Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 8151520, 10 pages
https://doi.org/10.1155/2022/8151520

Research Article

Extracting the Complete Travel Trajectory of Subway Passengers Based on Mobile Phone Data

Junwei Zhang,[1] Wei Wu,[2] Qixiu Cheng ,[1,3] Weiping Tong ,[1] Anish Khadka,[1] Xiao Fu ,[1] and Ziyuan Gu[1]

[1]*School of Transportation, Southeast University, Nanjing, China*
[2]*Zhejiang Institute of Communications Co., Ltd., Hangzhou, China*
[3]*Department of Logistics & Maritime Studies, The Hong Kong Polytechnic University, Hong Kong, China*

Correspondence should be addressed to Qixiu Cheng; qixiu.cheng@polyu.edu.hk and Weiping Tong; wptong@seu.edu.cn

Received 19 August 2021; Accepted 24 December 2021; Published 25 January 2022

Academic Editor: Elżbieta Macioszek

The usage of mobile phones has undergone tremendous growth in the past decades. Large amounts of mobile-phone signaling data (MSD) are generated while using various mobile phone applications. The large-scale MSD presents opportunities for transport planners to utilize it for better planning and management of the transportation system. In this paper, we use MSD to analyze subway passengers' travel behavior and extract their complete travel trajectories. The complete travel trajectories of subway passengers include their trajectories, both inside and outside the subway system. In the first stage, the MSD from the subway base stations is selected, sorted by time, and the rough trajectory in the subway system is extracted. The ground base stations around the subway station are then considered to correct the boarding and alighting subway stations in order to obtain a more detailed trajectory. In the second stage, the service range of the base station is determined according to the Thiessen polygon, and a temporal dynamic threshold is proposed to extract the passenger's stop point outside the subway system. Finally, the complete trajectories of subway passengers are obtained. The proposed algorithms are verified using a set of MSD collected in Suzhou, China. The results show that the proposed algorithms can effectively extract the complete travel trajectory of subway passengers.

1. Introduction

With the rapid development of information and communication technologies, we have entered the era of big data. Many new sources of data, including mobile phone signaling data (MSD), have emerged and have been used to study travel behavior. Compared with traditional travel data, MSD has many unique features and advantages, such as its wide coverage, its low cost, its reliability, and the fact that it can be monitored and processed in real-time [1, 2]. MSD also has some limitations. It lacks the psychological information which is needed to infer the travel purpose and explain people's travel behavior. However, its unprecedented coverage of population and geographic areas is conducive to collecting information on traffic behavior. MSD contains the user's location information, which is recorded as the associated base station ID. In addition to understanding the flow of passengers at subway stations, the complete trajectory of subway passengers is useful in extracting the origin and destination of subway passengers as well.

Extracting the trajectories of subway passengers has become a hot research topic as it can potentially assist the development and management of subway systems. Studies in the literature mostly considered questionnaire surveys and subway transaction records as the data source to extract trajectories [3–5]. Survey-based data can provide valuable insights into the respondents' sociodemographic characteristics, while transaction-based data usually contains information regarding which subway stations were used to enter and exit the subway system. However, these data cannot accurately reflect the entire trajectory of subway passengers, especially for the linkage part outside the subway system (the first mile and last mile). To this end, we aim to extract the entire trajectory of subway passengers based on MSD as it can be generated both inside and outside the subway system. Telecommunications companies set up several 4G or 5G base stations covering a large geographic area that enables users to stay connected to the network and enjoy mobile services, during which users can be localized, even when they are in the subway system. Thus, MSD can provide information for the complete trip of passengers. Besides the boarding and alighting subway stations, we can also obtain the en-route stations and transfer behavior and their travel trajectories outside the subway system.

The spatial resolution of MSD data depends on the service radius of each base station, which may vary significantly across the multiple telecommunications companies present in each specific region. Moreover, radio frequency signals between base stations and mobile phone devices are unstable. Consequently, when using MSD to extract travel trajectories, we cannot get an accurate geographic location. Thus, the paper proposes an effective method to reduce errors and improve the travel trajectory extraction.

1.1. Literature Review.

In the last decades, considerable effort has been made to identify and describe travel behavior, including travel duration, travel distance, travel modes, trip sequences or complex trip-chains, trip destinations, travel companions [6–10], etc. Past travel behavior studies have applied mobile phone data to detect stops and extract trips. In general, passenger stops are detected by merely counting the frequency of mobile phone data in each area [11, 12]. Location information is recorded as longitude and latitude of base stations, not passengers, which is inevitably inaccurate [1]. Therefore, scholars used clustering methods to identify stop points [12], such as the distance-based clustering method [13]. In addition to the detection of stop points, types of activities and travel modes are often explored. Activities were detected by frequency and duration, such as staying at home and working in the workplace [14]. Different travel modes have been detected by roughly estimating the speed based on the rate of change between connected base stations [15, 16]. However, identifiable travel modes are limited, including stationary, walking, and motorized travel modes [15].

In terms of extracting passenger travel trajectory, it can be estimated by map matching after the initial preprocessing that allocates MSD points to specific locations [17, 18]. Selecting the nearest road or transportation node as the origination or destination, the author matched the OD to a transportation network map and roughly estimated the travel trajectory [19]. The travel trajectory result from MSD can be used to analyze the travel characteristics of passengers, whether it is individual behavior at the microlevel or urban planning at the macrolevel [18, 20, 21].

