

Characteristics analysis for travel behavior of transportation hub passengers using mobile phone data

Gang Zhong^{1,2,3}  · Tingting Yin⁴ · Jian Zhang^{1,2,3} · Shanglu He⁵ · Bin Ran^{1,2,3}

Published online: 13 April 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract The travel behavior of passengers from the transportation hub within the city area is critical for travel demand analysis, security monitoring, and supporting traffic facilities designing. However, the traditional methods used to study the travel behavior of the passengers inside the city are time and labor consuming. The records of the cellular communication provide a potential huge data source for this study to follow the movement of passengers. This study focuses on the passengers' travel behavior of the Hongqiao transportation hub in Shanghai, China, utilizing the mobile phone data. First, a systematic and novel method is presented to extract the trip information from the mobile phone data. Several key travel characteristics of passengers, including passengers traveling inside the city and between cities, are analyzed and compared. The results show that the proposed method is effective to obtain the travel trajectories of mobile phone users. Besides, the

✉ Gang Zhong
anhuizhonggang@126.com

Tingting Yin
yttwen@163.com

Jian Zhang
jianzhang@seu.edu.cn

Shanglu He
slhemickey@126.com

Bin Ran
bran@seu.edu.cn

¹ Jiangsu Key Laboratory of Urban ITS, School of Transportation, Southeast University, Nanjing 210096, China

² Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing 210096, China

³ Jiangsu Province Collaborative Innovation Center for Technology and Application of Internet of Things, Nanjing 210096, China

⁴ Jiangsu Expressway Company Limited, Nanjing 210049, China

⁵ School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China

travel behavior of incity passengers and external passengers are quite different. Then, the correlation analysis of the passengers' travel trajectories is provided to research the availability of the comprehensive area. Moreover, the results of the correlation analysis further indicate that the comprehensive area of the Hongqiao hub plays a relatively important role in passengers' daily travel.

Keywords Mobile phone data · Travel behavior · Transportation hub · Digital travel trajectory · Correlation analysis

Introduction

Passenger transportation hub is the transfer center for passengers to transfer between different transportation modes. The rapid development of aviation, high-speed railway and metro has made the passengers' daily travel more convenient, which also has enhanced the importance of the passenger transportation hubs. A comprehensive passenger transportation hub may attract hundreds of thousands of passengers on a typical day. Researchers have paid their attention to monitoring and modeling the passenger flow inside the transportation hubs, utilizing the methods such as simulation (Zhang et al. 2008; King et al. 2014) and field research (Cheung and Lam 1998; Srikukenthiran et al. 2013). The travel behavior of hub passengers outside the hub but within the city scope is also important, because the transportation planners need such information to analyze the traffic demand related to the hub and design supporting traffic facilities, such as bus lines and metro lines in the city. It is also crucial to understand passengers' travel patterns and figure out whether the patterns are different from the general human mobility patterns. However, the travel behavior of the hub passengers in the urban areas has received little attention, mainly because it is difficult to collect the passengers' trip information. Travel surveys like questionnaires and telephone interviews are not cost effective methods because they are time and labor consuming. Insufficient information poses an obstacle to the understanding of the interaction relationship between the operation of the transportation hubs and the passengers' travel behavior.

In recent years, there are several new data sources which are used to study the human travel behavior, including GPS data, social network data, mobile phone data, etc. Compared to the traditional travel survey methods, the new data sources can record the users' locations on the digital maps which can trace the users' travel trajectories conveniently and timely. The position accuracy of the GPS data is at meter level which is higher than that of other data sources. However, the GPS data used for the travel behavior research is normally collected from a limited number of volunteers. Thus, the researchers usually use the GPS data to study the individuals' travel behavior and travel patterns (Shen et al. 2013). The social network data has a relatively larger sample size, while the samples tend to be biased since the subscribers are from only certain social groups (Sagl et al. 2014). As the involuntary location information generated by the entire users, the sample size of the mobile phone data is huge and the samples are unbiased. Therefore, the mobile phone data is suitable for studying the general human travel behavior. Furthermore, special categories of users, such as the hub passengers, can also be studied without the limit of the sample coverage, which makes the mobile phone data an ideal data source for this study.

The mobile phone data has already been used to collect the traffic information in previous studies, including the traffic state (Zhang et al. 2015; He et al. 2016) and the human travel behavior (Gonzalez et al. 2008; Calabrese et al. 2013). The data record is generated,

which contains a timestamp and an approximate location, when a mobile phone connects to the communication network for events like texting, calling, etc. To analyze the travel behavior, the users can be traced by studying the digital footprints they left in the wireless communication space. The privacy of the users is protected by encrypting the user ID in the data records. Furthermore, the ownership ratio of the mobile phones has increased sharply in the past decades, even in the developing countries, which guarantees the mobile phone data has a sample coverage which is large enough for studying the travel behavior.

Although the mobile phone data can't provide personal information of the users, such as age, gender, and income, it tends to be an effective way to characterize the general travel behavior of the hub passengers using the mobile phone data. This study tries to analyze the passengers' travel behavior of an international transportation hub in Shanghai, China. In the next section, a literature review is presented to briefly introduce the previous studies about extracting travel trajectories and analyzing human travel behavior using the mobile phone data. Next, the transportation hub and the mobile phone data in this study are described, due to that the information is necessary for the following research. Then, a method is proposed to identify the staying points and collect the trip information of the mobile phone users. The fifth section details the results of the travel behavior analysis. In the sixth section, the availability analysis for the comprehensive area is illustrated and analyzed. Finally, the conclusion and the future work are addressed.

Literature review

Travel trajectory extracting

With the rapid development of cellular technology, the mobile phone data has become an outstanding data source to analyze the travel behavior of users. The mobile phone data contains the temporal and spatial information of each active mobile phone, which is essential for extracting the travel trajectories of the users. To study the travel behavior, it is necessary to transform the mobile phone data into trip data (Asakura and Hato 2004). Therefore, stay locations where users conduct their activities should be identified based on the mobile phone data. In previous studies, there were two principal methods presented to solve the problem.

In *the first method*, the mobile phone records of each user are clustered in chronological order. The trips are then formed with the continuous stop clusters. In this method, the temporal intervals and the spatial distances between location points (i.e., the records) are calculated, so that the temporal and the spatial thresholds can be used to cluster the location points. Asakura and Hato (2004) first illustrated a labeling algorithm with the temporal and spatial thresholds to provide trip attributes to each point including moving and stopping. A cluster of continuous stopping points would be identified as a stay location. Phithakkitnukoon et al. (2010) divided the continuous location points into discrete clusters based on the distance criterion. A cluster was set as an activity location if the duration time of the cluster exceeded a certain time threshold. Calabrese et al. (2013) calculated the maximum distance and the maximum time interval between location points in a cluster to verify whether they satisfied the thresholds as an activity location.

Different from the first one, *the second method* disregards the time order when identifying stay locations, which usually needs multiple days of records. Alexander et al. (2015) and Jiang et al. (2017) adopted multiple days of records to divide the user's total location

points on these days into several clusters using the spatial criterion. The location points in the same cluster were close in space, while they may be far apart in time (even on different days). Then, the cluster was regarded as a stay location if at least one of the user's stay durations in the cluster exceeded the temporal threshold on the research days. Therefore, all stay locations on the research days were identified for the user, although he/she didn't have to stop at all the stay locations on each day. In a related study, Chen et al. (2014) proposed an innovative method using a logistic regression model to estimate the attributes of the clusters (staying or moving) based on a shape variable and a volume variable. This work testified that location points in the staying clusters were more likely to be distributed in all directions, whereas moving points were more likely to be distributed along a linear line.

In this study, it is not appropriate to use *the second method* to extract travel trajectories of the hub passengers. Since some of the hub passengers are not the city residents, it is difficult to find multiple days of their records in the city to support the study. They also tend not to visit a place for several times in the urban area. However, the previous studies related to *the first method* didn't consider the distribution characteristics of location points illustrated in Chen et al. (2014). The characteristics should be helpful to estimate the trip attitudes of the location points. Therefore, this study attempts to combine the temporal-spatial thresholds and the distribution characteristics of location points to propose a novel travel trajectory extracting method.

