A Travel Mode Identification Framework Based on Cellular Signaling Data

Jiatao Chen,1,2 Chen Xiong1,2 and Ming Cai1,2

1School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China
2Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China

Correspondence should be addressed to Ming Cai; caiming@mail.sysu.edu.cn

Received 23 November 2021; Revised 10 February 2022; Accepted 1 March 2022; Published 1 April 2022

Copyright © 2022 Jiatao Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The rapid development of telecommunication network has produced a large amount of spatial-temporal information of mobile phone users. GPS data are typically collected by smartphones apps, which are restricted to small samples of the population. Cellular signaling data (CSD) are usually collected by mobile network operators, which enables researchers to conduct travel behavior analysis of the entire population at a relatively low cost compared to GPS. However, extracting travel mode information from CSD is particularly challenging due to the noise data and low positioning accuracy. This paper proposes a travel mode identification framework based on CSD, which includes data cleansing and travel mode identification. In terms of data cleansing, oscillation sequence and drift data are mainly cleansed. For the oscillation sequence, this paper proposes a detection algorithm based on time window. For the drift data, this paper proposes a detection algorithm based on distance, velocity and frequency. In terms of travel mode identification, the task is divided into two dichotomous problems: motor and non-motor transport identification and public and private transport identification. Each dichotomous problem proposes an algorithm that does not rely on the ground truth dataset for model training. Finally, a ground truth dataset is constructed to verify all algorithms. The result shows that, in terms of data cleansing, the similarity between CSD after cleansing and the actual trajectory according to DTW improved by 101.13% on average. In terms of travel mode identification, the proposed method can achieve similar or even better accuracy than traditional supervised-learning algorithms (94% in motor and non-motor transport identification, 83.5% in public and private transport identification), which can be directly applied to large-scale population analysis scenarios.

1. Introduction

The travel mode of resident’s trips is an important part of traffic demand analysis and urban traffic planning. It can provide data support for transportation planning and policy makers’ decision-making, by identifying residents’ travel mode and understanding the utility and proportion of each travel mode.

Traffic survey is one of the most widely used methods for obtaining information about travel patterns. The respondents will provide their exact travel information, such as the departure place, destination, departure time, arrival time, and the travel mode they adopted [1]. It is mainly completed through face-to-face interviews, questionnaires, telephone interviews, and E-mail [2]. Although the traditional traffic survey can directly obtain rich social population information, it also has obvious limitations which hinder the progress of the research on social travel patterns [3], such as high collection cost, limited sampling rate, long survey duration [4]. To overcome the shortcomings of traffic survey, traffic scientists have been trying to develop methods to automatically identify residents’ travel characteristics.

The Global Position System provides high-precision and time-stamped location data. Traffic scientists gradually begin to replace the traditional traffic survey with GPS data to obtain the travel characteristics of urban residents, thus overcoming the shortcomings of traffic survey [5]. Researchers have adopted abundant methods on travel
mode identification based on GPS data, such as rule-based methods, random forest [6], Bayesian network [7], Neural network [8], and other learning algorithms. Due to the excellent positioning accuracy and intensive collection frequency of GPS data, these methods have achieved considerable accuracy (mostly over 85%). However, GPS data still has limitations that restrict large-scale application. For example, GPS data collection requires volunteers to carry a dedicated GPS logger or active the GPS positioning module on their smartphone, which limits the scale and duration of such investigations [9].

With the continuous development of mobile communication technology, mobile phones have changed residents’ communication habits in recent decades. Cellular Signaling Data (CSD) is gradually coming into the field of traffic researchers. In a few years, the world mobile-cellular telephone subscription has raised to 8.2 billion subscribers in 2020 – corresponding to penetration of 133.4% in the developed world and 99.3% in developing world [10]. In other words, almost everyone owns a mobile phone. As an ancillary product used by mobile operators for communication billing [11], CSD has natural advantages such as large sample size, wide coverage of the population and long collecting duration [12], which can make up for the limitation of GPS data mentioned above. In recent years, CSD has been adopted in many traffic research, such as population density analysis [13], Home and workplace identification [14–16], mobility pattern analysis [17–19], commuting analysis [20], and can be used widely in mobile applications [21]. However, CSD has two major defects [9]: low positioning accuracy and unstable collection frequency, which lead to a big difference in data quality between CSD and GPS. Therefore, it is impossible to directly apply the travel mode identification methods based on GPS data to CSD. Overcoming the two major defects of CSD remains a major challenge on travel mode identification based on CSD.

Travel mode identification based on CSD generally requires three steps: data cleansing, trajectory segmentation and travel mode identification. In terms of data cleansing, due to the signal drift and other problems, the raw CSD may be offset too far from the actual location. Furthermore, the coverage between cellular towers has overlapping areas, which leads to the oscillation problem [22, 23]. That is, when a mobile phone stays at one location, it may be handed over to different cell towers due to load-balancing or other operational purposes. This results in traces suggesting that the mobile phone bounces between multiple stations. This is known as oscillation. [5, 24]. There are few papers to describe the method of data cleansing in detail. Moreover, most of the papers also rarely evaluate the effect of data cleansing [9]. In terms of trajectory segmentation, most of the papers mainly divide CSD trajectory points into two types: staying points and moving points. The moving points between two staying points constitute a moving trajectory. In terms of travel mode identification, three types of methods are mainly adopted: rule-based method, supervised learning and unsupervised learning. The challenge here is in developing algorithms that can differentiate more modes, especially those presenting similar speed profiles and following the same network. Moreover, due to the lack of ground truth data, there are limited papers to verify the algorithm of travel mode identification.

In summary, the main contributions of this paper are as follows:

1. A self-developed CSD acquisition APP is used to collect CSD and GPS data of volunteers at the same time. The volunteers calibrate the real travel mode and stay status of their CSD. Then the ground truth dataset is constructed to evaluate the effect of the data cleansing algorithm and travel mode identification algorithm this paper proposed.