For subway passengers, geographic information matching algorithms can also be used to identify the transfer routes and stations by using MSD and subway-related data [22]. In addition to MSD, Wifi data can also accurately detect subway passenger flow and obtain the behavior of entrance, transfer, and exit. However, the percentage of passengers using Wifi varies. It is difficult to guarantee the accuracy of data, so additional equipment is needed to ensure the full coverage of Wifi signals [23]. Micropedestrian simulation is used to simulate the behavior of passengers in the subway station, but the simulation model cannot fully describe the complex travel behavior of passengers [24, 25]. Some scholars use automatic fare collection (AFC) system data to extract the passengers' route choices by clustering [26, 27]. AFC data is used along with the train operation data to develop a reverse model of passenger travel trajectory. The essence of this method is "apriority" and cannot accurately restore the travel of passengers [28, 29]. However, AFC data can only obtain the travel characteristics at the level of the subway station, not the specific travel routes of passengers in the rail transit system.

For rail transit, scholars usually use MSD to study the subway occupancy rate, the service scope of the subway station, and the land use around the subway station. The research of extracting passenger trajectory by MSD is mainly limited in theory or simulation. Few studies have used large-scale historical MSD to study the base station trajectory of passengers. The current research only considers passenger trajectory in the subway and does not integrate passenger trajectory outside the subway in their studies. The complete travel trajectories of passengers are useful to grasp the flow of passengers at subway stations and extract the origin and destination of subway passengers. Therefore, how to use MSD to extract the overall journey trajectory of subway passengers needs further analysis and research.

1.2. Objectives and Contributions.

Compared with other data sources, MSD has many advantages, such as large scale and low cost. However, due to the sparse layout of mobile phone base stations, MSD lacks stability, resulting in errors in identifying the entrance and exit stations. Besides, the time recorded by each base station includes the residence time and movement time within the base station, which leads to the identification error of the stop point.

The original mobile phone data provide the time and location information about the cellphone towers, from which the trajectory information of users is not available directly, especially the stop points. In this study, we propose a new algorithm to extract the complete trajectory of subway passengers, which contains the space-time information of the origin and destination outside and inside the subway system. The contributions of this paper are twofold. First, the ground base station (GBS) around the subway station is proposed to improve the passenger trajectory in the subway system. Secondly, a temporal dynamic threshold is proposed to improve the extraction accuracy of the passenger stop point outside the subway system.

The rest of this paper is structured as follows. The second section describes the specific problems and the relationship between the passenger travel trajectory, MSD, and mobile phone base stations. The third section discusses the extraction of subway passenger trajectory. Then, the paper introduces the identification of stop points outside the subway system. Experiments are carried out on real data sets to evaluate the performance of the algorithm. Finally, the discussion section is followed by the conclusion and recommendations.

2. Problem Description

MSD data is collected by telecommunications companies. The data consists of the user's unique code and position information. The position is recorded as the base station ID. MSD is generated whenever a user triggers an event, such as the use of mobile communications and the Internet, or passes through the service range of a base station [1]. The attributes of a mobile phone data record contain unique user code, location time, latitude, and longitude of the base station, as shown in Table 1. It should be emphasized that the location information contained in each piece of data is not the location of the user but the location of the mobile phone base station. The MSD is a static dataset. We can only know that the user has been around the base station at that time. Therefore, we cannot directly distinguish from the large amount of user data that represents when/where a trip starts and when/where it ends.

The signal coverage of the mobile phone base station in the subway system is different from that in the ground system. When subway passengers are in the subway system, their mobile devices generate the signal switching with the subway base station (SBS) instead of the ground base station (GBS). Data points in the MSD that were generated by SBS can be distinguished through the specific information about these two types of cellphone towers from mobile phone operators. This allows us to identify which individuals have, on any date, used the subway system. Thus, MSD generated by SBS in the subway system can be divided into three parts, as shown in Figure 1: (i) Change in the location area of mobile phones between SBS and GBS when entering and exiting the subway station; (ii) Upon usage of mobile phone in the subway; (iii) Change in the location area of mobile phones between the different SBS when subway moving.

For subway passengers, their complete paths comprise the travel in the subway system and the travel out of the subway system. These two parts of the travel path are extracted separately and are combined to get the passenger's complete travel trajectory. The basic procedure to extract the complete trajectory is shown in Figure 2. In the first stage, the MSD from the subway base stations is selected to extract the rough trajectory in the subway system while considering the ground base stations around the subway station to correct the boarding and alighting subway stations. In the second stage, the MSD from the ground base stations is selected, and a temporal dynamic threshold is proposed to extract the passenger's stop point outside the subway system. Finally, we can obtain the complete trajectory of subway passengers.

Table 1: Data tags of mobile phone 4G signaling data.

Number	Data Tag	Data type	Description
1	mdn	string	user unique code
2	start_time	string	data generation time
3	end_time	string	none
4	lat	string	base station latitude
5	lon	string	base station longitude
6	cell_id	string	base station code
7	lac	string	location code
8	city_id	string	city number
9	clndr_dt_id	int	date

3. Base Station Trajectory Extraction in the Subway System

As mentioned above, subway passengers can be distinguished by whether MSD occurs in the subway system. The total MSD generated by SBS is selected to detect the travel trajectory in the subway system. However, due to the sparse layout of the base station, when passengers board or alight at some stations, there is a possibility the MSD will not be generated. Thus, in this paper, after the MSD is sorted by time to obtain the position sequence of the base station, GBS around the subway station is used to correct this situation.

