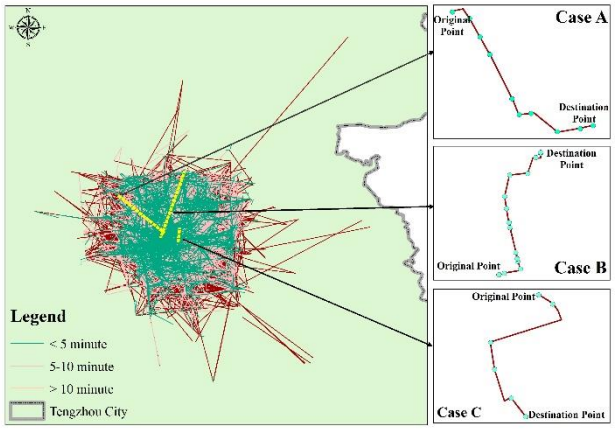

Figure 3. The week-long trajectories of shared e-bikes in Tengzhou city

To verify the superiority of the trajectory data cleansing method prosed in this paper, the cleaned route information from one shared e-bike in a same day are spatialized. As Figure 4 illustrated, it can be seen that the O-D points can

be derived well from continuous and disorder GPS trajectory points. 23 May, this shared e-bike was used 7 times, produced seven riding trajectories.

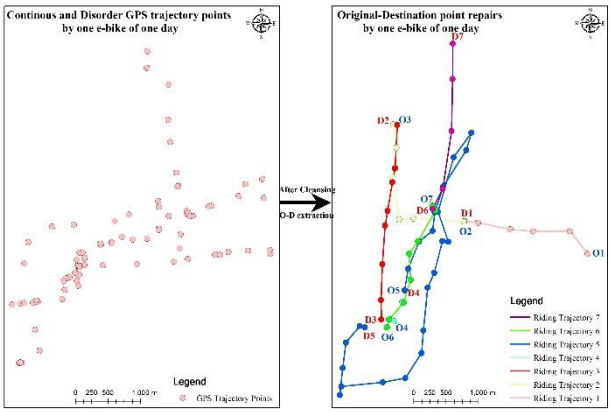

Figure 4. The cleansing method used in one shared e-bike in a same day.

Three trajectories (Case A, B and C in Figure) in the morning (working hours), late afternoon (off-work hours) and night (leisure time) were randomly selected and visualized combing actual road network in GoogleMap.

Fig. 4 describes the cleaned route information and its matching with the road network. Based on the actual road information of this route, the scrubbed trajectory segment is theoretically compatible with the moving object departing on the 25th of May at 07:03 from Yuanzhuang village, and then traveling along Pingxing North Road, Middle Pingxing Road, and Xingtan Road, before arriving at Tengzhou Central People's Hospital at 07:15.

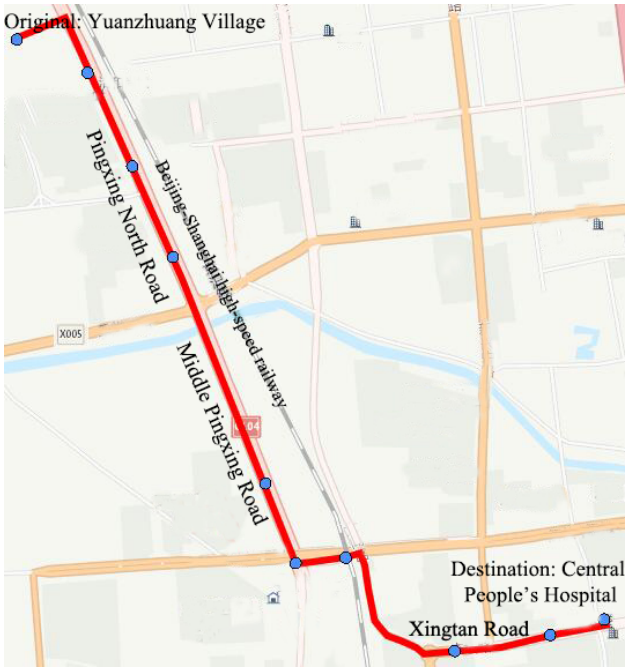

Figure 4. Trajectory of case A after matching with the actual road network in GoogleMap

Fig. 5 illustrates the route information of Case B and the matching of this trajectory with the road network. Based on the matching between the extracted trajectory segment and the road network, Case B is theoretically compatible with the moving object departing near an office area in Tengzhou City (the Tengzhou Bureau of Education, Tengzhou No. 1 Middle School and Tengzhou Central People's Hospital are all nearby), followed by a ride along Xingtan Road, Shanguo Middle Road, Shanguo North Road, and finally Beixin Middle Road, before arriving at its destination, the Huateng West District.

