Characterizing Passenger Flow for a Transportation Hub Based on Mobile Phone Data

Gang Zhong, Xia Wan, Jian Zhang, *Member, IEEE*, Tingting Yin, and Bin Ran

***Abstract*— As the vital node of a passenger transportation network, the transportation hub is the connection between multiple travel modes and the important port for the massive passenger flow to enter into or exit from a city area. Transportation operators need to understand the passenger flow pattern for hub management, transportation planning, and so on. However, it is difficult to use traditional methods, such as video detection, to provide such information. With the increasing number of mobile phone users, mobile phone data have shown remarkable potential in detecting the transportation information with high sampling coverage and low cost. This paper utilizes the mobile phone data to characterize the passenger flow of the Hongqiao transportation hub located in Shanghai, China. First, a temporal-spatial clustering method is proposed to identify the passenger active area of the Hongqiao hub in the wireless communication space. Second, a classification process is presented to extract different types of passengers in this transportation hub. Subsequently, the access characteristics of passengers in the city are studied for various time intervals. The results further verify the potential of using mobile phone data to monitor and characterize passenger flow related to the transportation hubs.**

***Index Terms*— Base station, mobile phone data, passenger transportation hub, temporal-spatial clustering, wireless communication.**

I. Introduction

PASSENGER transportation hub is the place where passengers transfer between different transport modes such as aviation, railway, highway, public transportation, etc. A comprehensive passenger transportation hub can become a distribution center with massive passenger flow. Thus, it is highly demanded to analyze the characteristics of the passenger flow for both transportation planning and management. Traditional methods, such as the video detection and the loop detectors, are focused on the traffic flows in critical sections.

However, these methods can't accurately detect the total number of passengers in the transportation hub. And it is hard to get the real-time information such as how many of them come from other cities. Moreover, the distribution of passenger flow in the city is completely out of the detection coverage of traditional methods. It is difficult to discover the travel behavior of passengers in the city. Lack of such information poses a challenge for transportation operators to monitor the operation state of the hub and optimize the transportation management.

In recent years, location data from social networks, mobile phones, GPS, etc. has been applied to characterize human activity [1]–[5]. Among these data sources, the application of mobile phone data has shown more potential for the extensive popularity of mobile phones, even in the developing countries. The sampling coverage of mobile phone data is more remarkable for research compared to that of other data sources. Furthermore, as involuntary user generated data, mobile phone data can provide relatively unbiased sample across the society [6]. Since the user ID is encrypted, it's no need to concern the privacy issue.

There are still some limitations of mobile phone data, such as lacking of detailed user information (sex, age, income, etc.). In spite of the limitations, mobile phone data has shown the superiority for researching the overall situation of human activity compared to the traditional methods. In order to further prove it, this study utilizes the advantages of mobile phone data to characterize the passenger flow related to a transportation hub of a megacity-Shanghai. First, the passenger active area of the transportation hub in the wireless communication space is identified based on the temporal-spatial clustering method we proposed. Second, passengers are distinguished from the staff members working in the transportation hub and classified into certain categories by the proposed classification method. The access characteristics of passengers in the city are analyzed at last. The remainder of the paper is organized as follows. In the second section, literature review about analyzing human activity based on mobile phone data is presented. The third section is the description of the mobile phone dataset and the transportation hub we studied. In the fourth section, the temporal-spatial clustering method is proposed to identify the passenger active area. The characteristics of passenger flow are analyzed in the fifth section. The conclusion and future work are addressed in the sixth section.

II. Literature Review

Previous research has sought to understand the temporal dynamics of urban activity and periodical human travel patterns by employing mobile phone data.

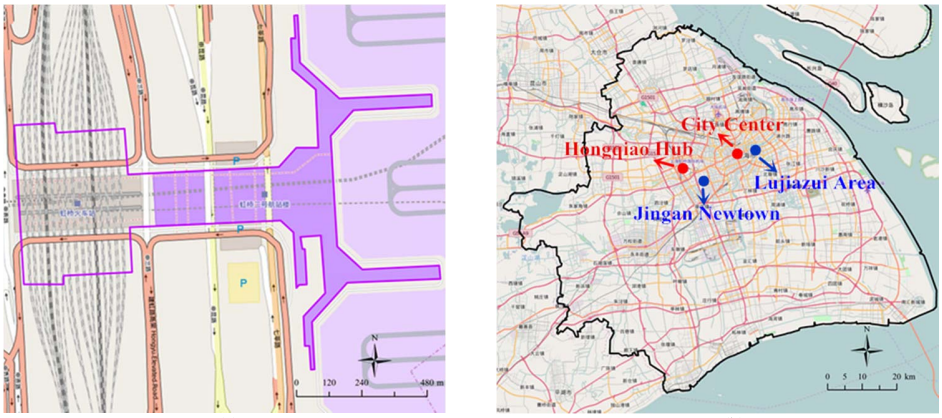
Manuscript received November 14, 2015; revised July 6, 2016; accepted September 6, 2016. This work was supported in part by the National Key Basic Research Development Program of China under Grant 2012CB725405 and in part by the Science and Technology Demonstration Project of Ministry of Transport of China under Grant 2015364X16030. The Associate Editor for this paper was P. Wang.

G. Zhong, J. Zhang, T. Yin, and B. Ran are with the School of Transportation, Southeast University, Nanjing 210096, China (e-mail: anhuizhonggang@126.com; jianzhang@seu.edu.cn; yttwen@163.com; bran@seu.edu.cn).

X. Wan is with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: wan5@wisc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2607760



(a) (b)

Fig. 1. Hongqiao transportation hub. (a) Geographic coverage of the hub. (b) Location of the hub in Shanghai.

With regard to urban activity, the concept of "mobile landscape" has been proposed for the graphic representation of human activity intensity using mobile phone data. Ratti *et al.* [7] utilized the mobile phone activity in the geographic cells to explore urban dynamics at different times in Milan. Becker *et al.* [8] further analyzed the distribution of residence area and modeled the typical flow of people between various parts of the city over time. Visual analytics approach was illustrated in the study of Sagl *et al.* [9] Besides the analysis of typical spatial-temporal human daily mobility patterns, detection of exceptional events by distinguishing visually outliers was also discussed in previous research. Krisp [10] made suggestions for the allocation of fire and rescue services based on the changing of population hotspots over time. In the research of Sagl *et al.* [6], Self-Organization Mapping (SOM) was proposed as an innovation to investigate the temporal dynamics of collective human activity. Voice calls, text messages, and total network traffic were chosen as input variables from the mobile communication profiles.

As for travel patterns, there are a number of studies focusing on utilizing mobile phone data to estimate Origin-Destination (OD) matrices. Iqbal *et al.* [11] propose a methodology to develop OD matrices using mobile phone call detail records and limited traffic counts. Based on the methodology, tower-to-tower OD matrices were converted to node-to-node OD matrices in the traffic network. Folak *et al.* [12] discussed how mobile phone data can be processed to inform a four-step transportation model. Rokib *et al.* [2] extracted OD matrices from mobile phone data and Foursquare data, and the results are compared to the existing OD matrix based on travel surveys.

Individual mobility patterns were also explored in some studies driven by mobile phone data. Calabrese *et al.* [13] used the average daily total trip length of mobile phone users to measure individual mobility. Gonzalez *et al.* [14] and Candia *et al.* [15] investigated patterns of calling activity at the individual level and explored the interplay between calling activity and mobility patterns. Wang *et al.* [16] focused on the spatial variability of individual location choice and the time of day effect was analyzed.