2. A set of CSD cleansing algorithm frameworks is proposed. Combined with GPS data of volunteers, the effect of data cleansing is evaluated by using a variety of track matching indexes. The result shows that more than 95% of noise data are detected successfully. The similarity between CSD after cleansing and the actual trajectory according to DTW improved by 101.13% on average.

3. A multi-level travel mode identification algorithm based on CSD is proposed. The accuracy of the algorithm is verified through the ground truth dataset. The result shows that the proposed algorithms can achieve similar or even higher accuracy (increase by 10%) than the traditional supervised-learning classification algorithm without relying on the ground truth dataset for model training.

The remainder of this paper is organized as follows. Chapter 2 mainly describes the data adopted in this paper. Chapter 3 is the literature review of data cleansing, trajectory segmentation and travel mode identification. Chapter 4, chapter 5 are the methodology of data cleansing and travel mode identification proposed in this paper, respectively. Chapter 6 mainly analyzes the experiment results. Chapter 7 is the conclusion.

2. Data Description

2.1. Mobile Phone Network Data. When a mobile phone is turned on, it needs to connect to the nearby cellular tower to meet the communication requirement (such as making a phone call). Then a position record will be generated for the cellular tower that the mobile phone is connected to. The trigger mechanism for generating location records is mainly divided into events or network updates. Based on the trigger mechanism, Mobile Phone Network Data can be divided into event-driven Mobile Phone Network data and network-driven Mobile Phone Network data. For a more detailed description of the technologies and standards used to derive the position of mobile phones, see [5, 12, 25].

2.1.1. Event-Driven Mobile Phone Network Data. The data will be recorded when phone calls, text messages and Internet requests are made. This type of data is mainly used by
mobile operators to collect fees. This kind of data can be divided into two kinds:

(1) Call Detail Records (CDRs): CDR includes details of phone calls and text messages sent by mobile phones. The information recorded by different operators varies, but it generally records the encrypted phone number, the timestamp when the event occurred, the duration of the event, the type of the event and the unique number of cellular towers.

(2) Internet Protocol Detail Records (IPDRs): IPDR includes detailed documentation of the use of communication network. The information recorded by different operators varies, but it generally records the encrypted phone number, the timestamp of the event, the information about the website visited and the unique number of cellular towers.

2.1.2. Network-Driven Mobile Phone Network Data. To ensure the quality of communication services, mobile phone networks need to monitor the location of mobile phones, resulting in creating this kind of data. Network-Driven Mobile Phone Network Data (also known as cellular signal-data or sighting data) stores location update information of mobile phones, mainly including the following types of network event trigger:

(1) Power on and power off: In this case, the unique serial number of the cellular tower that the mobile phone connected to at the beginning or the last is recorded.

(2) Location area update (LAU): It is updated when a mobile phone moves to another cellular location area.

(3) Handover: It is recorded when a mobile phone moves from the coverage area of one cellular tower to another one when the mobile phone is making a call or using the communication network.

(4) When making or receiving phone calls and text messages, as well as when accessing the Internet: Similar to CDR, the encrypted phone number, the timestamp of the event and the unique serial number of the base station connected by the mobile phone are recorded.

(5) Regularly update location: when a mobile phone is on standby for a long time, that is, when none of the above events is triggered, the mobile operator will send a request for regular location update to obtain the unique number of cellular tower that the mobile phone is connected to.

Obviously, the recording frequency of Network-Driven Mobile Phone Network Data is higher than the Event-Driven Mobile Phone Network Data. A study of CDR shows that for 100,000 people over six months, the average interval of CDR is about 8.2 h [17]. Recently, with the continuous development of the Internet, more and more people are using the Internet to get information (such as through specific mobile phone apps). This makes the recording frequency of Network-Driven Mobile Phone Network Data increase significantly. For example, Chin Jiaqi found that the medians of the interval are about 90s in the network-driven dataset used [26].

In this paper, we adopted Network-Driven Mobile Phone Network Data as research data.

2.2. Ground Truth Dataset. To verify the effect of the algorithm, it is necessary to collect the ground truth dataset manually. In this study, a self-developed APP is used to collect both CSD and GPS simultaneously from volunteers. The GPS information can be obtained from the mobile phone positioning module. The CSD information can be obtained by accessing the real-time connected cellular tower information in the mobile phone. While collecting data, volunteers made label calibration on their actual stay status (staying or moving) and the travel modes (walk, bus, drive). The ground truth dataset is shown in Table 1.

Isdn is the anonymous unique number of the volunteer. Date is the recorded time. GPS_Link and GPS_lat are the GPS longitude and latitude of the volunteer. Cell_Link and Cell_lat are the longitude and latitude of cellular tower that the mobile phone connects to. Stay_status is the actual stay status (staying or moving). Mode is the adopted travel mode (Walk, Public Transport, Private Transport). The ground truth dataset is composed of 216 signaling trajectories collected by 5 volunteers, including 106 walking trajectories, 85 bus trajectories and 25 driving trajectories. The volunteers have agreed that the collecting data will be used only for experimental verification. The volunteers’ IDs will be anonymous.

2.2.1. GPS Data. The GPS data is only used for verification. The collection result shows that the sampling rate of 5 seconds can basically restore the volunteers’ travel trajectory accurately. So the GPS data is collected with a sampling frequency of 5 seconds. The positioning accuracy is about 2 to 20 meters. To ensure GPS data matches the actual trajectory, all GPS data has been artificially corrected by volunteers before verification and analysis.