The passenger trajectory in the subway system can be further divided into two parts: at the subway station and on the subway. While passengers are in the subway trains, it is possible to determine the two stations between which specific MSD data points are generated. To do so, SBS is divided into the subway-station base station and the intersite base station. The specific information about these two types of base stations is offered by the mobile phone operators, which is used to differentiate them directly. Figure 3 shows the specific base station classification. The subway-station base station includes the corresponding subway station number. The intersite base station includes the previous subway station number and the subsequent subway station number. Therefore, in addition to the original field shown in Table 1, the processed MSD includes the following fields, as shown in Table 2.

By selecting the data with the SBS base station and arranging the base station sequence chronologically, the rough result of the subway passenger trajectory can be obtained. However, the layout and selection of the base station have a significant influence on MSD acquisition. For instance, some subway stations do not have enough base stations. Thus, passengers boarding or alighting at such stations do not have signal switching with the subway-station base station when passing in or out. Therefore, their trajectory in the subway system is incomplete as their starting or ending point is not located at the subway-station base station. In order to overcome this limitation, GBS within the range of 500 m around the subway station is added. Considering the GBS near the time point when passengers enter and exit the station, the result is modified to obtain the final travel trajectory in the subway system. If MSD appears on the GBS around the station for a period of time before (after) the passenger enters (leaves) the subway

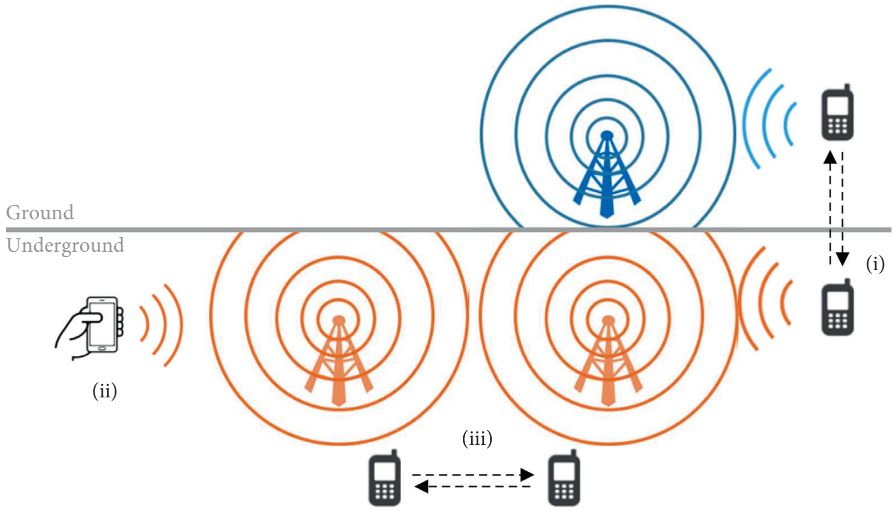

Figure 1: MSD generated when entering or leaving subway station.

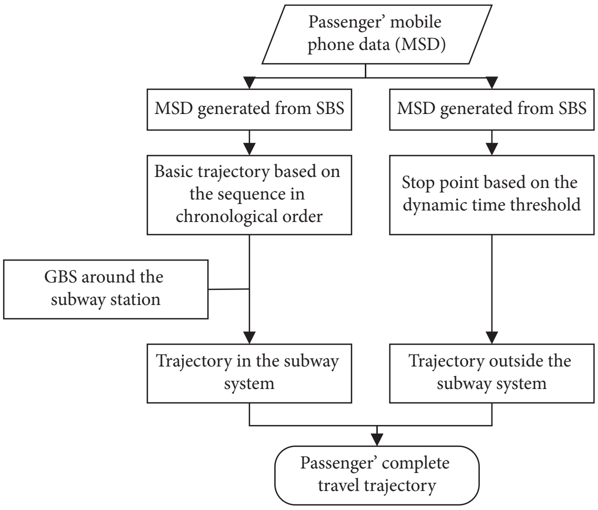
Passenger' mobile phone data (MSD)
MSD generated from SBS
MSD generated from SBS
Basic trajectory based on the sequence in chronological order
Stop point based on the dynamic time threshold
GBS around the subway station
Trajectory in the subway system
Trajectory outside the subway system
Passenger' complete travel trajectory

Figure 2: Extraction of complete travel trajectory.

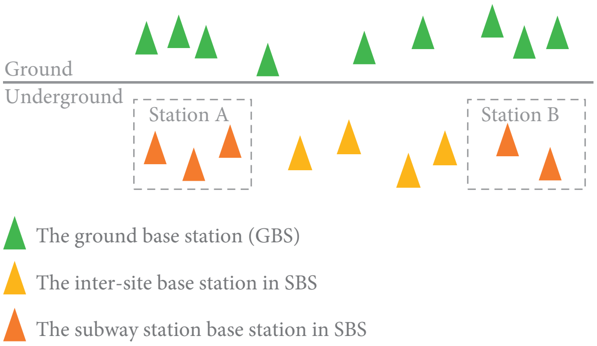
Ground
Underground
Station A
Station B
The ground base station (GBS)
The inter-site base station in SBS
The subway station base station in SBS

Figure 3: Base station classification.

system, then the station corresponding to the GBS is the starting or ending point of the passenger in the subway system.

The process of trajectory extraction in the subway system is as follows:

Step 1: (Data selecting).

Select MSD with SBS base station, and add the data tags indicated by Table 2.

Step 2: (Basic base station trajectory).

MSD is grouped according to the subway passenger and sorted according to the occurrence time field. The interval between adjacent data records is then calculated while extracting passenger travel time, metro station, and next recorded information. The passenger trajectory segment consists of adjacent stations. Obtain the basic trajectory of the subway system.

Step 3: (Eliminate the erroneous data).