The studies mentioned above characterized human activity, including collective human and individual, across the general urban landscape. However, the analysis of human activity of specific regions like the transportation hub is rarely involved. Our study tries to fill the gap based on clustering and visualization methods in previous research, and we also try to prepare basic information to further study the travel behavior of different types of passengers.

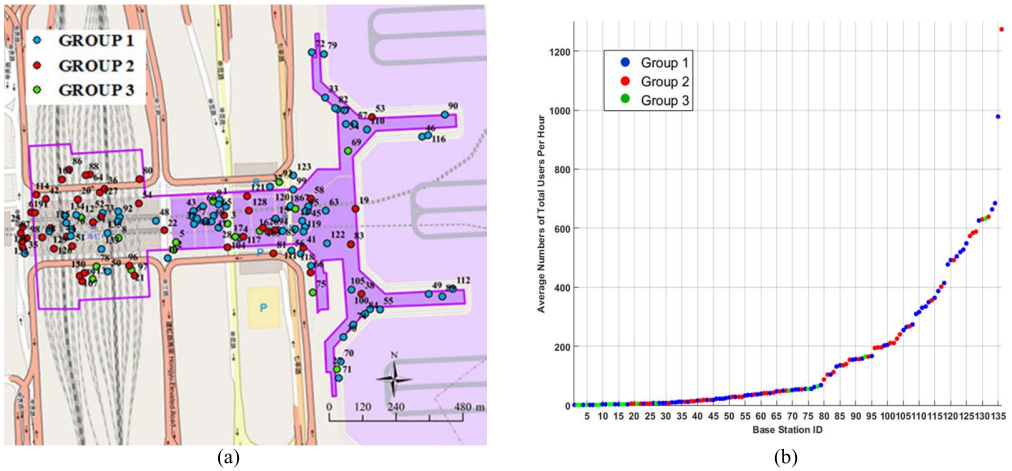
III. Data Description

A. Study Area

In this study, Hongqiao transportation hub in Shanghai of China is selected as the study area. It is an international integrated transportation hub that consists of an airport terminal, a railway station, a coach station, two subway stations, and several bus stations. According to the information on the website of Hongqiao central business district, the average daily number of inter-city travel passengers in 2013 is about 297,500, including 95,000 by aviation, 196,100 by railway, and 6,400 by highway [17]. As one of the largest transportation hubs in China, Hongqiao hub makes outstanding contributions to the economy of Yangtze River Delta besides transportation service. The study area is enclosed by the purple line in Fig.1 (a) which is about 50,000 square meters, and the location of Hongqiao hub in the city of Shanghai is shown in Fig.1 (b) which is about 12 kilometers away from the city center.

B. Mobile Phone Data

The mobile phone dataset used in this paper were collected from a major cellular service provider in China, including the data in the whole city of Shanghai on certain days in Nov. 2013. The total number of records is about 800 million for each day generated by about 18 million unique users. This study is concentrated on three continuous work days from Nov.27 to Nov.29. Each record is composed of a unique user ID, a Location Area Code (LAC), a Cell ID, a timestamp, and an event ID. The records can be generated by events like calling, texting, handover, location update, etc. And if there are no such events for a certain time, mobile phones will automatically connect to base stations to update their locations and



(a) (b)

Fig. 2. Groups of confirmed base stations on Nov.28. (a) Locations of all the confirmed base stations. (b) Average numbers of total users per hour of all the confirmed base stations.

generate records, which is called automatic location update. The time period to cause this location-updating event is one hour. And the latitude and longitude data of each base station located in Shanghai were also collected.

Data pre-processing is needed for the raw dataset to filter the erroneous data points including repeat data, ping-pong data and drift data. Repeat data is the type of totally same records which can be easily cleaned utilizing the database software. Ping-pong data is generated when a mobile phone switches between several adjacent base stations frequently in a short time, such as {A-B-A-B-A} or {A-B-C-B-A}. Whereas, drift data refers to the kind of data generated when a mobile phone jumps to a remote base station and reselects an adjacent base station after a short time. We set the temporal and spatial thresholds to filter the data records with the ping-pong pattern and the drift pattern, referring to the method proposed by Yang *et al.* [18].

IV. Identification of Passenger Active Area

A. Definitions

The physical boundary of the hub is pretty clear as the purple line in Fig.2 (a). However, the passenger active area of the hub in the wireless communication space, defined as '**hub communication area**', is not restricted by it. In consideration of the coverage of base stations, some base stations outside the physical boundary can also be connected by mobile phones of passengers in the hub. City expressways and park plots near the hub, where passengers are likely to appear, are covered by base stations too. All these base stations mainly serving for passengers compose the hub communication area.

An assumption is proposed in this paper that base stations inside the physical boundary and within 50 meters outside the physical boundary are all confirmed to be in the hub communication area. And these base stations compose '**the confirmed base stations set**'. There are 136 base stations in this set (shown in Fig.2 (a)). The average total number of unique user ID is around 197,950 on the chosen days collected from these base stations, and **the coefficient of variation (CV)** is 0.02 which illustrates that the number on each day is close.

An **active user** is a user whose mobile phone communicates with the base stations at least once in the investigating time period. For a specified base station, active users in a particular time period can be divided into three kinds.

- New users: active users who have records in the current time period and don't have records in last time period.
- Halt users: active users who have records in both current time period and last time period.
- Total users: all active users in the current time period.

The investigating time period used in this paper is one hour which is same as the time period for the automatic location update. The change of all three kinds of active user numbers over time can be expressed as time series. Due to the different cover ranges and user capacities of stations, the average numbers of total users per hour of different confirmed base stations are various during these days. The base stations are numbered order by these values in ascending sequence, and the specific numbers are displayed in Fig.2 (b).

Dynamic time warping distance (DTW) [19] is used to measure the difference between the time series of two base stations in this paper, which is called the **temporal distance** (TD) between them. As illustrated before, each base station has three groups of time series respectively representing the numbers of total users, halt users and new users over a day. TD between two base stations is divided into three categories for these three kinds of active users, which are T_TD, H_TD, and N_TD. Before TD is calculated, the time series of total users should be normalized into the range of 0 to 1 by considering the user capacity of each base station is not all the same. The time series of the other two kinds of active users also should be normalized based on their proportions to the total users.

The **spatial distance** (SD) between two base stations used in this study is the spherical distance between them on the earth. By knowing the latitude and longitude of base stations and the radius of the earth, SD can be calculated.

In the rest content of Section IV, we utilize the data on Nov.28 as examples, because the methods are similar for