Travel behavior analysis

There were a number of previous studies analyzing the users' travel behavior based on the collected travel trajectories. Gonzalez et al. (2008) discovered a vital rule governing people's movements that the distribution of the trip distances over all users was well approximated by a truncated power-law. Calabrese et al. (2013) used the distribution to test the trips they identified and found that the truncated power-law had slightly different coefficients for people living in different regions, which is attributed to different built environments. According to the basic distribution rule, the regularity and predictability of the human's mobility were explored by measuring the entropy of each individual's travel trajectory. The accuracy of the predict algorithms was surveyed to show the difference between predictability in practice and theory (Song et al. 2010). Besides the travel patterns, the trip information required in the travel survey was also studied, such as the origin–destination (OD) estimation. Previous studies had proposed methods to extract OD data over the city scope. Pan et al. (2006) proposed the procedure of setting up a cellular-based OD collecting system and collected the trip distribution between each OD pair. Zhang et al. (2010) designed a mathematic model to convert the cellular counts into equivalent vehicle counts according to the cellular probe trajectories. Calabrese et al. (2011) studied the OD flow in different time durations to analyze the periodical traffic patterns. Frias-Martinez et al. (2012) adopted certain temporal association rules to identify the home/work locations and proposed a method to estimate commuter trips. Fang et al. (2014) conducted case studies in four major cities in China and validated the feasibility of proposed methodologies. In addition, the data from social networks was also applied in the research of estimating OD matrix with the mobile phone data for comparison (Rokib et al. 2015).

This study focuses on a transportation hub to analyze the travel behavior of passengers, which is different from previous studies. Moreover, the mobile phone data records in this study contain more events than the call detail records used in previous studies (Calabrese et al. 2013; Gonzalez et al. 2008), such as handover, location update, etc. Thus, the data

records tend to have better temporal precision for characterizing the travel trajectories, while the data processing work is more complicated.

Datasets

Study area

This study is mainly focused on the area of the Hongqiao hub located in the city of Shanghai, China, which can be categorized into the core area and the comprehensive area. Figure 1a shows the location of the Hongqiao hub in the city and enlarges the two areas in the top left corner. The area circled by the blue line is the central urban area. Figure 1b illustrates the relation between different areas. The definition of each area is explained as follows:

The core area is the distribution center of passengers from different transportation modes, which contains an airport terminal, a railway station, a coach station, two subway stations, and several bus stations.

The comprehensive area is an area planned based on the core area with the supportive functions such as residence and commerce. This area is located adjacent to the central urban area. The main aim of the comprehensive area is to provide services for the hub passengers and promote the economic development utilizing the passenger attraction of the core area. In this paper, the two areas are separated to avoid misunderstanding, i.e., the comprehensive area doesn't cover the core area.

The communication area is an active area in the wireless communication space for passengers in the core area. Because the wireless signal is not restricted by the physical boundary, the mobile phones of passengers can connect the base stations around the core area. Thus, the area is defined as the base stations mainly serving for passengers in the core area. In our previous study (Zhong et al. 2017), this virtual area was proposed to obtain the passengers' information from the mobile phone dataset. This study continues to use the obtained information (as illustrated in *Mobile phone data* subsection) to study the travel behavior of hub passengers.

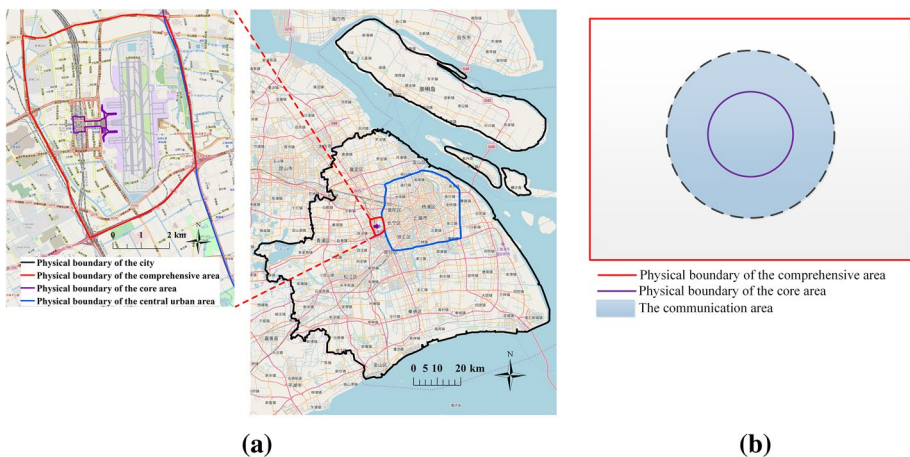


Fig. 1 Geographic information of Hongqiao transportation hub. **a** Location of the hub area in the city. **b** Sketch map of the hub area

The core area enclosed by the purple line is around 50,000 m², while the comprehensive area enclosed by the red line is around 26.26 km². As one of the largest transportation hubs in China, the Hongqiao hub serves for hundreds of thousands of passengers every day. It plays a significant role in the development of the city and the country. The passenger information on the official website shows that the average daily number of passengers in the whole hub area is around 684,300 in 2013 (Shanghai Hongqiao Central Business District 2013). The huge passenger volume indicates the importance of analyzing the travel behavior of passengers, which is necessary for planning the interchange facilities related to the hub such as bus lines, metro lines, etc. Moreover, the passengers' travel behavior also affects the development of the comprehensive area.

Mobile phone data

The mobile phone dataset was collected from roughly 18 million users in Shanghai over a period of 12 days from November 19 to 30 in 2013. The record is generated when a device connects to the cellular network for events like calling, texting, handover, location update, etc. There are five segments in each record, including an anonymous user ID, a Location Area Code (LAC), a Cell ID, a timestamp, and an event ID. The two segments, LAC and Cell ID, can be used to locate the users at the cellular tower level. More than 60,000 cellular towers can be found in Shanghai from the dataset, with a spacing gap ranging from under 50 m in the downtown area to around 1500 m in the suburb district, as shown in Table 1. Overall, most cellular towers are less than 200 m away from the closest neighbor towers. Besides, the Voronoi grid is employed to display the distribution of the approximate reception area of each tower (Gonzalez et al. 2008) in Fig. 2. In Fig. 2a, it can be found that the density of the Voronoi grid in the central urban area tend to be higher than that in the suburb area. Figure 2b displays the distribution of the grids in the comprehensive area of the Hongqiao hub.

In order to explore the travel behavior of passengers, the hub passengers should be first identified from the mobile phone users. In our previous study (Zhong et al. 2017), a method was introduced to identify *the communication area* of the hub and extract the users having appeared in the core area. Besides, a classification method was also presented to distinguish incity passengers from external passengers. The definitions of the two categories of passengers are listed as follow:

Incity passengers - people who travel inside Shanghai using the travel modes in the hub, mainly composing of residents.

External passengers - people who travel between Shanghai and other cities using the travel modes in the hub, mainly composing of tourists, businessmen, etc.

Based on the identification results, two datasets are extracted to research the travel behavior of passengers as follow:

Dataset D1 - This dataset contains the records of all passengers who have appeared in the core area of the hub on two workdays (November 20–21, 2013) and two non-workdays (November 23–24, 2013, weekend). The average number of unique user IDs is about 400,000

Table 1 The percentage of cellular towers in each spacing gap range

Range (m)	0–50	50–100	100–200	200–500	500–1000	1000–1500	> 1500
Percentage	39.42	28.76	20.28	9.88	1.44	0.18	0.04

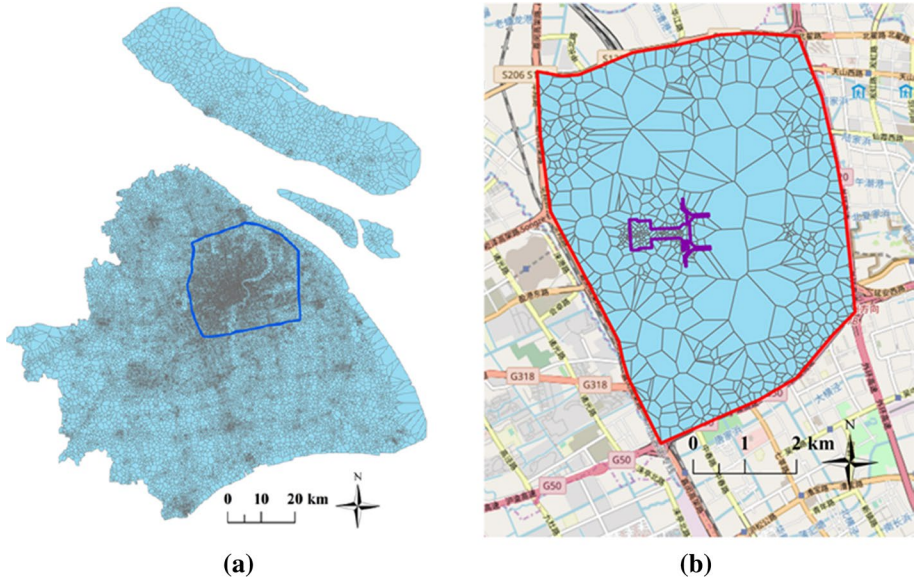


Fig. 2 The approximate reception area of each cellular tower shown as Voronoi grid. **a** Voronoi grids in the city of Shanghai. **b** Voronoi grids in the comprehensive area of the Hongqiao hub

on each day, including both incity passengers and external passengers. This study attempts to employ the dataset to explore the travel behavior of passengers in the following sections.