2.2.2. CSD Data. The CSD data is collected with a sampling frequency of 5 seconds. The total number of CSD is 290118. In actual scenarios, CSD is usually very sparse due to the uncertainty in the communication between mobile phones and cellular towers. We conducted a statistical analysis of the CSD data in Foshan, China provided by mobile network operator. The data average sampling interval is 162 seconds, the average number of CSD points per user per day is 533, as shown in Figure 1.

To verify the performance of the proposed algorithms in actual scenarios, a resampling method is adopted. We randomly hit the CSD points under different sampling rates from 5% to 100% to simulate the actual sampling rate in actual CSD data. And the performance of algorithms under different sampling rates will be testified.
3. Literature Review

3.1. Data Cleansing. There are generally two kinds of noise data that need to be cleansed in CSD, which are oscillation data and drift data.

3.1.1. Oscillation Data. When a mobile phone is at the overlapping area of the cellular towers, CSD will appear to connect to several cellular towers successively rather than to a single cellular tower within a short period of time, even if the mobile phone is not moving. This phenomenon is known as the oscillation problem [23]. When the oscillation problem occurs, some devices will generate CSD that two (or more) cellular towers switch back and forth at a high frequency. An oscillation data example is shown in Table 2. The CSD switches between cellular towers CT1 and CT2 five times in a few minutes. Oscillation data produces a series of data that cannot reflect the actual moving trajectory. Since oscillation data is a common phenomenon in CSD [27], if oscillation data is not processed, subsequent analysis will be affected [5, 23]. The cleansing steps of oscillation data can be divided into two steps: the detection of oscillation sequence and the selection of equivalent position.

In terms of oscillation sequence detection, current research can be divided into three kinds of methods: Pattern-based, Speed-based and Hybrid-based method. Pattern-based mainly defines the switching model of oscillation sequence. For example, Li, G. et al. [22] only deals with the simplest model of oscillation data A-A-B-A-A (A and B represent cellular tower). In this case, B is removed and all remaining points A are merged to reduce data redundancy. Li, X. [23] further considers two switching models A-B-A-B-A and A-B-C-A to detect oscillation sequences. Bayir et al. [28] introduces the concept of switching times and believed that there must be three or more switches between two cellular towers in an oscillation sequence. A sequence A-B-A-C-D-C-B would be detected as an oscillation sequence because it switched between A and B three times. Since an oscillation sequence switches between multiple cellular towers in a short period, the speed of switching between them can be extremely high. Thus, another method of oscillation sequence is generated – Speed-based. Oscillation sequence is detected by artificial speed threshold through the speed-based method. Ivan et al. [29] set the speed threshold to 200 km/h to distinguish between normal sequence and oscillation sequence. The third method, Hybrid-based, combines the advantages of Pattern-based and Speed-based methods to detect complex oscillation switching patterns and reduces the risk of misjudging actual movement as oscillation data [27].

After detecting the oscillation sequence, it is necessary to obtain a suitable position to represent the position of an oscillation sequence. A common way is by counting the number of each point in the oscillation sequence and take the maximum number one as the equivalent position [28]. Another way is to build up a scoring method [23], which evaluates the frequency and average distance simultaneously. The point with the highest score will be selected as the equivalent position.

3.1.2. Drift Data. Surprisingly, we have not found any literature on cleansing drift data. According to our observation of the CSD given by the operator, drift data is a common phenomenon, which is necessary to be cleansed. In the following paper, we will describe the phenomenon of drift data in detail, and proposed a drift data cleansing algorithm based on spatiotemporal and frequency characteristics.

3.2. Travel Mode Identification. The existing algorithms could be mainly divided into three categories: those that are based on prior knowledge, unsupervised learning and supervised learning.

For prior knowledge-based methods, speed, distance and route are generally employed as the classified standards. Methods in this category focus on overcoming the low resolution of CSD position accuracy and aim to reduce the bias between the calculated speed and real speed. However, these methods mainly considered speed as the gist of splitting the travel mode, as it could not distinguish travel modes with similar speed features. Abdelaziz et al. [30–34] proposed to recognize travel modes by analyzing the signal intensity of GSM and the switching rate among cellular towers to obtain mobile phones’ travel speed. By combining with the road network data, Qu Y et al. acquired the individual travel trajectory, and thus, achieved the aim of recognizing the modes of private transport, public transport and walking [35].
For unsupervised-learning-based methods, considering the information from origin-destination (OD) of signaling trajectories, Wang H et al. classified the individual travel mode by employing the K-means algorithm to cluster the travel time information from signaling trajectories that have the same OD [36]. However, assigning modes to the resulting clusters is often done manually (i.e., by human interpretation) and might not be an easy task if differences between clusters are very small or do not match prior knowledge [9].

For supervised-learning-based methods, many advanced classification algorithms have been gradually published in the computer field [37, 38]. With the analysis of a cellular tower’s non-redundant number and average sojourn time in each cell during one trip, Wu W et al. employed the naïve Bayesian classification model to distinguish the modes of air track and walk [39]. Chin K et al. use the Random Forest algorithm to split travel mode into subway, train, bike, car and walk and reach an average accuracy of 73% [40].

Due to the difficulty of collecting a large amount of ground truth data, it limits the application of supervised-learning classification algorithms in large-scale practical scenarios. Therefore, we mainly focus on researching travel mode identification method which not rely on ground truth data. The ground truth dataset is only used for validation and feature selection.

### 4. Data Cleansing

Since trajectory segmentation is not the focus of this paper, we directly extract trajectories for data cleansing and travel mode identification. The raw CSD contains a variety of noise data, which wrongly representing the mobile phone’s location. It is necessary to detect and recorrect noise data before further analysis. In this paper, we propose a cleansing algorithm framework to detect two main kinds of noise data in CSD, drift data and oscillation data. Drift data will be detected and cleared first, followed by ping-pong sequence identification and correction.