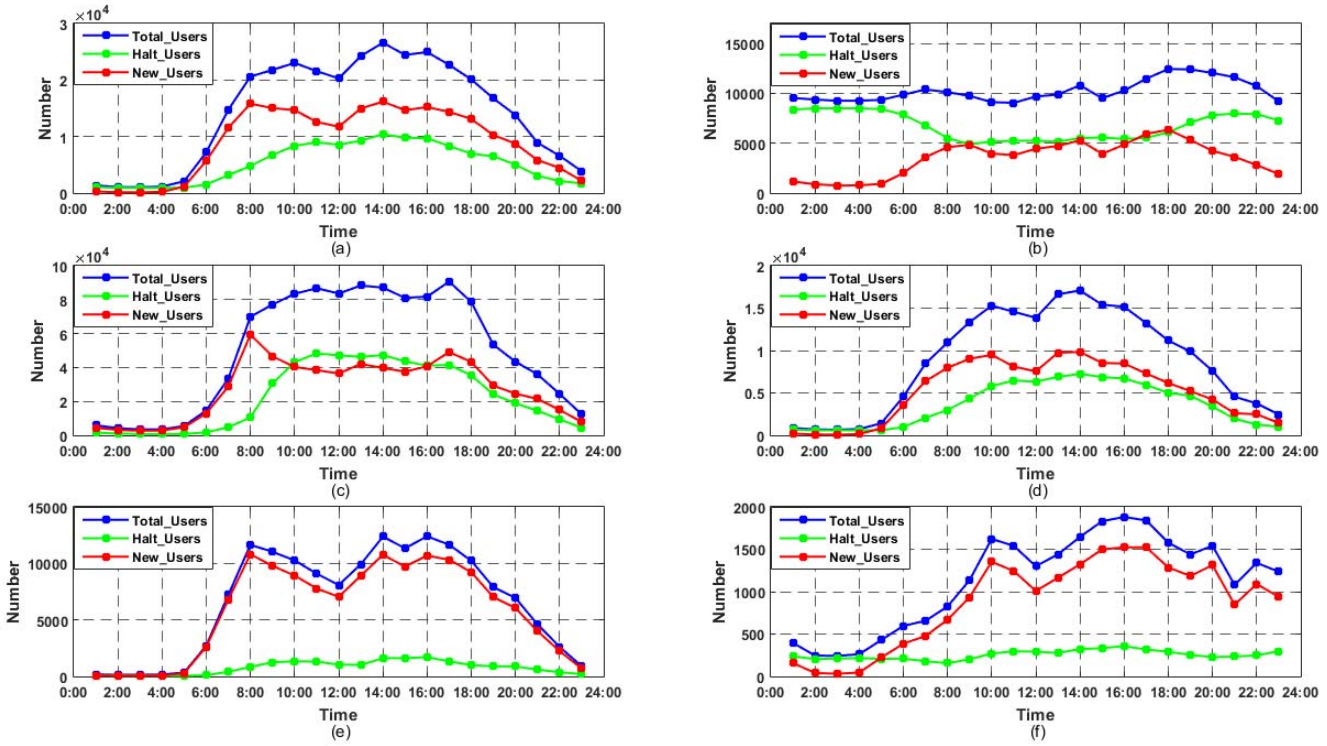


Fig. 3. Time series of active users. (a) Time series of active users of the whole Confirmed base stations set. (b) Time series of active users of the base stations chosen from the residential district. (c) Time series of active users of the base stations chosen from the commercial district. (d)-(f) Time series of active users of the base stations in Group1, Group2 and Group3.

researching on different days. When it comes to the identification results in Section IV-E, we refer to situations on all chosen days to minimize the data errors.

B. Characteristics of Active Users in Confirmed Area

In order to understand the characteristics of the active users in the hub communication area, we display the time series of user volumes of the confirmed base stations set on Nov.28 (as shown in Fig.3 (a)). The time series of whole base stations, respectively chosen from a residential district and a commercial district, are also displayed for comparison (Fig.3 (b, c)). We choose two famous districts in Shanghai to make sure that they are representative, which are Jingan Newtown and Lujiazui area. The land usage of each district is unitary so that the time series of the chosen base stations can demonstrate the characteristics of the districts. The specific locations of the two districts in Shanghai are shown as the blue points in Fig.1 (b).

Compared to the base stations in the residential district (Fig.3 (b)) and commercial district (Fig.3 (c)), a distinct characteristic of the confirmed base stations set is found that the proportion of halt users always maintains lower than that of new users. It implies that many users in the hub communication area tend to utilize the transfer function of the hub. Moreover, there are more halt users in the daytime than that at night in Fig.3 (a), which is different from the time series of the base station in the residential district. The reason is obvious that people more likely choose to stay at home instead of traveling in the nighttime. According to the timetables, there are usually no coaches, high-speed trains and flights scheduled at nighttime (0:00-6:00) in the hub. In Fig.3(c), the rush ours for new users are around 8:00-9:00 when people come to the commercial district for work. During the daytime (10:00-16:00), the numbers of halt users are higher than that of new users, because people tend to stay in their offices during this time period.

TABLE I
RESULTS OF THE TEMPORAL DISTANCES

District	T_TD(CV)	H_TD(CV)	N_TD(CV)
Hub	0.26(1.02)	0.45(0.73)	0.20(0.72)
Residence	1.21(0.45)	0.68(0.63)	0.97(1.07)
Commerce	0.93(0.64)	0.66(0.76)	0.57(1.29)

To further understand the difference, we calculate the temporal distances between the time series of the confirmed base stations set and the time series of each base station in the three districts (hub, residence, commerce). The average results and the corresponding CV are listed in TABLE I. It can be found that the average temporal distances for the confirmed base stations of the hub are much lower than that of the other two districts (especially T_TD and N_TD), which testifies that the base stations in the hub communication area have visible different characteristics. However, it is partial to simply use the time series of the confirmed base stations set to represent the characteristics, because the CV for the confirmed base stations are still large. We need to group the confirmed base stations according to the time-variance patterns of users.

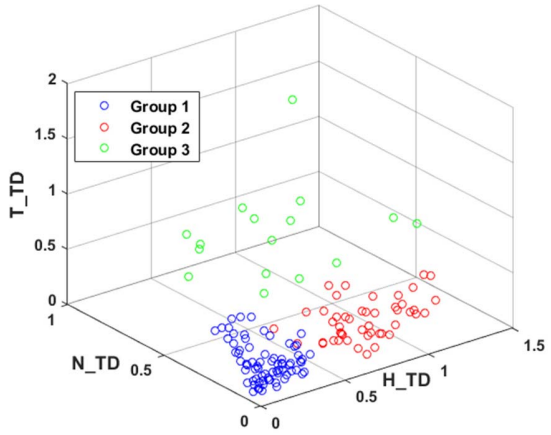


Fig. 4. The results of the grouping in 3-D space coordinate system.

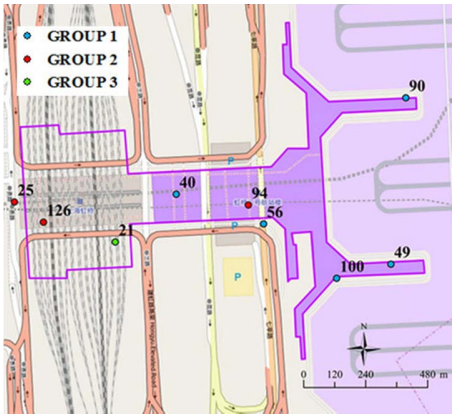


Fig. 5. The locations of the typical base stations.

C. Grouping the Confirmed Base Stations

We have analyzed the differences between base stations in different districts, utilizing the time series of the confirmed base stations set as references. With these references, the confirmed base stations are expressed by the temporal distances between the base stations and the whole set.

$$Base\ Station\ i = (T\_TD_i, H\_TD_i, N\_TD_i) \qquad (1)$$

Where T_TD_i, H_TD_i, N_TD_i are the three categories of temporal distances between base station i and the confirmed base stations set.

We take advantage of K-means method to group the base stations. Based on the comparison of the silhouette coefficient, the optimal number of the groups is 3. A three-dimensional space coordinate system is created to display the results of the grouping, as shown in Fig.4. The specific information of each group is shown in Fig.2.

The results illustrate that Group1 contains 73 base stations, Group 2 contains 47 base stations, and Group3 contains 16 base stations. The percentage of unique users in each group is respectively 58.8%, 63.2%, 10.2% compared to that of the whole set, and the total percentage is over 100% because a unique user in the whole set may appear in different groups.

The time series of the active users are displayed in Fig.3(d)-(f). It can be seen that the most significant difference between Group1 and Group2 is the proportions of the halt users. The proportions in Group1 are higher, which means there are also part of users tend to stay in the hub for a certain time to wait for flights or trains. Whereas, base stations in Group3 tend to have lower user capacities (seeing Fig.2 (b)), and the time series are more likely to be unstable compared to that of the confirmed base stations set (seeing Fig.4).