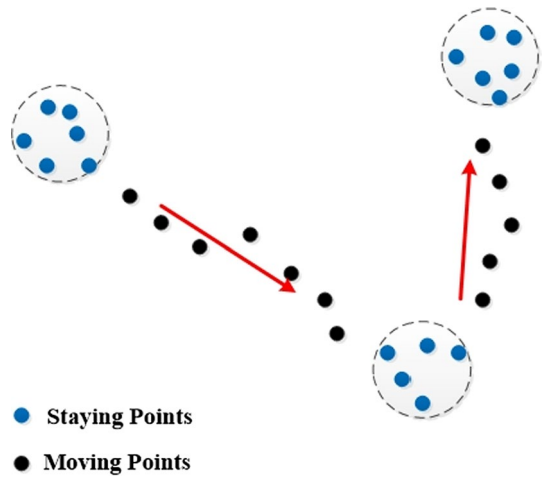
Dataset D2 - This dataset captures the records of 400,000 users randomly picked from the users who can be observed on each day in the mobile phone dataset. These users appeared in the city on 12 continuous days, which indicates they tend to have the trip characteristics of residents. The dataset contains the data of these users on four workdays (November 19–22, 2013), which is used to test the travel trajectory extracting method for validation and comparison.

The inter-event time, i.e., the time interval between two consecutive records, should be first analyzed for each user to manifest that the events generating records are frequent enough to characterize the passengers' movements. The average inter-event times of all users in D1 and D2 are respectively around 17 and 45 min. They are much lower than that in Calabrese et al. (2013) because the dataset in this study contains more events than the call detail records (CDRs).

Method of extracting passenger trajectory

The mobile phone data contains the spatial and temporal information of the users. It is necessary to transfer the information into travel trajectories that are normally used in travel behavior analysis. The attributes of the location points. (i.e., the data records) should be decided first, whether moving or staying. Chen et al. (2014) illustrated that staying points in one place were more likely to be distributed in all directions, whereas moving points were more likely to be distributed along a linear line. A cluster of staying points tend to be located in a circle with a limited radius as shown in Fig. 3. The attributes can be figured out by measuring the minimum radius of the circle that can enclose the tested location points.

Fig. 3 Distribution characteristics of the location points (staying points and moving points)



Based on the moving-or-staying identification, travel trajectories can be recognized from the mobile phone data.

Minimum enclosing circle problem

To find the minimum circle that can enclose the tested location points, a classical problem needs to be solved, i.e., minimum enclosing circle problem, which is an interesting problem having been studied for decades (Xu et al. 2003). Given a cluster of points on the Euclidean plane, $P = \{p_1, \dots, p_n\}$, the problem is to find the circle with the minimum radius that encloses all points in the cluster. In this study, the set of points in the problem is a set of location points in the mobile phone dataset. The randomized incremental algorithm (RIA) (De Berg et al. 2008) with the time complexity as $O(n)$ is utilized to solve the problem. The process of the algorithm needs to add a new point into the existing cluster first, which is similar to the process in our identification algorithm.

Moving-or-staying identification algorithm

Considering the distribution characteristics of the location points, our identification algorithm is developed following the work in Asakura and Hato (2004). Their algorithm is briefly introduced as following:

1. The point at the present time t has two possible attributes including staying and provisionally moving. The second one means the specific moving-or-staying status can't be decided at the moment. The attribute of the point at the following time $t+1$ should be identified next.
2. If the point t is staying, (\bar{x}, \bar{y}) is used to represent the averaged coordinate of the preceding N staying points till time t . The distance is calculated between the averaged position and the position of point $t+1$, (x_{t+1}, y_{t+1}) . Point $t+1$ is identified as staying point, if the distance is smaller than the given threshold, otherwise the attribute is provisionally moving.

3. If point t is provisionally moving, the distance between point t and point $t + 1$ is compared with the threshold. When the distance is smaller, both points are set as staying. Otherwise, point t is staying and point $t + 1$ is provisionally moving.
4. Time constraint is used to eliminate the stops with short stay time. The attributes of points at the stop are replaced by moving, if the staying time is shorter than the time threshold.

In step 2 of the algorithm, the distance between the averaged position of the preceding staying points and the position of point $t + 1$ is used to decide the attribute of point $t + 1$. However, the algorithm may make mistakes when the user moves slowly, such as walking along a street. The distance between the starting point and point $t + 1$ may be long enough for a trip, but point $t + 1$ can still be identified as staying because the distance gets shorter when the positions of the preceding points are averaged. The problem can affect the trip generation by neglecting a trip between two staying locations or dividing one trip into several segments with additional identified staying locations.

To avoid this problem, circles are utilized to limit the distribution of the staying points. The attributes of the location points are identified based on the radius of the minimum enclosing circle. If the minimum radius of the circle is short enough, the time interval of the cluster is also needed to be calculated. The attributes of the points are set as staying at one place if the time interval exceeds the threshold. Once the radius is longer than the critical radius after adding a new point to the cluster, the point is possibly moving. The attributes of the points are identified in chronological order.

For the i -th point p_i , there are three possible initial attributes including staying, provisionally moving and provisionally staying. The latter two attributes mean that there are two possibilities for this point, while the specific attribute of the point cannot be determined at this moment. More specifically, the i -th point is displayed as $p_i = \{x_i, y_i, t_i\}$, where (x_i, y_i) is the coordinates of the point in UTM reference frame and t_i is the timestamp of the point. The *first* and the *last* point of a user on a certain day are set as *staying* to study the travel behavior of both incity passengers and external passengers.

First case: p_i is staying

Since p_i is staying, there is a cluster of points staying at the same place with p_i which can be defined as the k -th staying cluster. The k -th staying cluster is denoted by:

$$P_S^k = \{p_{i-m+1}, \dots, p_i\}, m = 1, 2, 3 \dots \tag{1}$$

To identify the attribute of the $i + 1$ -th point, p_{i+1} is added into the cluster P_S^k temporarily. Then the radius of the minimum enclosing circle of the cluster, R_{i+1} , is calculated. When R_{i+1} is smaller than the threshold radius, the attribute of p_{i+1} is identified as staying. The k -th staying cluster is formally set as $P_S^k = \{p_{i-m+1}, \dots, p_{i+1}\}$. The staying place of the cluster is updated to the center of the minimum enclosing circle, (x_{i+1}, y_{i+1}) . If R_{i+1} exceeds the threshold radius, the attribute of p_{i+1} is set as provisionally moving while the k -th staying cluster remains unchanged, $P_S^k = \{p_{i-m+1}, \dots, p_i\}$. There is a temporary cluster to store the points whose attributes are temporarily uncertain, which is denoted as $P_T = \{p_{i+1}\}$. Then the algorithm goes to the next location point.

Second case: p_i is provisionally moving

The point p_{i+1} should be put into the temporary cluster $P_T = \{p_i\}$ to calculate the radius of the minimum enclosing circle. If R_{i+1} is still longer than the threshold radius, p_i is identified as moving and put into the moving cluster which stores the moving points between two consecutive staying clusters. The k -th moving cluster is denoted by:

$$P_M^k = \{p_{i-m+1}, \dots, p_i\}, m = 1, 2, 3 \dots \tag{2}$$

In the meantime, p_{i+1} is categorized as provisionally moving while the temporary cluster is set as $P_T = \{p_{i+1}\}$. Otherwise, the time interval between p_i and p_{i+1} (denoted as $\Delta t_{i,i+1}$) is measured, when R_{i+1} is within the threshold radius. If $\Delta t_{i,i+1}$ exceeds the threshold time interval, p_i and p_{i+1} are both identified as staying. They compose a new staying cluster $P_S^{k+1} = \{p_i, p_{i+1}\}$. If $\Delta t_{i,i+1}$ is shorter than the threshold time interval, the attributes are set as provisionally staying for the two points, $P_T = \{p_i, p_{i+1}\}$.