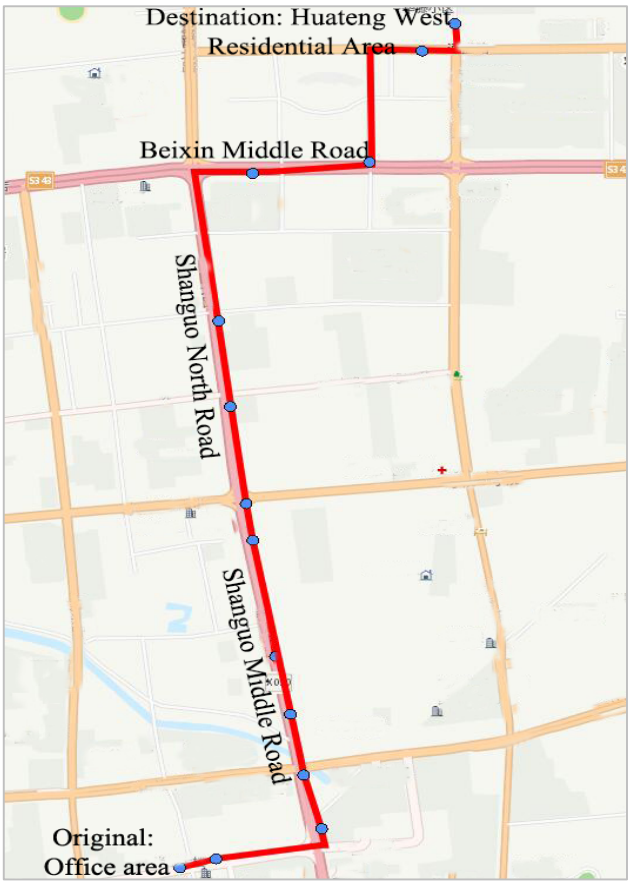

Figure 5. Trajectory of case B after matching with the actual road network in GoogleMap

Fig. 6 shows the route information of Case C and its match to the road network. Based on the matching between the trajectory and the road network in GoogleMap, the extracted trajectory is theoretically compatible with the moving object departing from the Central City A Unit Chun (residential – leisure area), and then traveling along Xingtan Road, Shanguo Middle Road, and finally Fuqian Road, before arriving at Chunqiuge Unit at 20:33.

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-22-2019 | © Authors 2019. CC BY 4.0 License.

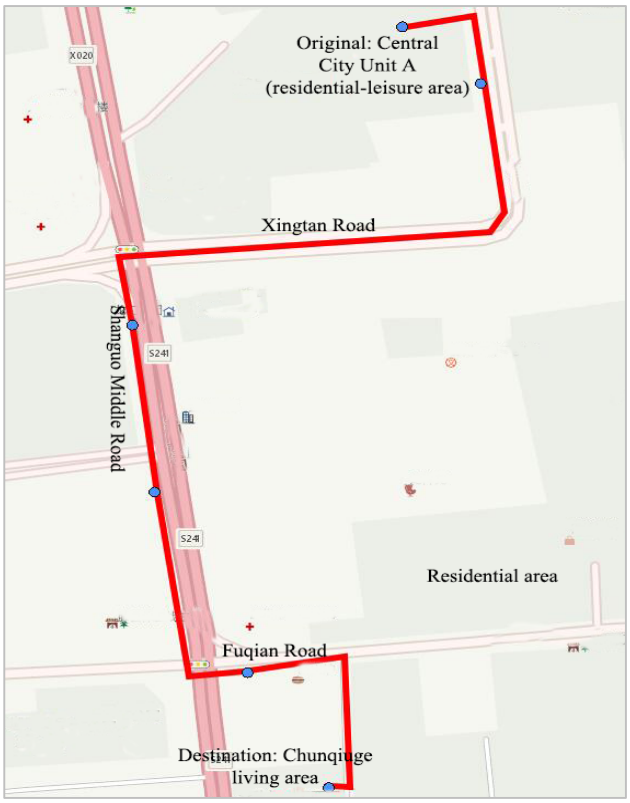

Figure 6. Trajectory of case C after matching with the actual road network in GoogleMap

In summary, our method for O-D pair extraction and cleansing from sparse trajectory data, which is based on the time difference, distance, and average speed between two sample points and the instantaneous speed of each sample point, has been proven to be viable. This method is especially suitable for continuous trajectory point data that do not possess riding status attributes.