D. Temporal-Spatial Clustering Method for Active Area Identification

In order to estimate whether a base station which is more than 50 meters away from Hongqiao hub physical boundary belongs to the hub communication area, a temporal-spatial clustering method is proposed. Procedures of this method are shown as follow.

1) Step1 (Choosing Typical Base Stations): According to the previous analysis, it is partial to simply use the time series of the confirmed base stations set to represent the characteristics of the hub communication area. To decrease the bias, a set of typical base stations need to be composed before the **temporal-spatial distance** calculation step. In Section IV-C, we have grouped the confirmed base stations to make sure that the base stations in the same group have relatively similar time-variance patterns. So the typical base stations are chosen following three principles.

1. Typical base stations should be chosen from each group to represent different time-variance patterns of different groups.

2. The representative base stations of each group should be chosen according to the distances to the corresponding group center in ascending order, which is calculated by the k-means method in Section IV-C.

3. The number of chosen base stations from each group is proportional to the number of total base stations in the group.

Thus, we rank the base stations of each group in an ascending sort order according to the distances to their group centers and choose the first 5, 3, 1 base stations respectively from Group1, Group2, and Group3. The specific information of the typical base stations is listed as follow:

Group1 {40, 49, 56, 90, 100},
Group2 {25, 94, 126},
Group3 {21}.

The nine base stations are used to represent the hub communication area in this study, and they compose the typical base stations set {T}, as shown in Fig.5.

$$\{T\} = \{21, 25, 40, 49, 56, 90, 94, 100, 126\}$$

2) Step 2 (Establishment of Temporal-Spatial Distance Model): In order to quantify the difference between a checked base station and the hub communication area from both temporal and spatial aspects, a **temporal-spatial distance** (TSD) model is built. Typical base stations are applied to represent the hub communication area in this model. The temporal and spatial differences between a checked base station and each typical base station are transformed into TD and SD. TD is used as the principal component. And it is considered that the typical base stations nearer to the checked one will contribute

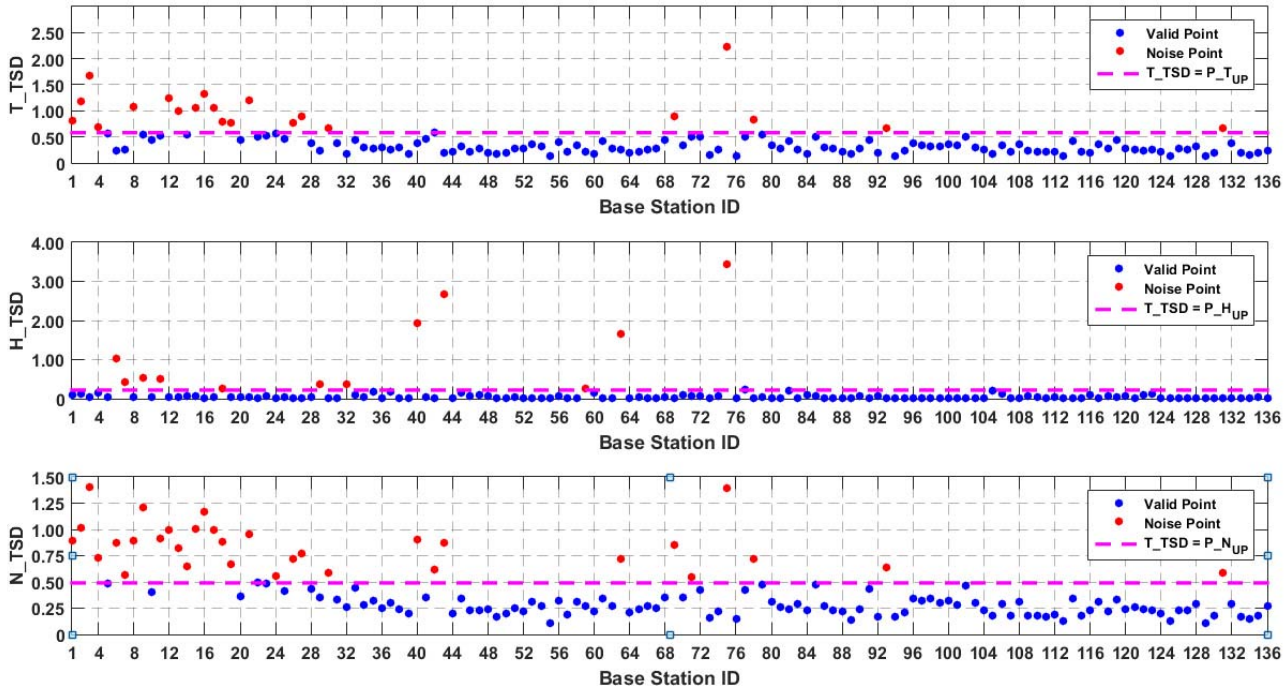


Fig. 6. TSD between each confirmed base station and the hub communication area.

more to the TSD than the further ones. So a simple weight method is taken to give each pair of TD a weight of the inverse of SD.

The final model is specified as:

$$TSD_i = \sum_j \left(\frac{1/SD_{ij}}{\Gamma} * TD_{ij} \right) \tag{2}$$

$$\Gamma = \sum_j \frac{1}{SD_{ij}} \tag{3}$$

If $i \in \{\mathrm{T}\}$, $j \in \{\mathrm{T}\} - \{\mathrm{i}\}$; if $i \notin \{\mathrm{T}\}$, $j \in \{\mathrm{T}\}$.

Where TSD_i is the TSD between base station i and the hub communication area; SD_{ij} is the SD between base station i and typical base station j; TD_{ij} is the TD between base station i and typical base station j; Γ is the sum of the weight.

3) Step 3 (Determination of the Clustering Criterion): The last step of the clustering method is to figure out whether a given TSD is beyond the threshold or not. Since the confirmed base stations are all considered to be in the hub communication area, the TSD between each confirmed base station and the hub communication area are calculated. The results are shown as the points in Fig.6. The average results of T_TSD, H_TSD and N_TSD respectively are 0.4159, 0.1389 and 0.4066. For each category of TSD, there are certain results far above these numbers. The reason for this phenomenon is that the time series of active users may present high volatility especially for base stations with relatively small user capacities. So it is not reasonable to set the maximum TSD results of the confirmed base stations as the marginal values. However, the threshold should make most of the confirmed base stations accord with the hypothesis that they belong to the hub communication area. In this paper, DBSCAN algorithm [20] is used to exclude the noise points from each category of TSD and group the remaining valid points as a valid cluster. We adopt the upper limit value of each valid cluster as the threshold for the corresponding category of TSD, which is shown as the red line in Fig.6. For base stations which are more than 50 meters away from the hub area, the clustering criterion is illustrated as:

$$T\_TSD \leq P\_T_{UP} \tag{4}$$

$$H\_TSD \leq P\_H_{UP} \tag{5}$$

$$N\_TSD \leq P\_N_{UP} \tag{6}$$

Where $P\_T_{UP}/P\_H_{UP}/P\_N_{UP}$ is the upper limit value of the valid cluster of *T_TSD/H_TSD/N_TSD* test results for the confirmed base stations.

For base stations that don't belong to the confirmed base stations set, satisfying this criterion means they are also in the hub communication area.