Third case: p_i is provisionally staying

The points in the temporary cluster $P_T = \{p_{i-m}, \dots, p_i\}, (m = 1, 2, 3 \dots)$ containing no less than two elements are with the same attribute, i.e., provisionally staying. The radius of the minimum enclosing circle is calculated after adding p_{i+1} into the cluster. In the case that R_{i+1} is longer than the threshold radius, the points in P_T are all identified as moving except p_{i+1} which is identified as provisionally moving. When R_{i+1} is within the threshold radius, the time interval between p_{i-m} and p_{i+1} need to be counted. If $\Delta t_{i-m,i+1}$ exceeds the threshold time interval, the points in the cluster P_T are identified as staying. The new staying cluster is $P_S^{k+1} = \{p_{i-m}, \dots, p_{i+1}\}$. Otherwise, the attributes of the points in the cluster P_T are still provisionally staying, $P_T = \{p_{i-m}, \dots, p_{i+1}\}$.

Generating travel trajectory

To generate the travel trajectory of a given user, the results of the identification need to be further processed. For a set of continuous staying points in the staying cluster P_S^k , they are all assumed to be located at one virtual place, i.e., the center of the minimum enclosing circle, (x_k^s, y_k^s) . In addition, the travel trajectory also needs to contain the temporal information. The arrival time (t_k^a) and the departure time (t_k^d) should be figured out for the user staying at the virtual place. In the work of Asakura and Hato (2004), the departure time was defined as the half of the observe times of the last point in a staying cluster and the first point in the next moving cluster. The arrival time was defined as the same way. The results of the method tend to be rough, especially when the average inter-event time of the data is large. Widhalm et al. (2015) proposed a method to estimate the arrival/departure time based on an assumed lower bound for the travel time between two staying clusters. However, the method tends to be less reliable when there is no certain rule to set the lower bound for the travel time. In this paper, a more detailed method is proposed to estimate t_k^a and t_k^d based on the information of the location points in the staying cluster and the adjacent moving clusters.

For the staying cluster $P_S^k = \{p_i, \dots, p_{i+b}\}$, the adjacent moving clusters are assumed to be $P_M^l = \{p_{i-a}, \dots, p_{i-1}\}$ and $P_M^{l+1} = \{p_{i+b+1}, \dots, p_{i+b+c}\}$ ($a, c=2,3,\dots, b=1,2,3,\dots$), as shown in Fig. 4. The average moving speed of the user in P_M^l can be calculated using the distance ($\Delta d_{i-a,i-1}$) and the time interval ($\Delta t_{i-a,i-1}$) between p_{i-a} and p_{i-1} , $v_l = \Delta d_{i-a,i-1} / \Delta t_{i-a,i-1}$. It is assumed that the average moving speed from the departure

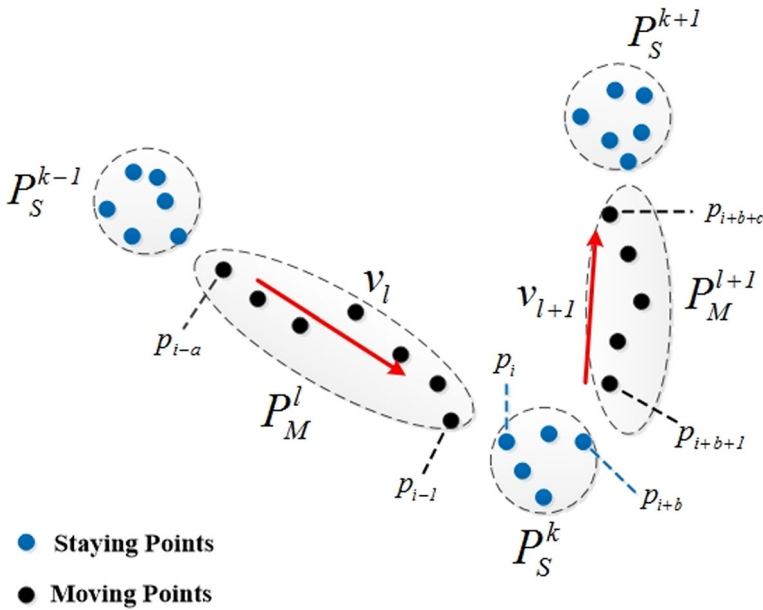


Fig. 4 Staying clusters and moving clusters

of P_S^{k-1} to the arrival of P_S^k is equal to v_l . Thus, the arrival time in theory (\tilde{t}_k^a) of P_S^k can be calculated. The actual arrival time is set as the minimum number of the theory value and the timestamp of p_i .

$$\tilde{t}_k^a = t_{i-1} + \frac{\sqrt{(x_{i-1} - x_k^s)^2 + (y_{i-1} - y_k^s)^2}}{v_l} \tag{3}$$

$$t_k^a = \min(\tilde{t}_k^a, t_i) \tag{4}$$

According to the same method, the departure time of P_S^k is calculated as:

$$\tilde{t}_k^d = t_{i+b+1} - \frac{\sqrt{(x_k^s - x_{i+b+1})^2 + (y_k^s - y_{i+b+1})^2}}{v_{l+1}} \tag{5}$$

$$t_k^d = \max(\tilde{t}_k^d, t_{i+b}) \tag{6}$$

However, the average moving speed between two staying clusters can't be obtained when there is no more than one point in the adjacent moving cluster between them. For the staying cluster $P_S^k = \{p_i, \dots, p_{i+b}\}$ having no adjacent moving clusters with more than one point, the method is simplified as follows.

If there is no point in the adjacent moving cluster,

$$t_k^a = t_i \quad (7)$$

$$t_k^d = t_{i+b} \quad (8)$$

If there is only one point in the adjacent moving cluster,

$$t_k^a = 1/2(t_{i-1} + t_i) \quad (9)$$

$$t_k^d = 1/2(t_{i+b} + t_{i+b+1}) \quad (10)$$

Based on the analysis, the k -th trip stop of the user is defined as $S_k = (x_k^s, y_k^s, t_k^a, t_k^d)$. Two consecutive trip stops compose one trip candidate of the user, while not all trip candidates are valid. Li (2015) classified the collected trip candidates into four categories to exclude the invalid ones, which is also used in this study to generate the valid trips. The travel trajectory of the user is composed of the valid trips in chronological order.

Sensitivity analysis for the spatial and temporal thresholds

The threshold radius and the threshold time interval are critical in the moving-or-staying identification algorithm. Therefore, the sensitivity analysis is necessary to choose appropriate threshold values. In this study, the data in D2 on November 20, 2013 (a workday) is used to do this work. The travel trajectories of users in D2 are extracted with different threshold values. The results of the average trip frequency are collected under these different conditions, as shown in Table 2.

As illustrated in the definition of dataset D2 in the *Datasets* section, the users in D2 tend to have the trip characteristics of residents. According to the travel survey of residents in Shanghai (Shanghai City Comprehensive Transportation Planning Institute 2010), the average trip frequency of the residents is 2.23. It can be found from Table 2 that the trip frequency is closest to the travel survey result when the threshold radius is set as 350 m and the threshold time interval is set as 30 min.

Result analysis of travel trajectory

The proposed method is first used to extract the travel trajectories of users in D2 to explore the travel behavior of residents in Shanghai for validation. In this study, the threshold radius is set as 350 m while the threshold time interval is set as 30 min. As an example, Table 3 shows the travel trajectory of a user through an entire day. The arrival time of S_1 and the departure time of S_5 are respectively the timestamps of the first and the last records on the day.