6. Conclusion

Data cleansing and O-D point pairs extraction are prerequisites for the analysis of urban spatial structures and human mobility characteristics based on LBS data. Although studies have been conducted using the spatiotemporal data of mobile phone signals, taxis and shared bicycles, there are no reports in the literature about the use of shared e-bike data. By analysing the characteristics of the shared e-bike data of this study in Tengzhou city, a novel method for data cleansing and O-D point pairs extraction that is suitable for trajectory data characterized by the shared e-bikes. The conclusions are as follows:

(1) Due to the limit of bicycle battery and GPS devices grasping, the raw data quality of shared e-bikes is inherently sparse, characterized by un-continuous, heterogeneous GPS points, and without riding status information, which is a common case in traffic data grabbing.

(2) By using the indices of time difference, distance, and average speed between two adjacent sample points and the instantaneous speed of each sample point, the classification of the endpoints of two adjacent trips from continuous and stateless raw data can be identified.

(3) During the procedure of the GPS trajectory data cleaning, the determination and adjustment of the threshold value of the cleaning index should be combined with the real road network, to ensure the reasonableness of the cleansing results.

(4) The method proposed in this paper combining with actual road network, is proved to have applicability for data cleansing and O-D pair extraction from sparse trajectory data that lack attribute information (like riding status information) and with non-uniform GPS information. According to the analysis of the experimental results of the week-long trajectory data of shared e-bikes in Tengzhou city, it was proved that our method has an extraction accuracy of 91% by assessing 100 randomly sampled trajectories. Therefore, our method is rational, and applicable for the extraction and cleansing of trajectory information from the big and sparse GPS data of shared e-bikes.

Due to the characteristics of shared e-bike trajectories, fragmentation-induced data loss is inevitable. Nonetheless, our method is designed to maximally restore the trajectories corresponding to the lost data, to obtain information about the human mobility patterns. However, the valid trips that were extracted from the experimental dataset were distributed in a non-uniform manner, suggests that the trips are also affected by constraints and factors other than those specified in this study. In the future work, we will perform the interpolation and cleansing/extraction of sparse data with lacking of attribute information, by using methods like similarity analysis, and incorporate factors like population characteristics and urban land-use. And conducting the algorithm validation by using field observations.

7. References

Alexander, L., Jiang, S., Murga, M., et al. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C, 2015, 58:240-250.

Alsger, A., Assemi, B., Mesbah, M., et al. Validating and improving public transport origin – destination estimation algorithm using smart card fare data. Transportation Research Part C, 2016, 68:490-506.

Campbell, A.A., Cherry, C.R., Ryerson, M.S., et al. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Transportation Research Part C: Emerging Technologies, 2016,67:399-414.

Cui, J.X., Liu, F., Janssens, D., et al. Detecting urban road network accessibility problems using taxi GPS data. Journal of Transport Geography, 2016,51:147-157.

Duim, E., Lebrão, M.L., Antunes, J.L.F. Walking speed of older people and pedestrian crossing time. Journal of Transport & Health, 2017,5:70-76.

Gao, Q.L., Li, Q.Q., Yue, Y., et al. Exploring changes in the spatial distribution of the low-to-moderate income group using transit smart card data. Computers, Environment and Urban Systems, 2018, 72:68-77.

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M., et al. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. Transportation Research Record Journal of the Transportation Research Board, 2013, 2343(-1):17-24.

Huang, J., Levinson, D., Wang, J.E., et al. Tracking job and housing dynamics with smartcard data. PNAS, 1815928115.

Janković, B., Nikolić, M., Vukonjanski, J., et al. The impact of Facebook and smart phone usage on the leisure activities and college adjustment of students in Serbia. Computers in Human Behavior, 2016,55(A):354-363.