#### 4.1. Drift Data

Drift data refers to the CSD generated when the cellular tower connected by the mobile phone has a large distance deviation from the actual location (generally 2 km to 10 km or even farther) due to the signal drift. In order to distinguish whether a point is a drift point or a normal point, velocity is introduced to detect drift points. Therefore, the drift point is defined as the signaling point that has long-distance and high-speed moving from the previous signaling point. The schematic diagram is shown in Figure 2.

The solid black line is the actual moving trajectory of a mobile phone. Cellular tower A-B-C-D-E are the cellular tower connected by the user during the moving process (signaling trajectory). Among the cellular towers, the distance between C and D is large (generally larger than 3 km). The moving speed between C and D is also much higher than the normal movement speed of human beings (such as the road speed limit of 120 km/h). Therefore, point D is defined as a drift point. The moving trajectory will be recorrected from A-B-C-D-E to A-B-C-E.

Through the observation of CSD provided by mobile operators, it is found that there are generally two kinds of drift data, which are defined as simple drift data and complex drift data.

#### 4.1.1. Simple Drift Data

Simple drift is similar to that shown in Figure 2. After a brief drift occurs during the movement, it is connected back to the normal points. A data example and schematic diagram of simple drift data are shown in Table 3 and Figure 3.

#### 4.1.2. Complex Drift Data

Complex drift data is much more complicated than simple drift data. In the process of movement, the cellular towers mobile phone connected to has a long duration of continuous drift. During the drifting period, several cellular towers which are far from the actual position are connected. A data example and schematic diagram of complex drift data are shown in Table 4 and Figure 4.
As mentioned above, the basic idea to detect drift points is to analyze the distance and speed between two consecutive signaling points. However, CSD has the defect of unfixed collection frequency. From the CSD provided by the operator, the interval between two consecutive signaling points can vary from 1 second to 1 hour. Due to this defect, some drift points may be misjudged as normal points because the interval is too long. In this case, even if the distance is very long, the calculated velocity is still within the range of reasonable movement speed. Moreover, when a drift point is misjudged as a normal point, it will become a reference point to judge the next point whether is a drift point or not. Under this circumstance, when a drift point is misjudged as a normal point, the following normal points may likely be misjudged as drift points and wrongly eliminated.

In conclusion, the main challenges of detecting drift points are as follows:

1. How to overcome the problem of misjudging drift point as normal point caused by the defect of unfixed frequency of CSD.

2. How to overcome the problem that a series of normal points are misjudged as drift points because of the misjudgment of a drift point to a normal point.

In order to overcome the above two challenges, the frequency of signaling points is introduced as an additional consideration besides distance and speed. According to the observation of data provided by the operator, in general, signaling data will not drift multiple times for a single point, which means the frequency of drift points is far lower than normal points. Therefore, the idea of the drift point detection algorithm is firstly to count the frequency of each signaling point. And regarding the point with particularly prominent frequency as the absolutely normal point (ANP), that is, these points would not be detected as drift points. On one hand, the introduction of ANP can make sure that some of the normal points would not be misjudged as a drift point, thus overcoming challenge 2 to some extent. On the other hand, when it is found that the absolutely normal point is misjudged as a drift point on the process of detecting, it means the reference point causing the misjudgment. It is very likely that the reference point is a drift point but was misjudged as a normal point. In this case, the reference point will be recorrected to drift point, thus overcoming challenge 1 to some extent.

### 4.1.3. Drift Data Detection Algorithm

A schematic diagram of drift data detection algorithm is shown in Figure 5 (see Algorithm 1).

Normal point \( n_i \) represents the point identified as normal point by the algorithm. Drift point \( d_{i,j} \) represents the \( j^{th} \) drift point identified by the algorithm with the normal point \( n_i \) as the fiducial point. For example, drift point \( d_{1,2} \) means that this point is the second drift point identified by the algorithm with the normal point \( n_1 \) as the fiducial point. Unvalidated point \( u_i \) represents the points have not been judged by the algorithm. The example shown in Figure 5 is
a signaling trajectory (sorted by time), assuming that the current fiducial point is normal point $n_2$ and the next point is unvalidated point $u_1$. Then the algorithm steps are as follow:

Step1. Distance threshold $D$, speed threshold $V$ and quantity threshold are preset. Each point’s frequency of signaling trajectory is counted. Points whose frequency exceeds quantity threshold are identified as absolutely normal points, and the set composed of these points is set as $ANP$.

Step2. Calculate the distance $Dis$ and the speed $Vec$ between the fiducial point $n_2$ and the next point $u_1$ ($Dis$ is the linear distance between these two points).

Step3. Determine whether $Dis$ and $Vec$ are both greater than the pre-set threshold. Then the next point $u_1$ is further determined whether is in the set $ANP$.

Step3.1 If the next point $u_1$ is not in the set $ANP$, it will be identified as drift point $d_{2,1}$. The unvalidated point $u_2$ will be set as the next point.

Step3.2 If the next point is in the set $AP$, which means normal point $n_2$ is a drift point that was misjudged as a normal point due to the long interval. The normal point $n_2$ will be modified to drift point $d_{2,3}$. The normal point $n_1$ will be set as the fiducial point again. All drift points (in this case $d_{2,1}$) previously identified with the normal point $n_2$ as fiducial point will be modified to unvalidated point $u_2$, and the first point of these points (in this case $d_{2,1}$) will be set as the next point.

Step4. Repeat step2 and step3 until all signaling points are identified by the algorithm.

4.2. Oscillation Data. As mentioned above, due to load-balancing or a mobile phone being at the overlapping area of the cellular towers, CSD will appear to connect to several cellular towers successively rather than to a single cellular tower within a short period of time, even if it is not moving. This phenomenon is known as the oscillation problem [39]. As shown in Figure 6, the mobile phone is in the boundary of the coverage of multiple cellular towers. An oscillation sequence with points A, B and C switching is generated. As an example, an oscillation sequence is shown in Figure 7, which generates an oscillation sequence of ‘A-B-C-D-C-D-B-C-B-E’, while the actual trajectory is ‘A-B-E’.