(i) Eliminate passengers of MSD records with SBS less than 2, travel time less than 5 min, or entering and leaving at the same station. A passenger traveling by subway multiple times can be determined when its adjacent data interval is greater than 60 minutes.

(ii) In addition to underground tunnels, subways can also operate above ground. Thus, SBS in such sections of the subway line is consistent with GBS. We propose speed thresholds, which relate to the passengers who travel by car on the ground. When the speed is less than the speed threshold, it is considered to be traveling by subway

$$\text{speed} = w * \frac{1000}{t}, \tag{1}$$

where w is the number of intermediate subway stations, t is the interval time, and the approximate distance between subway stations is considered to be 1000 m.

Step 4: (Complete base station trajectory).

The GBS in the range of 500 m around the subway station is added. Then the subway station

Table 2: Data tags added after processing.

Number	Column name	Data type	Description
1	stn_or_itl	string	subway-station base station or intersite base station
2	station_id	string	subway station code
3	before_itl	string	previous subway station number
4	after_itl	string	subsequent subway station number

corresponding to GBS, where the last data (first data) in the first 10 min (the last 10 min) of the original OD is located, is considered as the starting point (the endpoint) of the passenger in the subway system. Finally, the complete trajectory in the subway system is obtained.

4. Base Station Trajectory Extraction outside the Subway System

The complete trajectory of subway passengers can be divided into three parts: the travel before entering the station, in the subway, and after exiting the station. There exists a similarity between the first part and the last part in the trajectory extraction. Therefore, this paper focuses on travel after leaving the subway station to extract the passenger trajectory outside the subway system. By setting a spatial threshold and combining the data points from nearby base stations, a set of possible stop points are established. Finally, the temporal dynamic threshold is set to filter out the exact stop point set.

Figure 4 shows the spatial-temporal distribution schematic diagram about MSD. The left side is a latitude and longitude spatial position map. The right side is a spatial-temporal distribution map, where the vertical axis is the spatial distance, and the horizontal axis is the time interval.

Even when passengers stay somewhere and their mobile phones are powered on, the mobile device's signal might be handed off from their base station where they stay to stations in surrounding areas. Thus, a set of MSD generated contains the corresponding base stations changing in a certain spatial range, exhibiting an aggregation in space. The range is determined by the specific base stations. In Figure 4, it shows a circle in the spatial position map and an ellipse in the spatial-temporal distribution map (characterizing the long-time interval and the short spatial distance).

When passengers move, their MSD are updated because they are continually crossing the service range of base stations. The base station trajectory direction shown by this data is similar to the passengers' actual moving direction. The data points are dispersing obviously, like point 3–point 6 (Figure 4) in the spatial position map. In the spatial-temporal distribution map, it is represented by another ellipse with the characteristics of a short time and long-distance.

The spatial distance between mobile base stations and the passengers' residence time in the base station is defined as follows.

4.1. Spatial Distance.

4.1. Spatial Distance. The Haversine formula and spherical distance formula was to calculate the spherical distance based on the longitude and latitude of two points.

4.2. Residence Time. The residence time specifies the duration during which the passenger is within a set of stop points. In previous studies, residence time only considered the time interval between the last mobile phone data and the first within the stop point set [11, 30]. In this paper, the passengers' moving time in the base station is also considered. Thus, the residence time is calculated by the following formula:

$$t = t_1 - t_2, \tag{2}$$

$$t_2 = \frac{R}{V}, \tag{3}$$

where t_1 is the surface residence time. It refers to the time interval between the first data in this data point set and the first data in the next base station outside this data point set. Taking the stop point set K in Figure 4 as an example, the time interval between the $j+2^{\text{th}}$ point and the $j-1^{\text{th}}$ point is the surface residence time of the stop point set K.

t_2 is the moving time. R is the service range of the base station. V is the moving speed. When the passenger moves, the time interval between two mobile phone data with different base stations consists of two parts, the staying time and the moving time in the previous base station. As shown in Figure 5, $R1$ and $R2$ are the range radius of station A and station B. The staying time is the duration when passengers stop at station A, and the moving time is the duration during which passengers move from stop point C to the intersection point of these two base stations' service ranges.

Hence, in order to obtain the moving time, it is necessary to define the service range of the base station. We assume that the range is a circle and use the Thiessen polygon to determine the service range. The characteristics of the Thiessen polygon are as follows:

(1) There is only one central point data (base station) inside each Thiessen polygon;

(2) The distance between the central point and other corresponding points within the Thiessen polygon is the closest;

(3) The distance between the point on the edge of the Thiessen polygon and the central point on the two sides of the Thiessen polygon is equal.

By using ArcGIS software and inputting the base station points, the Thiessen polygon area corresponding to each base station can be determined. In this paper, we assume that the coverage area is circular. Therefore, the radius of coverage R can be obtained from the Thiessen polygon area.

The recognition process of stop point is as follows:

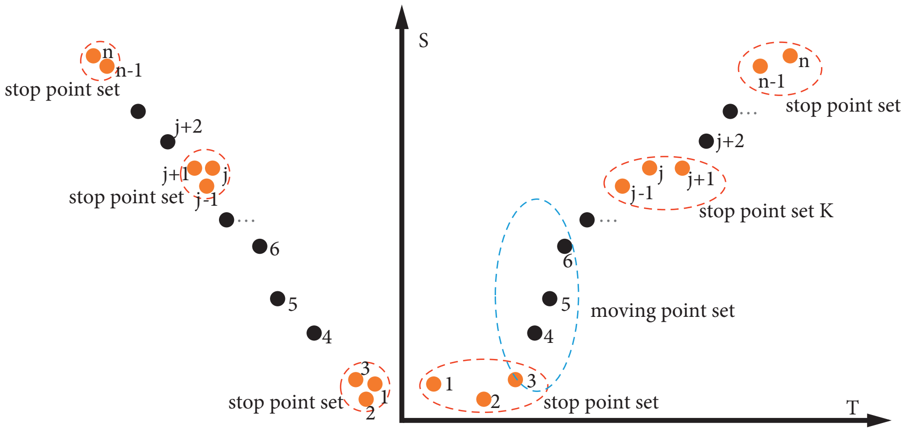

Figure 4: Movement of passengers after leaving the station.