Jensen, P., Rouquier, J.B., Ovtracht, N., et al. Characterizing the speed and paths of shared bicycle use in Lyon. Transportation Research Part D, 2010, 15(8):522-524.

Kim, J., Corcoran, J., Papamanolis, M. Route choice stickiness of public transport passengers: Measuring habitual bus ridership behaviour using smart card. Transportation Research Part C: Emerging Technologies, 2017,83:146-164.

Kou, Z.Y., Cai, H. Understanding bike sharing travel patterns: An analysis of trip data from eight cities. Physica A: Statistical Mechanics and its Applications, 2019,515(1):785-797.

Liu, L., Biderman, A., Ratti, C. Urban mobility landscape: Real time monitoring of urban mobility patterns. 2009.

Liu, Y., Kang, C., Gao, S., et al. Understanding intra-urban trip patterns from taxi trajectory data. Journal of Geographical Systems, 2012, 14(4):463-483.

Long, Y., Shen, Z. Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing[M]// Geospatial Analysis to Support Urban Planning in Beijing. Springer International Publishing, 2015.

Long, Y., Thill, J-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Computers, Environment and Urban Systems, 2015,53:19-35.

Long, Y., Zhang, Y., Cui C.Y. Identifying commuting pattern of Beijing using bus smart card data. Journal of Geographical Sciences, 2012,67(10):1339-1352.

Munizaga, M., Devillaine, F., Navarrete, C., et al. Validating travel behavior estimated from smartcard data. Transportation Research Part C, 2014, 44(4):70-79.

Nassir, N., Khani, A., Sang, G.L., et al. Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System. Transportation Research Record Journal of the Transportation Research Board, 2011, 2263(-1):140-150.

Schuessler, N., Axhausen, K.W. Processing Raw Data from Global Positioning Systems Without Additional Information. Transportation Research Record Journal of the Transportation Research Board, 2009, 2105(2105):28-36.

Shaheen, S., Zhang, H., Martin, E., et al. China's Hangzhou Public Bicycle. Transp Res Rec J Transp Res Board, 2247,2011:33-41,

Sørensen, A. Ø., Bjelland, J., Bull-Berg, H., et al. Use of mobile phone data for analysis of number of train travelers. Journal of Rail Transport Planning & Management, 2018,8,2:123-144.

Tang, J., Liu, F., Wang, Y., et al. Uncovering urban human mobility from large scale taxi GPS data. Physica A Statistical Mechanics & Its Applications, 2015, 438:140-153.

Tang, Y., Pan, H.X., Fei, Y.B. Research on Users' Frequency of Ride in Shanghai Minhang Bike-sharing System. Transportation Research Procedia, 2017,25:4979-4987.

Wolf, J.L. Using GPS data loggers to replace travel diaries in the collection of travel data[C]// Dissertation, Georgia Institute of Technology, School of Civil and Environmental Engineering. 2000:58--65.

Xu, J., Difference of urban development in China from the perspective of passenger transport around Spring Festival [J]. Applied Geography, 2017,84:85-96.

Yang, X.H., Cheng, Z., Chen, G., et al. The impact of a public bicycle-sharing system on urban public transport networks. Transportation Research Part A: Policy and Practice, 2018,107:246-256.

Zhang, Y.P., Mi, Z.F. Environmental benefits of bike sharing: A big data-based analysis. Applied Energy, 2018,220(15):296-301.

Zheng, Z., Rasouli, S., Timmermans, H. Evaluating the Accuracy of GPS-based Taxi Trajectory Records. Procedia Environmental Sciences, 2014, 22:186-198.

Zhong, C., Huang, X., Arisona, S.M, et al. Inferring building functions from a probabilistic model using public transportation data. Computers Environment & Urban Systems, 2014, 48(6):124-137.

Zhou, Z.G., Yu, J.J., Guo, Z.Y., et al. Visual exploration of urban functions via spatio-temporal taxi OD data. Journal of Visual Languages & Computing, 2018,48:169-177.

Xiao X.Q. A study about sparse trajectory similarities between vehicles in urban traffic [D]. Fudan University, 2014.

Zou L., Zhang Z.Z., Zhu L.X. Public transportation OD data collection based on mobile location technology. Computer and Communications, 2011, 29(5):122-126.