TABLE II
RESULTS OF THE TEST BASE STATIONS

Segment	Distance Range	Included Number	Included Percentage	Excluded Number	Excluded Percentage
1	50m-150m	23	85.19%	4	14.81%
2	150m-250m	13	68.42%	6	31.58%
3	250m-450m	8	53.33%	7	46.67%
4	450m-650m	10	58.82%	7	41.18%
5	650m-850m	3	30.00%	7	70.00%
6	850m-1050m	0	0.00%	11	100.00%
Total	50m-1050m	57	57.58%	42	42.42%

E. Passenger Active Area Identification Results

In consideration of the actual coverage of the base stations, the base stations within 50 to 1050m distance range away from the transportation hub physical boundary are tested using

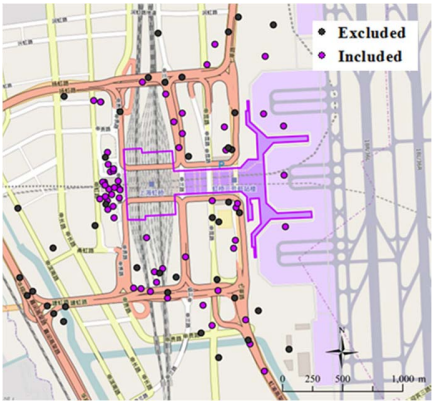


Fig. 7. Locations and test results of the test base stations.

our proposed clustering method. The range is divided into 6 segments for observing the change of the results with the distance increasing. We test the base stations on the three chosen days, and a base station is identified as 'Included' only when it get the 'Included' result on no less than two days. The test results are displayed in Table II. Numbers of 'Included' and 'Excluded' base stations are recorded, followed by their percentages in respective segments. Locations of the base stations are visualized in Fig.7. The passenger active area in the wireless communication space consists of the confirmed base station set and the test base stations whose results are 'included'.

In Table II, the results indicate that, in general, the probability for a base station belonging to the hub communication area decreases as the distance goes up. The randomness of coverage for base stations may cause fluctuations without affecting the overall trend. The results suggest that the distance range of the hub communication area is around 850 meters way from its physical boundary.

Different with the spatial clustering method in the previous study [21], the results for two base stations with close spatial distance may be different which is because the coverages of them are different. One of them can be connected by mobile phones in the hub, but the coverage of the other one may be not sufficient. For some micro base stations, the communication radius is just about 50-100 meters.

V. Characterizing the Passenger Flow of the Transportation Hub

A. Mobile Phone Users Classification for the Transportation Hub

In the hub communication area, the average number of unique active users on the chosen days is around 396,674, and the CV is 0.014. All the active users are assumed to represent people who have appeared in the transportation hub. According to the market share collected from the cellular service provider, the average number of people can be estimated, which is about 630,642.

People in the transportation hub can be simply classified as staff members and passengers. In order to understand the characteristics of passenger flow clearly, passengers need further classification.

- Incity passengers: people who travel inside Shanghai.
- External passengers: people who travel between Shanghai and other cities.
  - Arrival passengers: people who come from other cities and stay in Shanghai that day.
  - Departure passengers: people who head for other cities and don't return that day.
  - Arrival and Departure (AD) passengers: people who come from other cities and leave Shanghai on the same day.
    - Pass by (PB) passengers: AD passengers who don't leave the hub before departing.
    - Brief visiting (BV) passengers: AD passengers who travel in Shanghai before departing.
  - Departure and Arrival (DA) passengers: people who head for other cities and return on the same day.

For an active user in the hub communication area, the longest time interval between two continuous records in the area is around one hour due to the automatic location update. The distributions of records over the time in the hub communication area and the city of Shanghai are various for different kinds of people, which can be used to implement the classification. The procedures for identifying each kind of people from active users are listed as follows:

1) Step 1 (Identification of Staff Members): Before the identification of passengers, staff members need to be excluded from the active users first. Since staff members need to work in the transportation hub for certain hours (x hours) on each day, there are two criteria can be used.

1. The time interval between the first record and the last record in the hub communication area is no less than *x-1* hours.

2. The time interval between each two consecutive records in the hub communication area is no more than 1 hour.

2) Step 2 (Grouping of Mobile Phone Records in the Hub Communication Area): After staff members are identified, the remaining users are all considered to be passengers. A passenger may appear at the transportation hub more than once during the same day. During the time period between two appearances, the passenger may travel inside the city or head for another city (DA passengers). We set the minimum time interval as y hours for a roundtrip of a passenger between Shanghai and another city. In order to determine whether a passenger leaves the city between two appearances, it is necessary to group the records based on the distributions. If the time interval of two consecutive records generated in the hub communication area is more than y hours for a passenger, the records will be listed in different groups as shown in Fig.8 (a).

3) Step 3 (Identifying Passengers Based on Criteria): The identification is mainly based on the distribution of the records generated in Shanghai and the transportation hub. For the external passengers, there are no records when they are not in the city. Between the city boundary and the transportation hub, there is also some distance they may need to travel. The maximum time interval to cover this distance is set as z hours.

After understanding the time relationship between records generated in Shanghai and the transportation hub, three

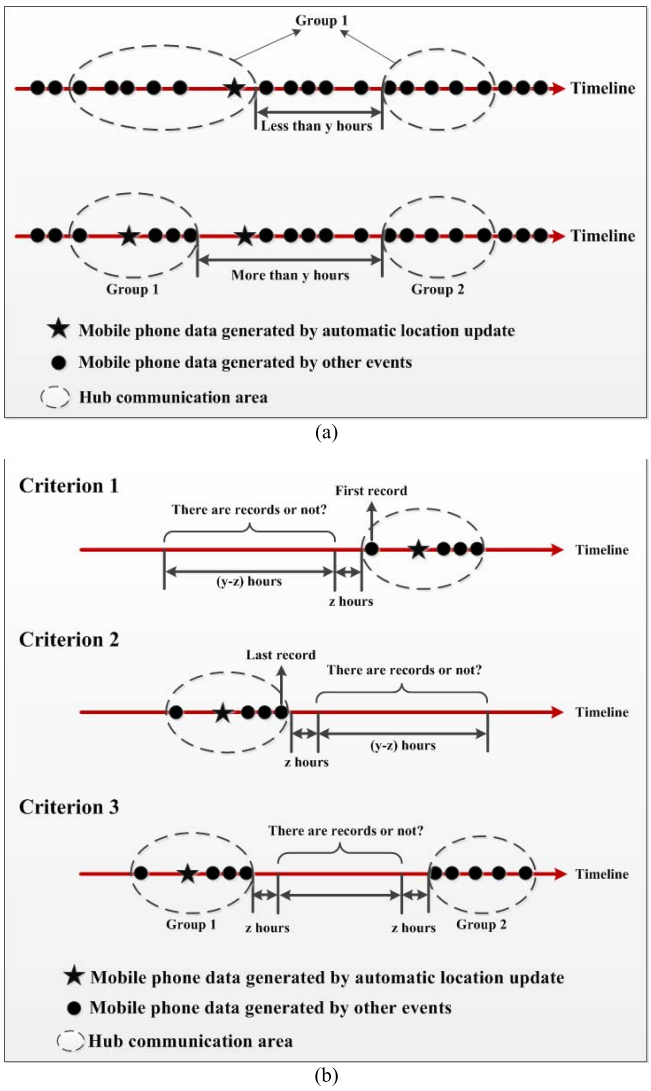


Fig. 8. Temporal distributions of records for Step 2 and Step 3. (a) Temporal distribution of records for different groups in Step 2. (b) Temporal distribution of the records for three criterion in Step 3.

criteria can be proposed to perform the identification as shown in Fig.8 (b).