The cleansing of oscillation data is mainly divided into two parts: the detection of oscillation sequence and the selection of equivalent position. In the aspect of detection of oscillation sequence, we proposed a stepwise search method based on time window. For the selection of the equivalent position, we proposed a clustering method based on the staying time at each cellular point in an oscillation sequence.
4.2.1. Oscillation Sequence Detection. A schematic diagram of oscillation sequence detection algorithm is shown in Figure 8 (see Algorithm 2).

The example shown in Figure 8 is a signaling trajectory (sorted by timestamp). L1, L2, L3, L4 and L5, respectively, represent different cellular towers. Then the algorithm steps are as follow:

Step1. A fixed time window $T_w$ is preset. L1 will be set as the fiducial point.

Step2. All signaling points within the time window started from the fiducial point will be extracted as S.

Step3. Find if the fiducial point L1 is present again in S.

Step3.1 If the fiducial point is present again. The last same fiducial point is taken as the last same point. The series of signaling points from the fiducial point to the last same point is identified as an oscillation sequence. Then the following point to the oscillation sequence will be set as the fiducial point.

Step3.2 If the fiducial point is not present again. The following point to the fiducial point will be set as the fiducial point.

![Figure 6: Schematic diagram of oscillation data.](image)

![Figure 7: An example of oscillation sequence.](image)
Step 4. Repeat step 2 and step 3 until the last point of signaling trajectory is set as the fiducial point.

4.2.2. Selection of Equivalent Position. After the detection of oscillation sequences, it is necessary to find an equivalent position for the sequence. As mentioned above, the generation of oscillation sequence is because the mobile phone is staying in the overlapping coverage area of multiple cellular towers. Therefore, the oscillation sequence implies the mobile phone’s position information at the coverage boundary of the cellular towers. Theoretically, the closer the mobile phone is to a cellular tower in an oscillation sequence, the stronger the signal the mobile phone receives from the cellular tower, the higher the proportion of staying time in the sequence. Therefore, this paper proposes a clustering method based on the duration at each signaling point in an oscillation sequence. The specific calculation equations are shown in Eq. 1 to Eq. 2 (see Algorithm 3).

\[
\ln g_{\text{mean}} = \frac{\sum \ln g_i \cdot \text{staytime}_i}{\sum \text{staytime}_i} \quad (1)
\]

\[
\text{lat}_{\text{mean}} = \frac{\sum \text{lat}_i \cdot \text{staytime}_i}{\sum \text{staytime}_i} \quad (2)
\]

5. Travel Mode Identification

After data cleansing and trajectory segmentation, the next step is travel mode identification. Each trajectory will be labeled with a single travel mode. In recent research, researchers usually directly identified fine-grained travel modes through a single algorithm or model based on spatial-temporal features. As mentioned above, CSD has the defect of low positioning accuracy which is relative to the distribution of cellular towers. Provided by the operator, the coverage of cellular tower is 200 – 500 m in urban areas while may achieve 800 – 1000 m in suburban areas. Low positioning accuracy can lead to bias between calculated spatial-temporal features and the actual spatial-temporal features. It will affect the final identification effect if the algorithm directly identifies fine-grained travel modes based on bias spatial-temporal features, especially those travel modes which have similar features such as public transport and private transport. Therefore, in this paper, we divide travel mode identification into two dichotomous problems. The first dichotomous problem is dividing trajectory into motor transport and non-motor transport. The second dichotomous problem is dividing motor travel into public transport and private transport. Each dichotomous problem will be designed with a specific algorithm to solve.

5.1. Motor and Non-motor Transport Identification. Compared with non-motor transport, motor transport usually has a longer travel distance and speed. As a result, there are obvious differences between non-motor transport and motor transport in spatial-temporal characteristics, even under the positioning accuracy which is only about 500 m. Therefore, in this paper, the clustering algorithm GMM based on spatial-temporal features is adopted to divide trajectory into motor transport and non-motor transport. The main steps of the algorithm are divided into two steps: feature engineering and clustering.

5.1.1. Feature Engineering. Common spatial-temporal features such as travel speed, travel distance and travel duration are initially adopted and divided into three classes: Distance, Time and Speed. Then Random Forest algorithm is adopted to select features with strong correlation. The preliminary selected features are as follows:

- Distance: CSD cannot reflect the actual travel distance due to the defect of low positioning accuracy. However, each signaling point still represents a location area. Therefore, travel distance still can be represented by the adjacent position relationship between signaling points. In this paper, accumulative distance, OD distance and the number of unique cellular towers. A schematic diagram of distance is shown in Figure 9.

   (1) Accumulative distance

   As shown in Figure 9, the two solid points are the origin and destination cellular towers of the trajectory. The hollow points are moving points which represent the location of cellular towers. A solid line \( l_i^{CT} \) represents the Euclidean distance between consecutive signaling points. Accumulative distance is the sum of all solid lines \( \sum l_i^{CT} \), which \( i \) represents the number of solid lines. Take Figure 9 as an example, \( i = 9 \).

   (2) OD distance

   OD distance is the Euclidean distance between origin and destination. Since there are oscillation sequences and drift data in raw CSD, and the cleansing algorithm cannot eliminate all noise data. Oscillation sequence or drift data may bring bias to the calculation of Accumulative distance.
Therefore, OD distance is introduced to eliminate the bias caused by noise data.

(3) The number of unique location points

The number of unique location points can represent the travel distance to some context. Theoretically, the longer the travel distance is, the more cellular towers to connect to. Therefore, this feature is introduced as one of the features of distance. Take Figure 9 as an example, the number of unique location points is 10.