The trip length distribution is studied and displayed in Fig. 5, which shows that the trip length ranges from 1 to 120 km. Although the parameters are slightly different, the

Table 2 The results of the average trip frequency under different conditions

	Temporal threshold value/min	Spatial threshold value/m			
		150	250	350	450
10		2.44	2.63	2.71	2.80
20		2.19	2.33	2.41	2.48
30		2.06	2.18	2.24	2.30

Table 3 An example of the travel trajectory

User ID	S_k	x_k^s	y_k^s	t_k^d	t_k^d
308184	S_1	351,551.6141	3,466,581.0516	00:58:16	08:15:38
308184	S_2	357,724.9042	3,457,110.4939	09:05:21	13:34:00
308184	S_3	360,777.5635	3,455,463.2075	13:47:15	14:43:05
308184	S_4	357,804.9576	3,457,007.4139	14:52:50	18:18:58
308184	S_5	351,658.3757	3,466,424.3204	19:07:18	23:07:23

Fig. 5 Trip length distribution of users in D2

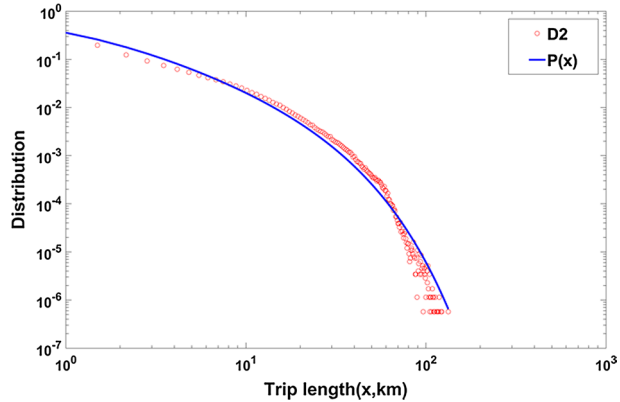


Table 4 The results of trip frequency on each research day

Method	11/19/2013	11/20/2013	11/21/2013	11/22/2013
Min_Circle	2.29	2.24	2.29	2.16
Asakura	2.31	2.27	2.31	2.17

distribution can also be well approximated by the truncated power-law proposed in Gonzalez et al. (2008). The main reason for the slightly different parameters is the built environment in different countries which is explained in Calabrese et al. (2013). The specific expression for the fitting distribution is $P(x) = (x + 0.98)^{-1.4} \exp(x/18)$ with the coefficient of determination (R^2) as 0.96.

To further validate that the proposed method is feasible, the average trip frequency per day of users in D2 on workdays are obtained using both our method and the method in Asakura and Hato (2004) (Asakura’s method). The average results on research days are respectively 2.25 and 2.27 using the equivalent spatial and temporal thresholds (700 m and 30 min in Asakura’s method), as shown in Table 4. According to the travel survey of residents in Shanghai (Shanghai City Comprehensive Transportation Planning Institute 2010), the average trip frequency is around 2.23.

The statistical significance is analyzed using t test for the results obtained by our method and Asakura’s method. There are three statistical tests that need to be conducted.

In the first test, one-sample t -test is used to compare the average trip frequency obtained by Asakura’s method and the result from the travel survey. The null hypothesis is there is

no significant difference between the average trip frequency obtained by Asakura's method and the result of the travel survey. The p -value ($p=0.367$) is larger than 0.05, which means the null hypothesis should be accepted when the significance level is 0.05.

In the second test, one-sample t -test is used to compare the average trip frequency obtained by our method and the result from the travel survey. The null hypothesis is there is no significant difference between the average trip frequency obtained by our method and the result of the travel survey. The p -value ($p=0.658$) is larger than 0.05, which means the null hypothesis should be accepted when the significance level is 0.05.

In the third test, paired-samples t -test is used to compare the results obtained by our method and Asakura's method. The null hypothesis is the average trip frequency obtained by Asakura's method is not significantly larger than the result of our method (*one-sided test*). The p -value ($p=0.008$) is less than 0.05, which means that the null hypothesis should be rejected when the significance level (α) is set as 0.05.

The first two tests prove that both methods are effective compared to the result of the travel survey. The result of the third test illustrates that the average trip frequency obtained by Asakura's method is significantly larger than the result of our method. Considering the average trip frequencies of our method, Asakura's method and the travel survey are respectively 2.25, 2.27 and 2.23, *the third test also indicates that the results of our method are significantly closer to the result of the travel survey than those from Asakura's method*. Based on all the three tests, the proposed method in this study is improved when using the result of the travel survey as reference.

Travel behavior of the hub passengers

As illustrated before, users of D1 are composed of incity passengers and external passengers who have appeared in the core area of the hub. To further evaluate the proposed method of extracting travel trajectories, the passengers' travel behavior of the Hongqiao hub is analyzed as an application. Several key travel characteristics of the two categories of passengers are compared. Moreover, productions and attractions are also studied for *the trips related to the core area of the hub* in this section. A concept is first introduced to analyze the importance of the transportation hub in passengers' daily travel.

The trips related to the core area of the hub (TRH) mean that the origin or destination points of the trips are located in the core area or the trips pass by the core area.

The travel characteristics studied in this paper mainly contain four categories: trip frequency, travel time, trip length, productions and attractions.

Trip frequency, *Travel time* and *Trip length* are three indexes characterizing the travel intensity of the passengers related to the transportation hub. The trip frequency can be used to quantify the general hub related travel demand, which can help the administrators to understand the operation situation of the hub and the traffic pressure of the surrounding transportation network. The departure time and the arrival time of the passengers can directly display the specific changes of the travel demand over the day. The trip length is an index related to the service range of the transportation hub containing different travel modes. *Productions and Attractions* provide a visualized way to characterize the service range of the hub and the travel demand in different areas of the city. All the indexes are useful for the transportation planning and operation related to the transportation hub.

Trip frequency

This index is used to study the number of trips per day for the users. The average trip frequency of the incity passengers in D1 is 2.43 (2.52 on workdays, 2.34 on weekends), slightly higher than the result of users in D2. However, a much lower result is obtained for the external passengers in D1, which is around 1.75 (1.82 on workdays, 1.68 on weekends). To find the reason for the results, the distributions of the trip frequency are analyzed for the incity and external passengers in D1, as shown in Fig. 6. Figure 6a displays that the curve of the external passengers peaks at $x=1$ instead of $x=2$ which is the peak point in the curve of the incity passengers on workdays. It can be explained as that the external passengers tend to pass by the core area of the hub to enter the city or exit the city without additional trips, while the incity passengers need to leave and return the places where they live. Moreover, the curve of the incity passengers declines slower than that of the external passengers. On weekends, the situation is similar despite that the values at $x=1$ and $x=2$ are close in the curve of the incity passengers (seeing Fig. 6b), which illustrates that the incity passengers in D1 tend to travel less on weekends.

The numbers of TRH are further explored. The average results of the incity and external passengers are respectively 1.23 (1.25 on workdays, 1.21 on weekends) and 1.18 (1.20 on workdays, 1.15 on weekends). It can be seen from Fig. 6c, d that most passengers travel one or two times through the core area of the hub per day. The proportions of the incity passengers are no less than that of the external passengers when $2 \leq x \leq 6$, which is the

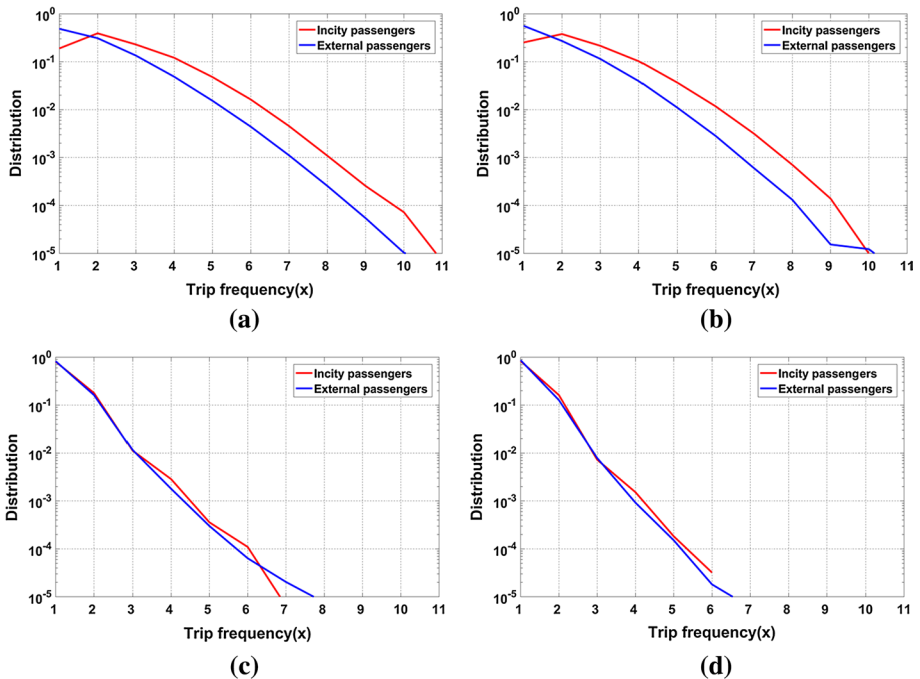


Fig. 6 Trip frequency distributions of passengers in D1. **a** Distributions of trips on workdays. **b** Distributions of trips on weekends. **c** Distributions of TRH on workdays. **d** Distributions of TRH on weekends

reason why the incity passengers have a higher TRH frequency. In fact, the distributions of the two curves are quite similar.