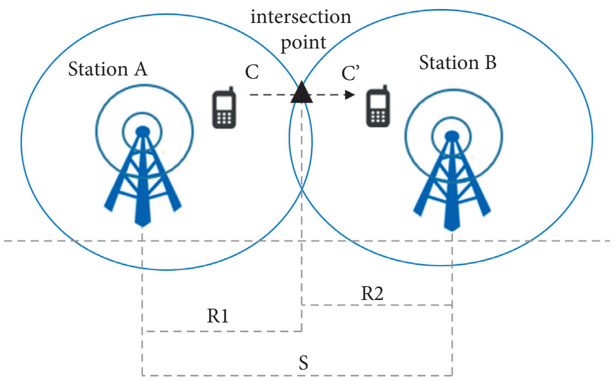

Figure 5: Movement of passengers after leaving the station.

Step 1: (Sort data).

The mobile phone data after exiting the subway station is sorted by time.

Step 2: (Select the alighting station).

Set the first data after leaving the subway station as the station data.

Step 3: (Establish a possible set of stop points).

For the i^{th} $(i = 1, 2, \ldots, n)$ data, a series of data in the continuous-time is searched with the spatial threshold, $D = 500$ m as a radius, considering the average radiation range of urban cell phone base stations [31, 32]. If the distance between the k^{th} $(k = i+1, i+2, \ldots, n)$ data and the i^{th} data is less than the spatial threshold D, the k^{th} data is added to the possible stop point set C_i. Update the spatial position of the i^{th} data to the gravity center of the current possible stop point set C_i. Then continue to judge the $k+1^{th}$ data. Until the distance between the m^{th} data and the i^{th} data does not meet the spatial threshold D, the search is stopped.

Step 4: (Determine the real set of stop points).

For the possible set of stop points C_i, if the maximum spatial distance between the data points is less than the space threshold D and the total residence time is greater than 4 min, mark the possible set of stop points C_i as a real set of stop points. In this step, considering the average speed of various traffic modes in the city, with the spatial threshold set to 500 m, 4 min is chosen as the time threshold [31, 32].

If the maximum spatial distance is greater than the spatial threshold D, assume that the points corresponding to the maximum distance are $k1$ and $k2$ ($k2$ occurs later than $k1$). Then $k2$ and the data points whose occurrence time is later than $k2$ are deleted from the possible set of stop points. The stop point identification is then performed again.

Step 5: (Continue to judge).

If the set of stop points has been determined in Step 4, start from the data, which is not marked as the stop point set. Execute Step 4 cyclically;

If it is not determined as the stop point set in Step 4, the i^{th} data is marked as a moving point, and Step 4 is cyclically executed from the $i+1^{th}$ data.

Step 6: (Determine the travel destination).

MSD, after leaving the subway station, is divided into a stop point set and a moving point set. On this basis, according to the time threshold, the residence time of the stop point set is checked successively to determine the travel destination. Within 500 meters, the residents who stay more than 20 minutes have enough time to complete the purpose of this trip, so they can be considered to have finished traffic travel. Therefore 20 min was chosen as the time threshold [31, 32].

If the i^{th} stop point set's residence time is more than 20 min, it is determined as the travel destination, and the cycle terminates.

If the i^{th} stop point set's residence time is less than 20 min, the judgment of the $i+1^{th}$ destination set is performed.

5. Case Study

Experiments were conducted using the mobile phone data obtained from a mobile network operator in Suzhou, China. The trajectory of subway passengers inside and outside the

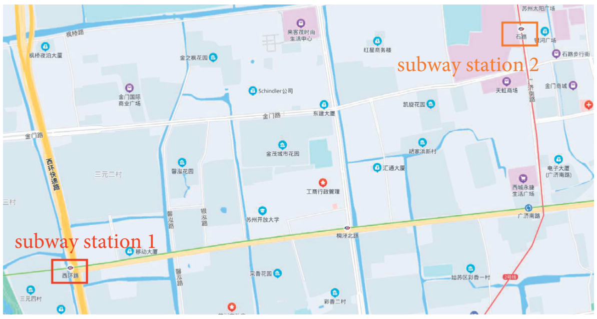

Figure 6: Part of the subway route map.

subway system was extracted. Using the extraction result of the trajectory, the passenger flow and OD between the subway station and the subway line can be statistically analyzed. In this study, the speed threshold of 18 km/h was roughly used to distinguish between cars and subways with mobile phone data. It should be noted that the threshold is only an initial setting to distinguish between cars and subway passengers with the complete mobile phone data. However, the focus of this study is mainly on extracting the complete travel trajectory of subway passengers when the mobile phone data of subway passengers have already been distinguished from all the data collected, rather than focusing on the travel mode identification. In future studies, one can also use some other advanced and sophisticated models to identify different travel modes with mobile phone data [33–35].

A part of the city's subway route map is shown in Figure 6.