Criterion 1: This criterion is used to identify whether a passenger comes from another city. The researched time interval is y hours to z hours before the first record generated on the research day in the hub communication area. If there are no records in the city during this time interval, the passenger is considered to come from another city.

Criterion 2: This criterion is used to identify whether a passenger head for another city. The researched time interval is z hours to y hours after the last record generated on the research day in the hub communication area. If there are no records in the city during this time interval, the passenger is considered to head for another city.

Criterion 3: This criterion is used to identify whether a passenger leaves the city during the time interval between two groups of records generated in the hub communication area. The researched time interval is z hours after the last record in the former group to z hours before the first record in the latter group. If there are no records in the city during this time interval, the passenger is considered to leave the city.

4) Step 4 (Identifying PB Passengers and BV Passengers): AD passengers can be identified after Step 3, and we need to further find whether the passengers just pass by the city. The criterion to identify PB passengers is that the time interval between each two consecutive records in the hub communication area is no more than 1 hour. The rest of AD passengers are identified as BV passengers.

In this paper, the parameters are respectively set as $x = 8$ hours, $y = 8$ hours, $z = 0.83$ hours (50 mins). After the classification, the number of each kind of people is also estimated based on the market share. The average results on chosen days are shown in Fig.9 (a), the standard deviations are shown in Fig.9 (b) and the time series of average results through the entire day are shown in Fig.9 (c). According to the results, the average number of passengers is estimated as 625,042 including 350,906 incity passengers and 274,136 external passengers. The standard deviation of each kind of passengers is relatively small compared to the average results (shown in Fig.9 (b)).

To validate the results, we find the passenger information in Nov. 2013 on the website of Hongqiao central business district [22]. In this month, the average number of passengers per day is around 704,300 with 401,000 incity passengers and 303,300 external passengers. The relative errors between our estimated results and the statistical results are respectively 12.7%, 14.3% and 10.6%. The main reason for the errors is that Hongqiao central business district is larger than the study area (core area of the district) in this study. And there is another airport terminal (T1 terminal) located about two kilometers away from our study area. So there are some incity and external passengers who are not calculated in our result. Thus, we further find the design passenger throughput of T1 terminal, which is around 27,400 per day (10 million per year) [23]. Adding this number, the result of our estimated external passengers is pretty close to the statistical result (relative error is 0.6%).

As shown in Fig.9 (a), the numbers of arrival passengers and departure passengers are close, and the situation is similar for the numbers of BV passengers and DA passengers. It can be found in Fig.9 (c), the peak values appear at around 8:00 and 17:00 in the time series of incity passengers, which reflects the importance of Hongqiao hub for commuting travel of incity passengers. Departure passengers are active from 6:00 to 19:00, however the active time period of arrival passengers is obvious latter which is from 10:00 to 21:00. For the time series of staff members, the time series stays steady during the work time in the day which is around 8:00-17:00.

B. Access Characteristics of Passengers

There are a mass of passengers entering and exiting the transportation hub at different times of the day. Based on the classification results, we attempt to analyze the access characteristics of these passengers to Hongqiao hub. It could provide sufficient transportation information for different groups including managers, passengers, and planners, which are difficult to collect by traditional methods. Moreover, we want

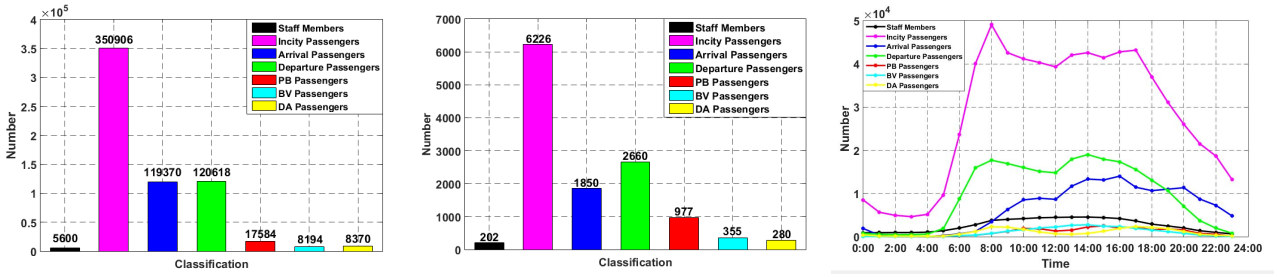


Fig. 9. Classification results of People in the transportation hub. (a) Average numbers of each kind of people. (b) Standard deviations of each kind of people. (c) Time series of each kind of people.

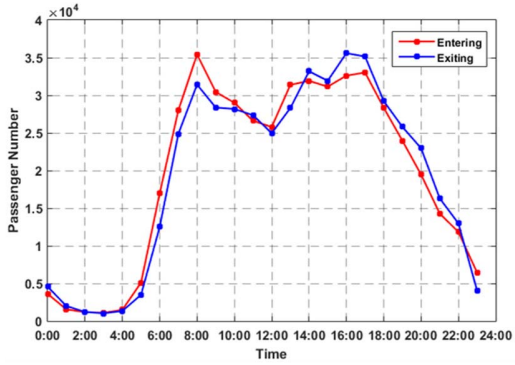


Fig. 10. The time series of passengers entering and exiting the transportation hub.

to further verify the potential of mobile phone data for researching passengers' activity with applying our research results to trace the travel information of passengers.

1) Number (The Number of Entering/Exiting Passengers Over the Day): For the managers of the transportation hub, information related to the number of passengers is necessary to understand the operation of the hub, arrange the monitoring of key areas, etc. Besides the information obtained in the classification results, the number of entering/exiting passengers over the day is also worth further researching.

In order to study the information, passengers entering and exiting the transportation hub needs to be classified first. The time period used for classification is one hour in this section. For instance, passengers in transportation hub between 8:00-9:00 are classified as follow.

1. If there are no records of the passengers in the hub communication area between 7:00-8:00, they are supposed to enter the transportation hub between 8:00-9:00. The time of the first record between 8:00-9:00 is assumed to be the entering time.

2. If there are no records of the passengers in the hub communication area between 9:00-10:00, they are supposed to exit the transportation hub between 8:00-9:00. And the time of the last record between 8:00-9:00 is assumed to be the exiting time.

Fig.10 shows the average number of passengers entering and exiting the transportation hub in time series on the chosen days. The results for the two types of passengers are pretty close which reflects that passenger flow is always in a state of movement. Rush hours are consistent with that in the classification results, and valley values in the daytime appear at 12:00-13:00. The transportation hub is much less active at night until 6:00 in the morning.

2) Distance (The Travel Distances of Passengers Within Certain Time in the City): Before the entering or after the exiting, some passengers need to travel in the city by urban transportation like public transportation, taxies, etc. It takes some time for passengers to cover the travel distances. Passengers need such information to arrange their schedule to catch a train or attend a meeting on time.

In this part, passengers entering or exiting the transportation hub between 8:00-9:00, 12:00-13:00 and 16:00-17:00 are chosen for researching the information. For passengers entering the hub, the travel distances in the city are calculated within a certain time interval before entering. For passengers exiting the hub, the travel distances in the city are calculated within a certain time interval after exiting. There are four values set for the time interval, which respectively are 15 mins, 30 mins, 45 mins, and 60 mins. In this paper, the travel distance of a passenger is measured by the Euclidean distance between the base stations of their first and last records within a study time interval, and the travel speed is computed as the travel distance divided by the time between the records. We propose two indexes, maximum travel distance (MTD) and average travel distance (ATD). MTD is the maximum value of the travel distances of all entering/exiting passengers for the specific time interval, while ATD is the average value.

In order to exclude passengers on the trains and the coaches, the maximum travel speed of passengers is set to 100km/h.