Time: The travel duration of the trajectory.

Speed: There are three types of speeds the correspond to three kinds of distances mentioned above: Accumulative speed, OD speed and Switching rate of cellular towers.
(1) Accumulative speed $V_{CT}$

Accumulative speed is calculated by Eq.3.

$$V_{CT} = \sum \frac{d_{CT}^i}{t_D - t_O}$$  \hspace{1cm} (3)

$\sum d_{CT}^i$ represents the accumulative distance, $t_D - t_O$ represents the travel time.

(2) OD speed $V_{OD}$

OD speed is calculated by Eq.4.

$$V_{OD} = \frac{l_{OD}}{t_D - t_O}$$  \hspace{1cm} (4)

$l_{OD}$ represents the OD distance, $t_D - t_O$ represents the travel time.

(3) Switching rate of cellular towers $V_N$

Switching rate of cellular towers is calculated by Eq.5.

$$V_N = \frac{N}{t_D - t_O}$$  \hspace{1cm} (5)

$N$ represents the number of unique location points, $t_D - t_O$ represents the travel time.

In this paper, the Random Forest algorithm is adopted as the feature selection method. The ground truth dataset is used as the sample data to extract features with strong correlation from the 7 features mentioned above.

5.1.2. Clustering. Compared with K-Means, GMM is essentially a density estimation algorithm. GMM introduces the concept of probability, which can effectively deal with the data in the overlapping region of the cluster. The cluster can also be divided into elliptical shapes to adapt to the different distribution of all kinds of data. Therefore, in this paper, the GMM is adopted to cluster the signaling trajectory based on the features selected by the Random Forest algorithm.

5.2. Public and Private Transport Identification. The main difference between public and private transport is the route and the duration. In general, considering the situation with the same OD, compared with private transport, public transport normally has a fixed travel path. Moreover, walking to a bus stop, waiting for a bus at the bus stop and stopping at each bus stop will all lead to the increase of the public transport duration. Therefore, in the case of the same OD trip, public transport usually has a longer travel distance and longer travel time than private transport. On the basis of the above cognition, the travel route and travel time are adopted as the main indicators to distinguish public and private transport. Private and public transport identification can be divided into three steps: navigation data request, trajectory matching and comparison of travel time.

5.2.1. Navigation Data Request. With the continuous development of electronic map, more and more people rely on electronic navigation (such as Google Map). Through the path planning API provided by the electronic map operator, in the case of a given specific OD, all the possible public transport plans and their corresponding travel routes and estimated travel time can be obtained. In the same way, the most recommended private transport plan and its corresponding travel route and estimated travel time can be obtained, too. The schematic diagram of public transport plans under the given OD provided by path planning API is shown in Figure 10.

5.2.2. Trajectory Matching. After requesting the navigation data, all possible public transport plans can be obtained. By evaluating the similarity between each possible public transport path and signaling trajectory, we can analyze whether the signaling trajectory is likely to be public transport or not. In this paper, a trajectory matching method is proposed to evaluate the similarity between the public transport path and signaling trajectory.

The matching between signaling trajectory and navigation public transport path is essentially the matching between each signaling point and navigation public transport path. The similarity between point and trajectory is based on the distance between point and trajectory, which is defined as shown in Figure 11.
\( p \) is a signaling point. \( A \) is a possible travel path. \( a_i \) represents position point on \( A \). \( d_i \) is the distance between the position point \( p \) and \( a_i \). Then the distance \( D \) between \( p \) and \( A \) is measured by the distance between \( p \) and the nearest point \( A \) to \( p \).

\[
D(p, A) = \min_{a_i \in A} d(p, a_i) = d_i
\]  

Considering that each cellular tower has a certain radiation range, distance threshold \( D_{max} \) can be set to determine whether it is possible for the signaling point in signaling trajectory to receive the signal of the public transport path. If the shortest distance between a signaling point and public transport path is bigger than \( D_{max} \), that is, the cellular tower could not receive the signal of the public transport path.

Considering the interference of noise data, when 80% of the shortest distance between signaling points and public transport path are less than \( D_{max} \). Then the public transport path is considered as a possible travel trajectory, the matching is succeeded. Otherwise, the matching is failed. A schematic diagram is shown in Figure 12.

\[ D(P, A) = D_{P, A}(p_i, A) \rightarrow \begin{cases} 
\text{match succeeds,} & \text{if } 80\% \text{of } D(P, A) \leq D_{max} \\
\text{match fails,} & \text{otherwise}
\end{cases} 
\]  

(7)

If all the possible public transport paths could not match the signaling trajectory, then this trajectory will be identified as private transport. Considering that the route of driving can be more random and unpredictable, public and private transport may have the same travel path. Therefore, if some of the possible public transport paths match the signaling trajectory, travel time will be further compared with the public transport estimated travel time and private transport estimated travel time provided by the API.
5.2.3. Comparison of Travel Time. As mentioned above, in the case of the same OD trip, public transport usually has a longer travel distance and longer travel time than driving a car. Therefore, the estimated travel time of public and private transport are introduced as a reference. By comparing them with the travel time recorded by the CSD, the successful matching public transport routes are further determined as possible travel paths. If the estimated travel time of all matching possible public transport plans $T_{public}$ is not closer to travel time than the estimated travel time $t$ of the private transport plan $T_{private}$, then this trajectory will be identified as private transport. Otherwise, it is public transport.

$$\Delta t_{private} = t - T_{private}$$
$$\Delta t_{public} = |t - T_{public}|$$

$$\begin{cases} 
\Delta t_{private} = |t - T_{private}| & \text{public transport, if } \Delta t_{public} \leq \Delta t_{private} \\
\Delta t_{public} = |t - T_{public}| & \text{private transport, if } \Delta t_{public} > \Delta t_{private}
\end{cases}$$

6. Experiment

6.1. Data Cleansing. The evaluation of data cleansing is mainly divided into three parts, which are the evaluation of drift data, oscillation data and overall cleansing effect. Taking a single trajectory as an example, the noise data is artificially calibrated by comparing CSD with GPS. The calibration result is shown in Figure 13. The brown line is GPS trajectory. The blue points are normal points. The red diamond points are oscillation sequences. The green diamond points are simple drift points. The orange diamond points are complex drift points.