Travel time

The distributions of the departure time and arrival time of TRH are analyzed for different kinds of passengers. To accomplish the target, the time of a day is divided into 48 equal intervals (i.e., 30 min in each interval) to calculate the proportions of users in each time interval. The first records and the last records of users on each research day are set as staying, as illustrated in the trips generation method. Since lots of users have their first records in 0:00–1:00 and the last records in 23:00–24:00, the proportions in these two time periods have less reliability. Figure 7 displays the distributions from 1:00 to 23:00. On workdays, the distributions of the incity passengers present obvious morning and evening tides. The rush hours in the arrival time curve appear later than those in the departure time curve (seeing Fig. 7a, b). For example, the rush hours in the morning are around 6:30–8:30 in the arrival time curve, while the rush hours in the morning are around 8:00–10:00 in the departure time curve. The reason is that the core area of the hub serves for the commuting trips of part of the incity passengers. It can also be found that the phenomenon of morning and evening tides disappears on weekends in Fig. 7c, d, which is also an evidence. In Fig. 7d, the curve of the incity passengers peaks at around 17:30–18:30 which is the time

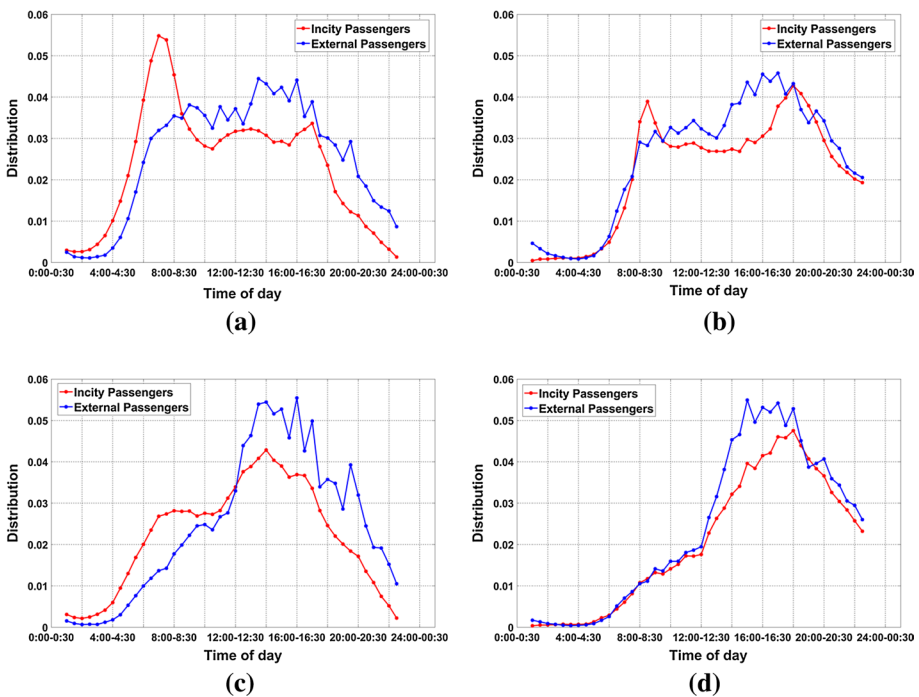


Fig. 7 Departure time and arrival time distributions of passengers' TRH in D1. **a** Departure time distributions on workdays. **b** Arrival time distributions on workdays. **c** Departure time distributions on weekends. **d** Arrival time distributions on weekends

interval when the incity passengers return home or begin nightlife on weekends. The distributions of the external passengers present a trend as rising-fluctuating-declining on both workdays and weekends, which is mainly influenced by the time schedules of the external traffic modes (aviation and railway).

Trip length

Trip length is a crucial index to analyze the travel behavior of human. It has been stated that the distribution of the trip length of users in D2 can be well approximated by the truncated power-law proposed in Gonzalez et al. (2008). To analyze the travel behavior of passengers, the length distributions of TRH are further studied for users in D1.

As shown in Fig. 8a, the distributions of both the incity passengers and the external passengers on workdays still have fat-tailed feature. However, only the curve of the incity passengers can be well approximated by the truncated power-law $P(x) = (x + 358)^{-0.45} \exp(x/14.7)$ with $R^2 = 0.96$. It can be seen that the curve of the external passengers increases during [4, 9 km] and fluctuates during [10, 50 km], which goes against the truncated power-law. The distribution of the incity passengers' TRH is consistent with the residents' travel patterns: most of the time they travel only over short distances, whereas occasionally they take longer trips. However, the curve of the external passengers illustrates that the probability they travel long distances [5, 50 km] to their destinations is not much lower than that of the short distances [1, 3 km], when utilizing the transportation modes of the hub. The phenomenon is mainly because that the external passengers usually travel between the hub and the boundary of the city when entering or exiting the city using ground traffic. The external passengers don't have home based trips in the city, which also leads to the irregularity. The situation on weekends is quite similar to that on workdays (Fig. 8b).

Productions and attractions of TRH

The indexes of travel characteristics have been analyzed to compare the travel behavior of the incity passengers and the external passengers. Productions and attractions should be further explored to study the influence of the hub on passenger's travel behavior. For the following analysis, a net with 20×20 grid cells is created to cover the area of the city, as shown in

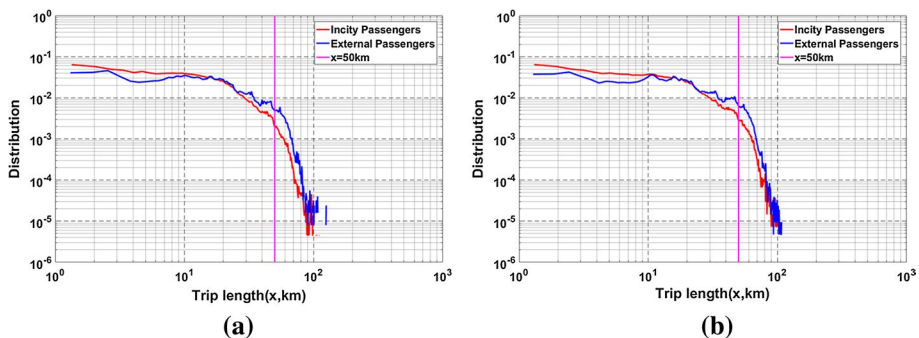


Fig. 8 Trip length distributions of passengers' TRH in D1. a TRH on workdays. b TRH on weekends

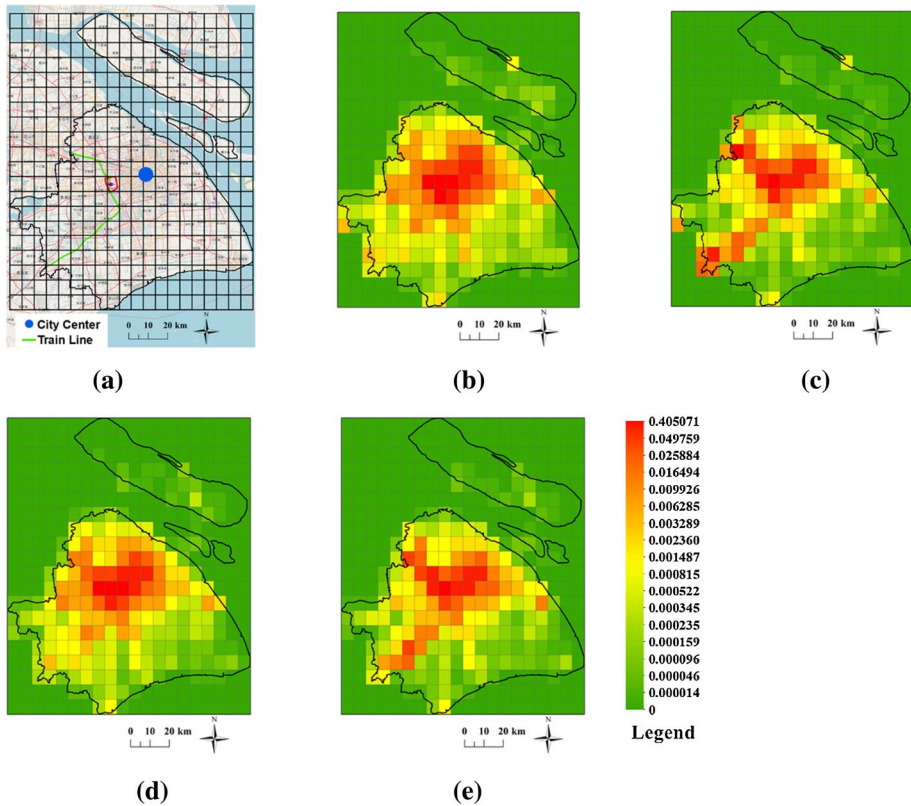


Fig. 9 Productions and attractions of passengers' TRH in D1. **a** A net with 20×20 grid cells covering the city area. **b** Productions of incity passengers' TRH. **c** Productions of external passengers' TRH. **d** Attractions of incity passengers' TRH. **e** Attractions of external passengers' TRH

Fig. 9a. Average proportions of productions and attractions of TRH are counted for each cell to display the distributions of origins and destinations in the city on the research days.