In Figure 6, subway station 1 is located in a residential area, where the passenger flow is mainly from nearby residential areas. Subway station 2 is located in a commercial area, where the passenger flow comprises mainly of commute or recreational trips.

Figures 7 and 8 show the daily passenger flow volume of subway stations located in commercial and residential areas, respectively. The vertical axis represents the number of passengers and the horizontal axis represents the hourly time points from 6 : 00 to 22 : 00. It is evident that in the commercial subway station, the outbound passenger flow is greater than the inbound passenger flow during the morning rush hour, while it is the opposite during the evening rush hour. However, in the case of a residential subway station, it is the opposite during both morning and evening peak hours. Besides, between the two peaks, the outbound and inbound passenger flow volume in residential areas is significantly lower than that in commercial areas, which is consistent with the expected traffic volumes for such types of areas. However, it should be noted that this analysis does not rely on any true validation; this is merely a high-level verification if the data produced by our method seems sound. While testing against real-world data would be best, this data was unfortunately not available, therefore making this comparison impossible.

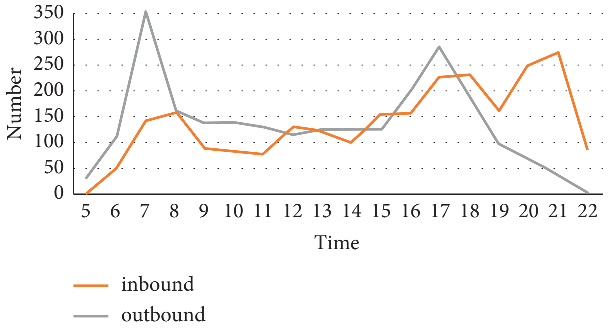

Figure 7: Passenger flow statistics of the subway station in the commercial area.

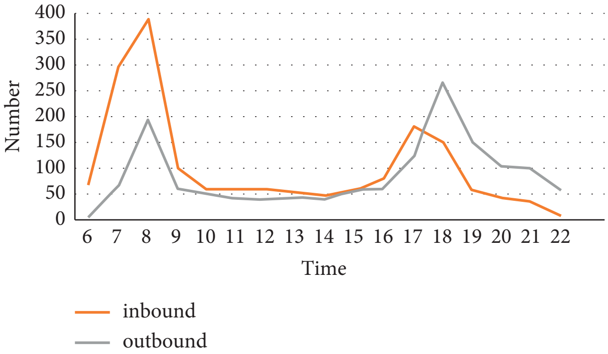

Figure 8: Passenger flow statistics of the subway station in the residential area.

The full-day mobile phone data of a subway passenger is selected randomly, and the data after leaving the subway station is filtered out to process.

Figure 9 is a three-dimensional map of the data points before and after processing, where *lat* indicates the latitude and *lon* indicates the longitude. As can be seen from Figure 9, when the passenger stops, the mobile phone data moves within a small latitude and longitude range. After processing, five data point sets are obtained. Table 3 is the result of the specific point set recognition, with three stop point sets and two moving point sets. The stop point set 3 satisfies the time threshold and is, therefore, determined as a travel destination.

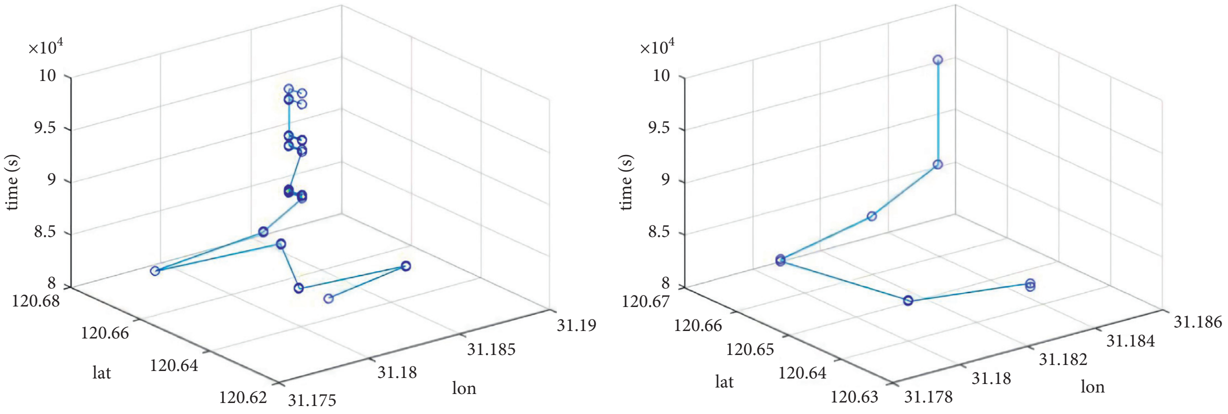

Figure 9: Data point distribution before and after processing.

Table 3: Stop point recognition result.

Point	Stay time (s)
Stop point 1	499
Moving point 1	66
Stop point 2	378
Moving point 2	58
Stop point 3	4415

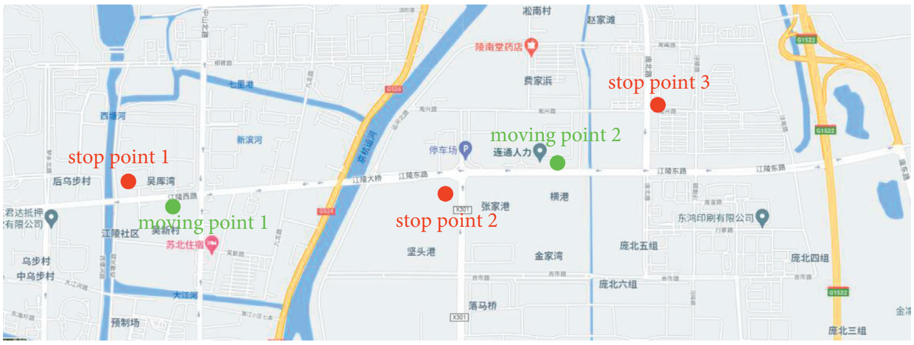

Figure 10: Map matched stop point recognition result.