6.1.1. Drift Data. By comparing CSD with GPS, a total of 9 drift points is artificially calibrated, including 4 simple drift points and 5 complex drift points. The drift data cleansing algorithm involves three preset thresholds: distance, speed and frequency. Based on the analysis of signaling data and characteristics of urban travel, the distance threshold is set to 2000 m, the speed threshold is set to 120 km/h. The setting of the frequency threshold is related to the time span of signaling data. Theoretically, the longer the time span is, the higher the threshold setting should be. Since the cleansing target is a single trajectory, the frequency threshold is set to 3.

After cleansing the drift data with the algorithm proposed in this paper, the cleansing result is shown in Figures 14 and 15. 4 simple drift points and 3 complex drift points are cleansed. There are still 2 complex drift points that not be cleansed. The main reason is that the time interval between the normal point and drift point is too long, even though the distance between them is bigger than 2000 m, the calculated speed is less than 120 km/h. The frequency of CSD collection will affect the drift data cleansing algorithm.

As mentioned above, drift data cleansing has the risk of misjudging normal points as drift points, resulting in the accidental deletion of normal points (challenge 2). To avoid this situation, we introduce the concept of ANP. To test the
validity of this method, we compare the retention of normal points with or without ANP. The test result is shown in Figure 14, which shows that there are 103 normal points in the original trajectory. Only 8 normal points remain after cleansing without ANP, while 98 normal points remain with ANP. The test result shows that the introduction of ANP can effectively avoid deleting normal points by mistake.

6.1.2. Oscillation Data. By comparing CSD trajectory with GPS trajectory, a total of 59 oscillation points is artificially calibrated. The oscillation sequence detection algorithm involves the preset threshold of the time window. By analyzing the characteristics of oscillation sequences, we set the time window threshold to 5 minutes. The evaluation of oscillation data cleansing is mainly divided into two parts. The first part is the detection effect of oscillation sequences, that is, how many oscillation points are recognized. The second part is the evaluation of equivalent position points. The test results show that the oscillation sequence detection algorithm based on the time window proposed in this paper successfully detects 57 oscillation points. To further verify the proposed algorithm’s performance, the pattern-based, speed-based, and hybrid-based methods are also adopted to be compared. In terms of pattern-based, oscillation patterns are set to A-B-A and A-B-C-A. In terms of speed-based, the speed threshold is set to 120 km/h. In terms of hybrid-based, oscillation patterns are set to A-B-A and A-B-C-A and the threshold is set to 120 km/h. The comparison result is shown in Figure 16. The result shows that the proposed method has a better performance in oscillation sequence detection.

In terms of equivalent position points, as shown in Figure 17, the red diamond point is an oscillation sequence identified by the algorithm in this paper, and the green diamond point is the equivalent position obtained from the oscillation sequence. It can be observed that the equivalent position is more consistent with the GPS trajectory than the oscillation sequence.

6.1.3. Overall Cleansing Effect. The purpose of data cleansing is to make the signaling trajectory more similar to the GPS trajectory. The trajectory similarity index is introduced to evaluate the overall effect of data cleansing. The main purpose is to calculate the trajectory similarity index by comparing the GPS trajectory with the signaling trajectory before and after data cleansing, respectively, to determine whether the signaling trajectory after cleansing is more similar to the GPS trajectory. To reflect the degree of optimization, we also calculated the ideal trajectory similarity index, which is the similarity between GPS trajectory and signaling trajectory whose noise data is all artificially deleted. DTW (Dynamic Time Warping) is selected as the evaluation index of trajectory similarity [41]. The calculation method is shown in Eq.9. The smaller the calculated DTW is, the more similar the two trajectories are. To further describe the improvement degree of trajectory similarity, we define an improvement coefficient IC. The equation is shown in Eq.10. The greater the calculated IC is, the more the

Figure 17: Schematic diagram of equivalent position of oscillation sequence.

Figure 18: Cleansing result of DTW.
improvement degree is.

\[
DTW(A, B) = \begin{cases} 
0, & \text{if } n = 0 \text{ and } m = 0 \\
\infty, & \text{if } n = 0 \text{ or } m = 0 \\
 d(\text{Head}(A), \text{Head}(B)) + \min & (\text{DTW}(A, \text{Rest}(B)), \\
\text{DTW}(\text{Rest}(A), B), \\
\text{DTW}(\text{Rest}(A), \text{Rest}(B))) 
\end{cases}
\]

\[
IC = \frac{DTW_{\text{before cleansing}} - DTW_{\text{after cleansing}}}{DTW_{\text{before cleansing}} - DTW_{\text{ideal}}} \times 100\% \tag{10}
\]

The result is shown in Figure 18. After data cleansing, DTW decreases compared with that before cleansing, indicating that the trajectory after data cleansing is more consistent with the GPS trajectory. The improvement coefficient IC reaches 85.06%, which shows a great improvement in trajectory similarity.

6.1.4. More Experiment. To further verify the effectiveness of the algorithm, we further calibrated the noise data for 10 trajectories and evaluated using the aforementioned indexes. The result is shown in Table 5.