From Fig. 9b, d, it can be found that the productions and attractions of the incity passengers spread over the city with the core area of the hub as the center. Most productions and attractions appear around the hub area, whereas the distributions spread all over the city. The finding conforms to the fat-tailed distribution of the trip length of incity passengers: they usually travel short distances, while they occasionally travel far from the hub with the aid of urban transportation modes. There is also a special phenomenon that the decrease of the distributions is quite slow from the hub area to the city center area, which states that the city center acts a vital role in incity passengers' daily travel. The core area of the hub establishes a connection between the city center and the origins/destinations of the incity passengers.

However, the situation becomes different when it comes to the distributions of the external passengers, as shown in Fig. 9c, e. Although the proportions of productions and attractions are still high around the hub area, the phenomenon that the distributions of the origins/destinations are focused around the city center seems to be more apparent. Moreover, the productions and attractions near the city boundary become higher, especially the

areas on the top left corner and the bottom left corner. The reason should be that these two areas cover the junctions of the city boundary and the train line. The railway is one of the most important external transportation modes for passengers. It can be further found that the productions and attractions all tend to be higher in the areas along the train line. The phenomenon is reasonable because the sampling time interval of the mobile phone data is uncertain, which may lead to that the first or last record of a user is not generated around the junctions. The Euclidean distances between the core area of the hub and the junctions are measured, which are respectively around 27 and 54 km. The results explain the reason why the distributions of the trip length of the external passengers fluctuates until around 50 km.

Correlation analysis of passengers’ travel trajectories

As illustrated in the definitions of the study areas, the comprehensive area of the Hongqiao hub is a subsidiary area with the functions of residence and commerce. It is like a new town built in the city based on the passenger attraction of the core area. Such situation is pretty common in China with the development of the high-speed railway. However, it is difficult to estimate whether these areas are useful for passengers or whether passengers choose to stay in these areas for the services. In this section, the correlation analysis method is applied to analyze the travel trajectories to understand the availability of such new towns, which is helpful to evaluate the development status of the subsidiary area surrounding the hub.

The extracted travel trajectories of the passengers are composed of the trip stops. For a passenger, the travel trajectory on a typical day is defined as a *Transaction* in the correlation analysis. Thus, the transaction database *T* contains the travel trajectories of all the hub passengers. The *Item* is defined as a trip stop in the travel trajectories which is classified according to the located areas. The specific definitions are listed as follow:

- Item 1 (com)*—the trip stop located in the comprehensive area of the hub,
- Item 2 (core)*—the trip stop located in the core area of the hub,
- Item 3 (city)*—the trip stop located in the city area excluding the transportation hub area,
- Item 4 (out_city)*— the trip stop located outside the city area, which is a category of virtual trip stop to complete the external passengers’ trajectories when they are not in the city.

The association rule $Item\ i \Rightarrow Item\ j\ (1 \leq i, j \leq 4)$ means that the passenger stops in both areas expressed in *Item i* and *Item j* on a same day. Several important concepts in the correlation analysis are expressed as follow:

$$\text{support}(Item_i) = \frac{\text{count}(Item_i \subseteq T)}{|T|} \tag{11}$$

$$\text{support}(Item_i \Rightarrow Item_j) = \frac{\text{count}(Item_i \cup Item_j)}{|T|} \tag{12}$$

$$\text{confidence}(Item_i \Rightarrow Item_j) = \frac{\text{support}(Item_i \Rightarrow Item_j)}{\text{support}(Item_i)} \tag{13}$$

$$\text{lift}(Item_i \Rightarrow Item_j) = \frac{\text{confidence}(Item_i \Rightarrow Item_j)}{\text{support}(Item_j)} \quad (14)$$

where $|T|$ is the number of all transactions in T , $\text{count}(Item_i \subseteq T)$ is the number of transactions containing $Item_i$, $\text{count}(Item_i \cup Item_j)$ is the number of transactions containing both $Item_i$ and $Item_j$.

To research *the availability of the comprehensive area*, the association rules related to *Item 1* need to be analyzed. In this paper, *the availability is defined as* whether passengers tend to stop in the comprehensive area and whether such trip stops are related to the other categories of stops in passengers' daily travel.

By applying the correlation analysis method to the passengers' travel trajectories extracted from dataset D1, the availability of the comprehensive area is studied. Three association rules are analyzed in this study including: *Item 2 (core) ⇒ Item 1 (com)*, *Item 4 (out_city) ⇒ Item 1 (com)*, *Item 4 (out_city) ⇒ Item 2 (core)*, as shown in Table 5. Several findings can be summarized from the results:

1. The comprehensive area of Hongqiao hub plays a relatively important role in passengers' daily travel. The percent of passengers staying in the hub area (the comprehensive area or the core area) during their daily travel equals $\text{Support}(com) + \text{Support}(core) - \text{Support}(core \Rightarrow com) = 0.448$, explaining that almost half of passengers tend to stay in the hub area. Among these passengers, the comprehensive area is a more preferred choice with the result that $\text{Support}(com)$ is more than double of $\text{Support}(core)$. Moreover, $\text{Lift}(core \Rightarrow com) < 1$ illustrates that *com* and *core* tend to be mutually exclusive, i.e., most passengers stay in only one area of the hub (either the comprehensive area or the core area).
2. The availability of the comprehensive area is balanced for the incity passengers and the external passengers. It can be found that *out_city* and *com* tend to be mutually independent from the result $\text{Lift}(out_city \Rightarrow com)$ is close to 1. The association rule *out_city ⇒ com* is unable to be established, although $\text{Support}(out_city \Rightarrow com)$ is considerable. Thus, passengers having trip stops located outside the city area (the external passengers) is not related to having trip stops located in the comprehensive area.
3. More proportion of the external passengers tend to stay in the core area than that of the incity passengers. $\text{Lift}(out_city \Rightarrow core) = 1.333 > 1$ explains that the occurrence probability of *core* is lifted when *out_city* occurs, i.e., the proportion of the external passengers staying in the core area is higher than that of the incity passengers. It is reasonable because the transfer waiting times of the incity transportation modes (bus, metro, taxi, etc.) tend to be shorter than those of the external transportation modes (train, aviation, etc.).

Table 5 The results of correlation analysis for passengers' travel trajectories

	<i>com</i>	<i>core</i>	<i>out_city</i>	<i>core ⇒ com</i>	<i>out_city ⇒ com</i>	<i>out_city ⇒ core</i>
Support	0.344	0.117	0.456	0.013	0.164	0.071
Confidence	–	–	–	0.111	0.360	0.156
Lift	–	–	–	0.323	1.047	1.333

Conclusion

In this paper, a complete method is presented to study the travel behavior of the hub passengers in the city area using the mobile phone data. A detailed application based on the Hongqiao transportation hub in Shanghai is conducted to validate the proposed method. The contributions are mainly focused on four major aspects.

1. This study provides a convenient and feasible way to study the hub passengers' travel behavior with the mobile phone data. The case study results of the Hongqiao hub further testify that the method is effective and the mobile phone data is an excellent data source to support the hub passengers' travel behavior research.
2. This study proposes a systematic and novel approach to collect the trip information from the mobile phone data based on Asakura's work (2004). A more detailed method is also introduced to collect the departure/arrival time of each trip. Since the trip information is essential in the travel behavior study, the approach is useful for the urban planners and the transportation researchers. The average trip frequency of users in D2 is compared to the travel survey result, which has validated the method.
3. The analyses of the travel behavior indexes can support the transportation administrators to operate the hub and optimize the related transportation planning. For example, the travel time distributions of the incity passengers show an obvious tide pattern on workdays. Therefore, the incity travel modes, such as the metro lines, can reduce the departure intervals in the rush hours to transfer the passengers more efficiently. Besides, the productions and attractions of TRH display the passenger connections between the areas inside the city and the transportation hub. More attentions need to be paid to the areas having obvious passenger connections with the hub, especially the suburb areas. Travel modes directly connecting the areas and the hub can be designed to facilitate the passengers.
4. The correlation analysis of the passengers' travel trajectories can be used to evaluate the development status of the subsidiary area surrounding the hub. The administrators can adjust the functions of such new towns based on the analysis results. If there is a low percentage of passengers choosing to stay in the new towns, more investment may be needed to build more shopping malls, hotels, restaurants, etc. The information can also be useful to estimate whether the new towns are necessary for the designed hubs in other cities.