Figure 10 is the map matching of the stop point recognition result. Stop point 1 is in the subway station. Stop point 2 is close to the road intersection. Stop point 3 is the travel destination, where there is a furniture store. On the contrary, both moving points are on the side of the road.

6. Conclusion

This paper proposed an algorithm to extract the complete trajectory of subway passengers using MSD. The complete travel trajectory includes the trajectories, both inside and outside the subway system. The algorithm contains two stages. In the first stage, the subway passengers' MSD was selected by the subway base station. Further, the basic trajectory inside the subway system based on the time series was supplemented by adding GBS around the subway station. In the second stage, the service range of the base station was divided by the Thiessen polygon. Based on the spatial threshold and the temporal dynamic threshold, the recognition algorithm for the passenger's stop point was established. Thus, the trajectory of subway passengers outside the subway system was obtained. The algorithm was demonstrated through a real data set collected in Suzhou, China. We confirmed that the passenger's flow had an apparent tidal phenomenon, while the trend was converse between the subway stations in the residential area and commercial area. This result, the directionality of the flows, follows what is expected for a commercial and residential area. We use the extracted user trajectories to further extract urban traffic results such as metro passenger volumes, urban OD traffic, and traffic volumes across urban corridors. These results are compared with real city data, thus validating the accuracy of the method. However, without access to any other real data, such as the direct real user trajectory data, further true validation cannot be performed, which is one of the paper's main limitations.

Compared to traditional questionnaire data, MSD is massive and objective, facilitating a more accurate extraction of urban residents' mobile behavior. At the same time, due to the real-time attribute of MSD, the passenger flow of the subway line and station can be detected and controlled at any time, which contributes to reducing the potential safety hazards caused by passenger crowding. In addition, the complete trajectory of the passenger can be obtained by using MSD. We can gain some extra insights into passenger behavior, like the origin and destination outside the subway system and their transfer stations. Further, the organization of subway stations can be optimized. Therefore, for future research directions, multimodal interchange can be considered—further research into the transport options for passengers outside the metro system [36, 37].

There are still a few aspects of this paper that need to be further improved: (1) Further collection of real data is considered to compare the passenger travel trajectories extracted and the actual passenger travel trajectories. (2) Due to the unavailability of further specific information on mobile phone base stations and the layout of the city, the same threshold was set throughout the city. However, the layout of mobile phone base stations differs in different areas of a city in an actual scenario. If the same threshold is set, due to the difference in the service range of base stations, the accuracy will vary. (3) Through example studies, we can obtain urban traffic trip statistics that match the reality by using the proposed method, but it is undeniable that the current trajectory extraction is rather crude. Therefore, future consideration is given to adding cell tower triangulation to the existing method to further improve the accuracy of user trajectories. (4) Contrary to using the data from a single telecommunications company for empirical analysis, in actual situations, multiple telecommunications companies exist. The results only focus on the residents with the service of a specific service provider. Besides, some urban residents (such as the elderly, children) do not use mobile communication devices. Therefore, the current results fail to incorporate urban residents of all ages and backgrounds.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the General Project (71771050) and the Key Project (51638004) of the National Natural Science Foundation of China, the Natural Science Foundation of Jiangsu Province in China under Grant no. BK20180402, and the Department of Science and Technology of Zhejiang Province in China under Grant no. 2020C01057.

References

[1] Z. Wang, S. Y. He, and Y. Leung, "Applying mobile phone data to travel behaviour research: a literature review," *Travel Behaviour and Society*, vol. 11, pp. 141–155, 2018.

[2] J. Huo, X. Fu, Z. Liu, and Q. Zhang, "Short-Term estimation and prediction of pedestrian density in urban hot spots based on mobile phone data," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[3] E. Chen, Z. Ye, C. Wang, and M. Xu, "Subway passenger flow prediction for special events using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1109–1120, 2020.

[4] X. Xie, X. Zhang, J. Chen, Y. L. Wang, and W. J. Chu, "The discrete choice model of urban rail transit passengers' route choice," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, no. 2, pp. 127–131, 2014.

[5] Y. Sun and R. Xu, "Rail transit travel time reliability and estimation of passenger route choice behavior: analysis using automatic fare collection data," *Transportation Research Record*, vol. 2275, no. 1, pp. 58–67, 2012.

[6] Q. An, X. Fu, D. Huang, Q. Cheng, and Z. Liu, "Analysis of adding-runs strategy for peak-hour regular bus services," *Transportation Research Part E: Logistics and Transportation Review*, vol. 143, Article ID 102100, 2020.

[7] X. Fu, W. H. Lam, B. Y. Chen, and Z. Liu, "Maximizing space-time accessibility in multi-modal transit networks: an activity-based approach," *Transportmetrica A: Transport Science*, 2020.

[8] D. Lei, X. Chen, L. Cheng, L. Zhang, S. V. Ukkusuri, and F. Witlox, "Inferring temporal motifs for travel pattern analysis using large scale smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 120, Article ID 102810, 2020.