For each time interval, the average results of two indexes are calculated on the chosen days, and the standard deviations are also displayed. In Fig.11 (a, e), it shows that the MTD of passengers are around 50-60km within one hour. According to the distances, the passengers with the MTD tend to travel between the hub and the townships in the city. And they are not the major component of the whole passengers, which leads to slightly higher standard deviations of MTD within 45 mins and 60 mins (as shown in Fig.11 (b, f)).

The ATD in Fig.11 (c, g) keep increasing from 15 mins to 60 mins which are all around 9-12km within one hour. And the results are close to the distance between Hongqiao hub and the city center, which is about 12km. It can be seen from Fig.11 (d, h) that the standard deviations of ATD are pretty small, which explains the travel behavior of passengers have

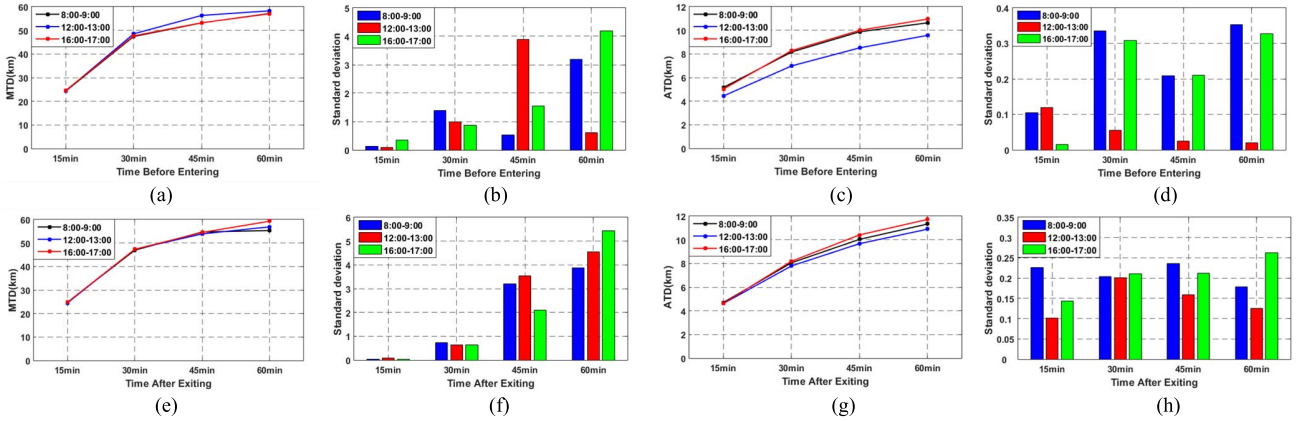


Fig. 11. Travel distances in the city within different time intervals. (a) MTD for people entering the transportation hub. (b) Standard deviation of MTD in (a). (c) ATD for people entering the transportation hub. (d) Standard deviation of ATD in (c). (e) MTD for people exiting the transportation hub. (f) Standard deviation of MTD in (e). (g) ATD for people exiting the transportation hub. (h) Standard deviation of ATD in (g).

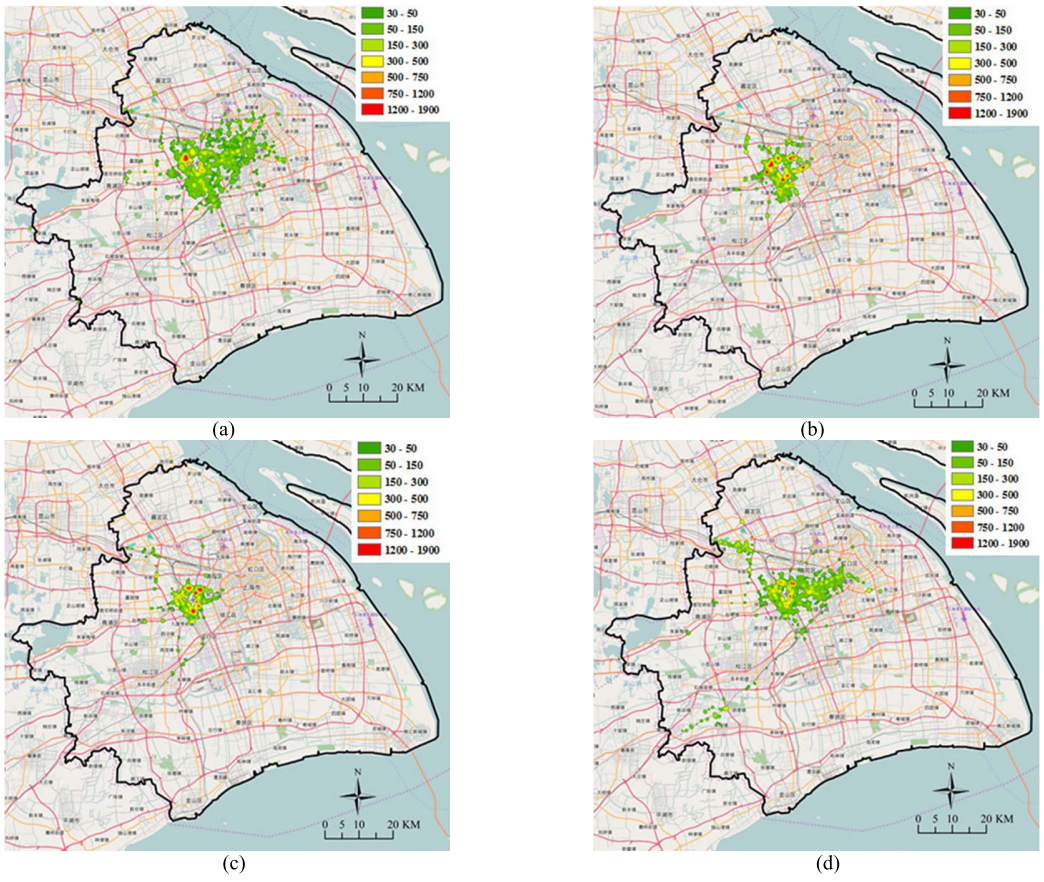


Fig. 12. Distributions of passengers within 15 mins and 60 mins in the city. (a) Distribution of passengers within 60 mins before entering times. (b) Distribution of passengers within 15 mins before entering times. (c) Distribution of passengers within 15 mins after exiting times. (d) Distribution of passengers within 60 mins after exiting times.

the characteristics of stability and tendency. Moreover, it is obvious that passengers entering/exiting the transportation hub at rush hours have longer average travel distances. The reason for this phenomenon may be that the length of commuting trips at rush ours tends to be longer.

3) Distribution (The Distribution of the Passenger Flow in the City): Transportation planners have long been trying to understand where passengers come from and head for. The distribution of the passenger flow in the city can be used to support the transportation planning, considering the importance of Hongqiao hub.

The distribution within 15 mins and 60 mins is visualized for passengers entering or exiting the transportation hub between 8:00-9:00 on Nov.28. For the sake of visualization, kernel density method is used to estimate the density of passengers with the assumption that the communication radius of base stations is set to 850m. The density is the number of passengers per square kilometer, and we only display the

density higher than 30 to show the main moving trend of passengers. It can be seen from Fig.12 that passengers tend to come from or go towards the area of city center which is consistent with the results of average travel distances.