Besides these contributions, some findings about the passengers' travel behavior related to the Hongqiao hub should also be illustrated.

1. The average trip frequency of the incity passengers is higher than that of the external passengers.
2. The travel time distributions of the incity passengers show an obvious tide pattern on workdays, while the distributions of the external passengers don't present this pattern.
3. Different from the distributions of the TRH length of the incity passengers, the distributions of the external passengers don't conform to the general human travel rule, i.e., the truncated power-law proposed in Gonzalez et al. (2008).
4. The productions and attractions of the external passengers are higher in the areas along the train line, which verifies the last finding and explains the reason in a visualized way.

5. The comprehensive area of the Hongqiao hub plays a relatively important role in passengers' daily travel according to the results of the availability analysis.

Our results show that it is an effective way to apply the mobile phone data to study the interactive relationship between the operation of the transportation hub and the passengers' travel behavior. Although the mobile phone data has plenty of advantages in the macroscopic travel behavior research, there are still some weaknesses.

1. The location accuracy of the mobile phone data is at the cellular tower level which is lower than that of the GPS data. Therefore, it is difficult to study the microscopic travel behavior, such as travel route choice, travel mode, etc., using the mobile phone data.
2. The check-in data from the social network can provide the subscribers' activity information besides the location information, which is also a weakness of the mobile phone data. It is difficult to use the mobile phone data to study the trip purposes of the users.
3. Due to the privacy policies, the personal information (such as age, income, profession, etc.) can't be collected from the mobile phone data. The personal information can only be collected from the methods with volunteers such as the travel survey, the GPS data, etc.

In the future, the additional data from other sources can work as complementary samples to study the hub passengers' travel behavior more comprehensively, such as the individual travel patterns, the trip purposes of the passengers, the prediction of the travel trajectories, etc. Moreover, the passengers' travel demand patterns related to the hub can also be further explored.

Acknowledgements This study is partially supported by the Information Technology Research Project of Ministry of Transport of China (No. 2015364X16030) and the National Natural Science Foundation of China (No. 61620106002). The support provided by China Scholarship Council (CSC) during a visit of G. Zhong to UW-Madison is acknowledged.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. C* **58**, 240–250 (2015)
- Asakura, Y., Hato, E.: Tracking survey for individual travel behaviour using mobile communication instruments. *Transp. Res. C* **12**(3), 273–291 (2004)
- Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating origin–destination flows using mobile phone location data. *IEEE Pervasive Comput.* **10**(4), 36–44 (2011)
- Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transp. Res. C* **26**, 301–313 (2013)
- Chen, C., Bian, L., Ma, J.: From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transp. Res. C* **46**, 326–337 (2014)
- Cheung, C.Y., Lam, W.H.: Pedestrian route choices between escalator and stairway in MTR stations. *J. Transp. Eng.* **124**(3), 277–285 (1998)

- De Berg, M., Van Kreveld, M., Overmars, M., Schwarzkopf, O., Overmars, M.H.: *Computational Geometry: Algorithms and Applications*, 3rd edn, pp. 86–89. Springer, Berlin (2008)
- Fang, J., Xue, M., Qiu, T.Z. (2014): Anonymous cellphone-based large-scale origin–destination data collection: case studies in China. In: *Proceedings of Transportation Research Board 93rd Annual Meeting*, Washington, D.C., No. 14-1567 (2014)
- Frias-Martinez, V., Soguero, C., Frias-Martinez, E.: Estimation of urban commuting patterns using cellphone network data. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, Beijing, China, pp. 9–16 (2012)
- Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779–782 (2008)
- He, S., Cheng, Y., Ding, F., Zhang, J., Ran, B.: Extended Kalman filter-based freeway traffic state estimation using cellphone activity data. In: *Proceedings of Transportation Research Board 95th Annual Meeting*, Washington, D.C., No. 16-3728 (2016)
- Jiang, S., Ferreira, J., González, M.C.: Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. *IEEE Trans. Big Data* **3**(2), 208–219 (2017)
- King, D., Srikukenthiran, S., Shalaby, A.: Using simulation to analyze crowd congestion and mitigation at Canadian subway interchanges: case of Bloor-Yonge Station, Toronto, Ontario. *Transp. Res. Rec.* **2417**, 27–36 (2014)
- Li, H.: Using mobile phone data to analyze origin–destination travel flow dynamics for city of Pasadena, CA and surrounding area. In: *Proceedings of Transportation Research Board 94th Annual Meeting*, Washington, D.C., No. 15-0804 (2015)
- Pan, C., Lu, J., Di, S., Ran, B.: Cellular-based data-extracting method for trip distribution. *Transp. Res. Rec.* **1945**, 33–39 (2006)
- Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C.: Activity-aware map: identifying human daily activity pattern using mobile phone data. In: *International Workshop on Human Behavior Understanding*, Istanbul, Turkey, pp. 14–25 (2010)
- Rokib, S.A., Karim, M.A., Qiu, T.Z., Kim, A.: Origin–destination trip estimation from anonymous cell phone and foursquare data. In: *Proceedings of Transportation Research Board 94th Annual Meeting*, Washington, D.C., No. 15-2379 (2015)
- Sagl, G., Delmelle, E., Delmelle, E.: Mapping collective human activity in an urban environment based on mobile phone data. *Cartogr. Geogr. Inform. Sci.* **41**(3), 272–285 (2014)
- Shanghai City Comprehensive Transportation Planning Institute. The fourth travel survey of residents in Shanghai. Shanghai (2010)
- Shanghai Hongqiao Central Business District. Passenger flow information of Hongqiao hub in 2013. http://www.shhqcbd.gov.cn/html/shhq/shhq_2013/Info/Detail_6403.htm (2013). Accessed 15 May 2015
- Shen, Y., Kwan, M.P., Chai, Y.: Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *J. Transp. Geogr.* **32**, 1–11 (2013)
- Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. *Science* **327**(5968), 1018–1021 (2010)
- Srikukenthiran, S., Fisher, D., Shalaby, A., King, D.: Pedestrian route choice of vertical facilities in subway stations. *Transp. Res. Rec.* **2351**, 115–123 (2013)
- Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C.: Discovering urban activity patterns in cell phone data. *Transportation* **42**(4), 597–623 (2015)
- Xu, S., Freund, R.M., Sun, J.: Solution methodologies for the smallest enclosing circle problem. *Comput. Optim. Appl.* **25**, 283–292 (2003)
- Zhang, Q., Han, B., Li, D.: Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. *Transp. Res. C* **16**(5), 635–649 (2008)
- Zhang, Y., Qin, X., Dong, S., Ran, B. (2010): Daily OD matrix estimation using cellular probe data. In: *Proceedings of Transportation Research Board 89rd Annual Meeting*, Washington, D.C., No. 10-2472 (2010)
- Zhang, J., He, S., Wang, W., Zhan, F.: Accuracy analysis of freeway traffic speed estimation based on the integration of cellular probe system and loop detectors. *J. Intell. Transp. Syst.* **19**(4), 411–426 (2015)
- Zhong, G., Wan, X., Zhang, J., Yin, T., Ran, B.: Characterizing passenger flow for a transportation hub based on mobile phone data. *IEEE T. Intell. Transp.* **18**(6), 1507–1518 (2017)

Gang Zhong is a Ph.D. candidate at the Research Center for Internet of Mobility, Southeast University. His research is mainly focused on mining information from location data, such as mobile phone data, to analyze travel behavior and support urban planning.

Tingting Yin holds a master degree in transportation engineering from Southeast University, China. She is now working in Jiangsu Expressway Company Limited. Her research interests now include transportation systems planning and traffic control of the expressway.

Jian Zhang holds a Ph.D. degree in transportation engineering from Southeast University, China. He is the vice director of the Research Center for Internet of Mobility, Southeast University, where he focuses on transportation application of mobile phone data and connected vehicles researching.

Shanglu He holds a Ph.D. degree in transportation engineering from Southeast University, China. She is now working as a lecturer in School of Automation, Nanjing University of Science and Technology. Her work is mainly focused on the area of data fusion, especially about sensing freeway traffic state with multi-source data including cellular handoff data, microwave data, etc.

Bin Ran is a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, USA, and the Director of the Research Center for Internet of Mobility, Southeast University, China. Currently, his research interests cover transportation big data analysis, transportation systems planning and connected and automated vehicles.