[9] C. Lyu, X. Wu, Y. Liu, and Z. Liu, "A partial-fréchet-distance-based framework for bus route identification," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[10] Y. Zuo, X. Fu, Z. Liu, and D. Huang, "Short-term forecasts on individual accessibility in bus system based on neural network model," *Journal of Transport Geography*, vol. 93, Article ID 103075, 2021.

[11] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, 2015.

[12] C. Chen, L. Bian, and J. Ma, "From traces to trajectories: how well can we guess activity locations from mobile phone traces?" *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 326–337, 2014.

[13] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie, "Mining Individual Life Pattern Based on Location History," in *Proceedings of the 2009 Tenth International Conference on mobile Data Management: Systems, Services and Middleware*, pp. 1–10, IEEE, Taipei, Taiwan, 2009, May.

[14] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González, "Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities," *Transportation Research Record*, vol. 2526, no. 1, pp. 126–135, 2015.

[15] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks*, vol. 6, no. 2, pp. 1–27, 2010.

[16] T. Feng and H. J. Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transportation*

Research Part C: Emerging Technologies, vol. 37, pp. 118–130, 2013.

[17] Y. Asakura and E. Hato, “Tracking survey for individual travel behaviour using mobile communication instruments,” *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 3-4, pp. 273–291, 2004.

[18] R. Ahas, A. Aasa, S. Silm, and M. Tiru, “Daily rhythms of suburban commuters’ movements in the Tallinn metropolitan area: case study with mobile positioning data,” *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 1, pp. 45–54, 2010.

[19] T. Tettamanti, H. Demeter, and I. Varga, “Route choice estimation based on cellular signaling data,” *Acta Polytechnica Hungarica*, vol. 9, no. 4, pp. 207–220, 2012.

[20] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools, “Building a validation measure for activity-based transportation models based on mobile phone data,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6174–6189, 2014.

[21] M. Y. Ansari, A. Mainuddin, and A. Ahmad, “Spatiotemporal trajectory clustering: a clustering algorithm for spatiotemporal data,” *Expert Systems with Applications*, vol. 178, Article ID 115048, 2021.

[22] J. Lai, Y. Chen, Y. Zhong, D. Wu, and Y. Yuan, “Travel route identification method of subway passengers based on mobile phone location data,” *Journal of Computer Applications*, vol. 33, no. 2, pp. 583–586, 2013.

[23] Z. Chen, H. Zou, H. Jiang, Q. Zhu, Y. C. Soh, and L. Xie, “Fusion of WiFi, smartphone sensors and landmarks using the Kalman filter for indoor localization,” *Sensors*, vol. 15, no. 1, pp. 715–732, 2015.

[24] M. Liao and G. Liu, “Modeling passenger behavior in nonpayment areas at rail transit stations,” *Transportation Research Record*, vol. 2534, no. 1, pp. 101–108, 2015.

[25] D. King, S. Srikukenthiran, and A. Shalaby, “Using simulation to analyze crowd congestion and mitigation at Canadian subway interchanges: case of Bloor-Yonge Station, Toronto, Ontario,” *Transportation Research Record*, vol. 2417, no. 1, pp. 27–36, 2014.

[26] X. Wu, H. Dong, S. Gao, W. Li, and Q. Zhang, “Extracting metro passengers’ route choice via AFC data utilizing Gaussian mixture clustering,” in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1933–1938, IEEE, Maui, HI, USA, November 2018.

[27] X. Chen, Y. Wang, Y. Wang, X. Qu, and X. Ma, “Customized bus route design with pickup and delivery and time windows: model, case study and comparative analysis,” *Expert Systems with Applications*, vol. 168, Article ID 114242, 2021.

[28] B. Si, M. Zhong, J. Liu, Z. Gao, and J. Wu, “Development of a transfer-cost-based logit assignment model for the Beijing rail transit network using automated fare collection data,” *Journal of Advanced Transportation*, vol. 47, no. 3, pp. 297–318, 2013.

[29] F. Zhou and R. H. Xu, “Model of passenger flow assignment for urban rail transit based on entry and exit time constraints,” *Transportation Research Record*, vol. 2284, no. 1, pp. 57–61, 2012.

[30] F. Wang and C. Chen, “On data processing required to derive mobility patterns from passively-generated mobile phone data,” *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 58–74, 2018.

[31] S. Qin, Y. Zuo, Y. Wang, X. Sun, and H. Dong, “Travel trajectories analysis based on call detail record data,” in *Proceedings of the 2017 29th Chinese Control and Decision Conference (CCDC)*, pp. 7051–7056, IEEE, Chongqing, China, 2017, May.

[32] H. Huang, Y. Cheng, and R. Weibel, “Transport mode detection based on mobile phone network data: a systematic review,” *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 297–312, 2019.

[33] M. A. Shafique and E. Hato, “Travel mode detection with varying smartphone data collection frequencies,” *Sensors*, vol. 16, no. 5, p. 716, 2016.

[34] L. Wu, B. Yang, and P. Jing, “Travel mode detection based on GPS raw data collected by smartphones: a systematic review of the existing methodologies,” *Information*, vol. 7, no. 4, p. 67, 2016.

[35] Z. Lu, Z. Long, J. Xia, and C. An, “A random forest model for travel mode identification based on mobile phone signaling data,” *Sustainability*, vol. 11, no. 21, p. 5950, 2019.

[36] D. Huang, Y. Gu, S. Wang, Z. Liu, and W. Zhang, “A two-phase optimization model for the demand-responsive customized bus network design,” *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 1–21, 2020.

[37] Q. Cheng, S. Wang, Z. Liu, and Y. Yuan, “Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data,” *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 422–438, 2019.