VI. Conclusion

In this paper, characteristics of passenger flow related to Hongqiao transportation hub are analyzed based on mobile phone data in the city of Shanghai, China. The contributions are mainly focused on four major aspects. First, a temporal-spatial clustering method is proposed to identify the passenger active area of Hongqiao hub in the wireless communication space. This is different from the clustering methods in the previous studies [6], [21] for that the time series of active user numbers are also used to characterize the base stations. Second, passengers are identified from mobile phone users in the entire hub and classified into certain types. The time series of passenger numbers through the entire day are also analyzed. Third, access characteristics of passengers are analyzed to provide various kinds of information for passengers and transportation managers which are difficult to collect by traditional methods. Fourth, this study further verifies the potential of mobile phone data in researching human activity of special regions and prepares basic information for studying travel behavior of the hub passengers in the city.

In conclusion, the results indicate several key characteristics which are useful for different objects including managers, passengers, and planners. First, the passenger active range of Hongqiao hub in the communication space is about 850 meters outside the physical boundary. Second, mobile phone data shows the potential for monitoring the numbers of passengers. And rush hours of the transportation hub are around 8:00-9:00 and 16:00-17:00. Third, the maximum travel distances of passengers in the city are between 50 km to 60 km within one hour, while the average travel distances are around 10 km to 12 km within one hour. Passengers tend to have longer travel distances at rush hours.

Future work of this study includes the following. Data from other sources can also be utilized to improve the quality of results. The travel behavior for different types of passengers inside the city can be further explored. Based on the travel patterns, different objectives for passengers traveling to this city can be estimated. Moreover, additional research is also needed for the transportation hubs with different sizes and locations in other cities.

References

[1] J. A. Carrasco, B. Hogan, B. Wellman, and E. J. Miller, "Collecting social network data to study social activity-travel behavior: An egocentric approach," *Environ. Plan. B, Plan, Des.*, vol. 35, no. 6, pp. 961–980, Dec. 2008.

[2] R. SA, M. A. Karim, T. Z. Qiu, and A. Kim, "Origin-destination trip estimation from anonymous cell phone and foursquare data," in *Proc. 94th Transp. Res. Board Annu. Meeting*, Washington, DC, USA, Art. ID 15-2379, 2015.

[3] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *J. Urban Technol.*, vol. 17, no. 1, pp. 3–27, Apr. 2010.

[4] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: Learning from GPS history data for collaborative recommendation," *Artif. Intell.*, vols. 184–185, no. 2, pp. 17–37, Jun. 2012.

[5] Y. Shen, M. P. Kwan, and Y. Chai, "Investigating commuting flexibility with GPS data and 3D geovisualization: A case study of Beijing, China," *J. Transp. Geogr.*, vol. 32, pp. 1–11, Oct. 2013.

[6] G. Sagl, E. Delmelle, and E. Delmelle, "Mapping collective human activity in an urban environment based on mobile phone data," *Cartogr. Geogr. Inf. Sci.*, vol. 41, no. 3, pp. 272–285, Feb. 2014.

[7] C. Ratti, S. Williams, D. Frenchman, and R. M. Pulselli, "Mobile landscapes: Using location data from cell phones for urban analysis," *Environ. Plan. B, Plan. Des.*, vol. 33, no. 5, pp. 727–748, Oct. 2006.

[8] R. A. Becker *et al.*, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 18–26, Oct. 2011.

[9] G. Sagl, M. Loidl, and E. Beinat, "A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic," *ISPRS Int. J. Geo Inf.*, vol. 1, no. 3, pp. 256–271, Nov. 2012.

[10] J. M. Krisp, "Planning fire and rescue services by visualizing mobile phone density," *J. Urban Technol.*, vol. 17, no. 1, pp. 61–69, Mar. 2010.

[11] M. S. Iqbal, C. F. Choudhury, P. Wang, and C. M. González, "Development of origin-destination matrices using mobile phone call data," *Transp. Res. C, Emerg. Technol.*, vol. 40, no. 1, pp. 63–74, Mar. 2014.

[12] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiretta, and M. C. González, "Analyzing cell phone location data for urban travel: Current methods, limitations and opportunities," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2526, pp. 126–135, Jan. 2015.

[13] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 301–313, Jan. 2013.

[14] M. C. Gonzalez, C. A. Hidalgo, and L. A. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[15] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and L. A. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *J. Phys. A, Math. Theory*, vol. 41, no. 22, May 2008.

[16] M. Wang, C. Chen, and J. Ma, "Time-of-day dependence of location variability: Application of passively-generated mobile phone dataset," in *Proc. 94th Transp. Res. Board Annu. Meeting*, Washington, DC, USA, Art. ID 15-3159, 2015.

[17] (2013). *Shanghai Hongqiao Central Business District, Passenger Flow Information of Hongqiao Hub*. [Online]. Available: http://www.shhqcbd.gov.cn/html/shhq/shhq_2013/Info/Detail_6403.htm

[18] P. Yang, T. Zhu, X. Wan, and X. Wang, "Identifying significant places using multi-day call detail records," in *Proc. 26th IEEE ICTAI*, Limassol, Cyprus, Dec. 2014, pp. 360–366.

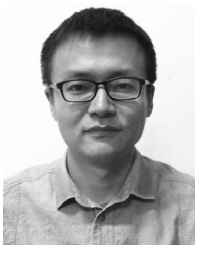
[19] K. Wang and T. Gasser, "Alignment of curves by dynamic time warping," *Ann. Stat.*, vol. 25, no. 3, pp. 1251–1276, Jun. 1997.

[20] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 1–6.

[21] H. Dong *et al.*, "Urban traffic commuting analysis based on mobile phone data," in *Proc. 17th IEEE ITSC*, Qingdao, China, Oct. 2014, pp. 611–616.

[22] (Nov. 2013). *Shanghai Hongqiao Central Business District, Passenger Flow Information of Hongqiao Hub*. [Online]. Available: http://www.shhqcbd.gov.cn/html/shhq/shhq_2013/Info/Detail_6352.htm

[23] (Nov. 2014). *Shanghai Hongqiao Central Business District, Reconstruction of T1 Terminal*. [Online]. Available: http://www.shhqcbd.gov.cn/html/shhq/shhq_xwzx_mtjj/Info/Detail_6859.htm



Gang Zhong received the B.S. degree from Southeast University, Nanjing, China, in 2012. He is currently working toward the Ph.D. degree with the Research Center for Internet of Mobility, Southeast University.

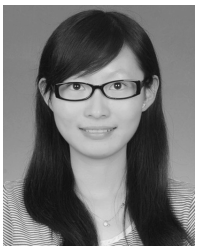
His research interests are related to the use of georeferenced mobile phone data in applications of urban analysis and transportation system planning.

Xia Wan received the Ph.D. degree from Southeast University, Nanjing, China, in 2011. She is working toward pursuing another Ph.D. degree with University of Wisconsin–Madison, WI.

Her research interests include traffic flow under connected vehicles environment and the transportation application of mobile phone data.

Jian Zhang (M'13) received the Ph.D. degree from Southeast University, Nanjing, China, in 2011.

He is the Vice Director of the Research Center for Internet of Mobility, Southeast University. He is also a member of the American Society of Civil Engineers. His research interests include transportation application of mobile phone data, connected vehicles, and public transportation system.



Tingting Yin received the B.S. degree from Southeast University, Nanjing, China, in 2013. She is currently working toward the M.S. degree with the Research Center for Internet of Mobility, Southeast University.

Her research interests are related to urban public transportation systems under the connected vehicle environment.

Bin Ran received the Ph.D. degree from University of Illinois, Chicago, USA, in 1993.

He is a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, WI, USA, and the Director of the Research Center for Internet of Mobility, Southeast University, Nanjing, China. He is one of the co-founders of the Chinese Overseas Transportation Association, and he was the first Chairman. He has authored or co-authored over 90 articles in international journals, including *Transportation Science, Transportation Research Part B*, and *Transportation Research Part C*.