<table>
<thead>
<tr>
<th>ID</th>
<th>evaluation of drift data cleansing</th>
<th>evaluation of oscillation sequence cleansing</th>
<th>DTW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before cleansing</td>
<td>after cleansing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simple drift point</td>
<td>Complex drift point</td>
<td>Normal point</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>14</td>
</tr>
</tbody>
</table>

Summary: Average drift data cleansing rate: 88.06% average normal points retention rate: 97.54% average oscillation sequence detection rate: 93.08% average IC: 101.13%

6.2. Travel Mode Identification

6.2.1. Motor and Non-motor Transport Identification. As mentioned above, the Random Forest algorithm is used to select the high-relative features. Ten-fold cross validation is adopted and the average correlation degree of features obtained by the Random Forest algorithm is shown in Figure 20.

The three most relevant features are selected as the clustering features, which are accumulative distance, the number of unique cellular towers and accumulative speed. According to prior knowledge, the travel distance of non-motor transport is usually limited, while the travel distance of motor transport can be long or short. We assume that the temporal and spatial characteristics generated by CSD can be divided into the data generated by Gaussian distribution corresponding to three travel modes: the short-distance non-motor transport, short-distance motor transport and long-distance motor transport. Therefore, the cluster number of GMM is set to 3, which corresponds to three types of transport, respectively. The commonly used evaluation indexes of machine learning are used for evaluation, which are accuracy, precision, recall and F1 Score. The clustering result is shown in Table 6.

The result shows that the overall accuracy of the GMM clustering method can reach 94%, which verifies the hypothesis. We further apply traditional supervised classification algorithms to compared with GMM, including Logistic...
Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and BP Neural Networks (BP). Each algorithm’s parameters are optimized by grid searching, and ten-fold cross validation is adopted to test the accuracy. The test result of each method is shown in Figure 21.

The result shows that GMM can achieve the same effect as the traditional supervised-learning algorithms without the need of ground truth dataset, which is more suitable for large-scale scenarios analysis.

To verify the performance of the proposed algorithms in actual scenarios, we randomly hit the CSD points under different sampling rates from 5% to 100% to simulate the actual sampling rate in actual CSD data. The proposed method performance under different sampling rates is shown in Figure 22. The proposed method still performs well under the sampling rate of 5%, which is close to the general data size.

6.2.2. Public and Private Transport Identification. First of all, the distance threshold $D_{max}$ for trajectory matching needs to be preset. Considering that the coverage range of a cellular tower is generally 300-500 m in urban areas and 800-1000 m in suburbs, $D_{max}$ is set to 800 m. The commonly used evaluation indexes of machine learning are also used for evaluation, which are accuracy, precision, recall and F1 Score. The identification result is shown in Table 7.

### Table 6: The clustering result of Motor and non-motor travel identification.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Mode-identified</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motor</td>
<td>108</td>
<td>2</td>
<td>98.2%</td>
</tr>
<tr>
<td>Non-motor</td>
<td>11</td>
<td>95</td>
<td>89.6%</td>
</tr>
<tr>
<td>Precision</td>
<td>90.8%</td>
<td>97.9%</td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

We test the accuracy of the algorithm under different distance threshold $D_{max}$. The test result is shown in Figure 23. It can be seen that under different $D_{max}$, the overall accuracy of the algorithm is stable at about 80%. The algorithm reaches the highest accuracy of 83.6% when $D_{max}$ is set to 800 m. The identification accuracy of trajectories under different OD distance is shown in Figure 24. With the increase of OD distance, the difference between public and private transport becomes more obvious, the recognition accuracy tends to rise.

Similarly, we further apply traditional supervised classification algorithms to compared with the proposed method,
including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and BP Neural Networks (BP). Each algorithm’s parameters are optimized by grid searching, and ten-fold cross validation is adopted to test the accuracy. The test result of each method is shown in Figure 25.

Due to the high similarity of spatial-temporal characteristics between public transport and private transport under CSD positioning precision, the classification effect of each supervised learning method is limited. The accuracy can only reach about 75%. In contrast, by considering the travel route and travel time, the method proposed accuracy in this paper increased by 10%, which has a better identification effect.

To verify the performance of the proposed algorithms in actual scenarios, we randomly hit the CSD points under different sampling rates from 5% to 100% to simulate the actual sampling rate in actual CSD data. The proposed method performance under different sampling rates is shown in Figure 26. The proposed method still performs well under the sampling rate of 5%, which is close to the general data size.

### 7. Conclusion

In this paper, we propose a travel mode identification framework, including data cleansing and travel mode identification. The framework is verified by a ground truth dataset. In terms of data cleansing, the algorithms this paper proposed can successfully detect more than 90% of the noise data including oscillation sequences and drift data. According to DTW and the improvement coefficient, the similarity between CSD after cleansing and the actual trajectory improved by 101.13% on average. In terms of travel mode identification, the result shows that the proposed algorithms...
can achieve similar or even higher accuracy than the traditional supervised-learning classification algorithm without relying on the ground truth dataset for model training (94% in motor and non-motor transport identification, 83.5% in public and private transport identification). In terms of public and private transport identification, the accuracy of the method this paper proposed increases by 10% compared with traditional supervised-learning algorithms. Furthermore, verification of each algorithm under different sampling rates from 5% to 100% is adopted to simulate the algorithm performance in actual CSD. The results show that under the actual sampling frequency, the algorithms this paper proposed can still perform well. The framework can be easily implemented and applied to the analysis for large-scale populations scenarios.

Data Availability

The cellular signaling data used to support the findings of this study have not been made available because of data privacy and protection.

Conflicts of Interest

The author(s) declare(s) that they have no conflicts of interest.

Acknowledgments

We acknowledge the financial support provided by the National Key R&D Program of China under Contract No. 2020YFB1600400, the National Natural Science Foundation of China under Contract No. 12002403 and the National Natural Science Foundation of China under Contract No. U1811463.

References


