Shan et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:215
https://doi.org/10.1186/s13638-019-1535-9

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

# Modeling and implement of mobile phone user location discrimination based on heuristic strategy

Qingchao Shan[1], Honghui Dong[1], Limin Jia[1*], Hua Yuan[2] and Hui Zhang[3]

**Abstract**

With the all-pervading mobile devices and continuing advancement of big data technologies, mobile phone data research has been gaining widespread popularity in the past few years. Dealing with the implausible location caused by cell handover phenomenon in the communication system is one key problem of user mobility profile building based on mobile phone call detail records (CDRs) data. In this paper, we propose a location discrimination model aiming at CDRs data, where heuristic strategies for the characteristic of the oscillation phenomenon from practical CDRs and handover categories are added to distinguish the stay points, passing points, and oscillating points. A whole month of CDRs data from one communication operator is employed to select parameters and validate the model on the Spark platform. The experiment results betray that the proposed model can identify the false locations effectively. Compared with the threshold models, the result of the proposed model is more reasonable both in the population aggregate level and individual level. Besides, the model can retain more user's trajectory points than clustering algorithm, so it can improve the quality of user mobility modeling.

**Keywords:** Location discrimination, Heuristic strategy, Handover, Mobile phone data, Cell oscillation phenomenon, Big data mining

## 1 Introduction

Mobile phones have become ubiquitous devices in the world. Each mobile phone connecting to the cellular network (GSM, CDMA, GPRS, UMTS, LTE, and so forth) generates digital traces which may serve as representative of traffic indicators and human behavior. Call detail records (CDRs) from mobile phones contain spatial-temporal of anonymized subscribers. Since CDRs are automatically collected by cell phone carriers for billing purposes, compared with data from traditional manual survey, mobile phone data have advantages in effortless collection, large-scale data, wide coverage, low collection cost, and good real-time performance, which have been studied largely in the transportation field, but it causes many challenges to be studied in depth [1, 2]. González et al. [3] analyzed human movement pattern based on mobile phone base station data and concluded that the individual trip mode tends to be in single spatial probability distribution. Song et al. [4] proved that the upper limit of human trip behavior predication accuracy can reach 93% based on trajectory entropy formed by mobile phone data. In recent years, mobile phone data has been widely adopted to study people's movement characteristics [5–7], activity characteristics [8], traffic demands [9], origination destination (OD) features [10], traffic flow [11], disease prevention, and so on. In addition, technologies related to big data provide important technical support for the research and application of massive mobile phone data.

Despite these advantages, handover problem is very challenging because the mobile device's real location is not known. In order to ensure that mobile station is connected to communication base station while it is stationary or moving, necessary handover should be implemented in mobile communication system [1, 12]. However, due to dynamic changes in signal strength and various transmission conditions, significant noise can be observed in CDRs data. Geographical environment and

* Correspondence: jialm@vip.sina.com
[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
Full list of author information is available at the end of the article

buildings impact signal transmission, and overlap and load balancing in cellular signal coverage areas cause a mobile phone to be assigned to multiple base stations. Even if the mobile phone is stationary, it will also be connected to different base stations instead of the nearest base station, causing false displacement [2, 12]. In order to ensure the quality of network services during movement, handover will be more frequent. In addition, the layout of base stations will also affect the handover [12]. Handover is an important source of CDRs data problem. The aim of this paper is to detect and alleviate trajectory distortion problem by heuristic strategies, so that the accuracy of path reconstruction can be improved in the position data provided by mobile operators.

The remainder of the present study is structured as follows: in Section 2, handover issue and related resolutions of previous research on mobile phone data are summarized. In Section 3, data processing methods are introduced, then a detailed introduction is presented to explain the framework of the proposed model, and the parameters selection. In Section 4, the analysis results of some users' trajectories obtained through this model are shown at the individual and aggregate level and compared with those obtained based on other methods. In Section 5, the conclusion is given.

## 2 Related work

A key issue in the application of call detail records (CDRs) data to user mobility modeling is the handling of false location caused by signal handover and caused oscillation, which has a great impact on the accuracy of study results. Signal handover between mobile phones and base stations is a very common phenomenon in the field of cellular communications and is also an important technology [1, 12].

### 2.1 Classification and characteristics of handover

Common handover in communication network can be roughly divided into four categories: intra-cell handover, handover between different cells in the same base station controller (BSC), handover between different BSCs in the same mobile switching center (MSC) and between different MSCs. Handover in CDRs data is also mainly divided into four categories [1, 12]:

1. Ping-Pong handover: frequent handover of sequences of a user between two adjacent consecutive base stations. Letters are adopted to represent base station number (the same below), and the handover mode is {ABA}, {ABAB...}, and the like.
2. Back handover: handover of sequences of a user failing back to the stations where the sequences switch out during moving. The handover mode is usually {ABA}, {AB...A}. This mode is similar to Ping-Pong handover, except that the former passes through one or more different base stations and then switches back to the base station where it switched out before, and the latter switches between only two base stations.
3. Multipath handover: the situation where the starting and ending base stations of some segments in the sequences of a user are the same, but the sequences pass through different base stations during the starting and ending base stations, and the handover modes show multiple paths such as {AB...D}, {AC...D}.
4. Handover location fluctuations: fluctuation of location and time differences obtained by multiple measurements within a certain range when sequence handover in stations are the same.

### 2.2 Cell layout

The coverage radius of early communication base station signals was large, which was usually several kilometers, even more than 10 km in sparsely populated areas, so the handover was not very frequent. With proliferation of mobile phone users and the increasing demand for network capacity, "cell splitting" technology is adopted to achieve capacity expansion in communication system and microcell is emerged with coverage radius of several hundred to several tens of meters, which causes frequent handover. In general, the communication network assigns fast-moving mobile phones to a macro-station, and slow-moving mobile phones to a micro-station. In this way, a mobile phone may frequently cross the cell during a call [12]. In addition, there are still wide differences between CDRs data for uncertain time granularity and spatial granularity greatly affected by the deployment density of the base station [2].

González and Song et al. did not discuss mobile phone data handover and oscillation. Some studies adopted global position system (GPS) data processing method (such as Kalman filter method) for mobile phone data, and many studies adopted time threshold [6, 8, 10, 13], distance threshold [6, 8, 9, 13], and speed threshold [13] for oscillation point discrimination, and some studies also additionally adopted rules [13] or clustering algorithms [8, 14, 15]for it. However, threshold methods cannot be adopted to effectively deal with handover outside the thresholds; the clustering algorithm will cause loss of trajectory points in effective trip discrimination, and the method based on time window ignores trips within the window time [9].

Fiadino et al. [16] presented a study on trajectory reconstruction, where they used a "Ping-Pong" suppression (PPS) method that ignores events where the device connects back to the previous cell within a predefined time

window (transformation of subsequence in the event history ABA→AB). Vajakas et al. [17] used a novel technique for improved "Ping-Pong" effect suppression by compensating for some cell shape distortions based on temporal cell-to-cell transit statistics.

In short, handover is not designed for the collection of traffic data, and false location caused by handover affects the accuracy of the study. Therefore, various factors should be considered for location discrimination.

## 3 Method

### 3.1 Data description

CDRs data: records-related data generated when a communication carrier directly provides services for situations where a mobile phone makes a call, sends and receives a text message and so on [2]. In this paper, modeling study is carried out by taking CDRs data of a certain operator in Beijing in February 2015 as an example. The data format is shown in Table 1. Field TIMESTAMP: UNIX timestamp, the total number of seconds from 00:00:00 GMT on January 01, 1970 to the time when data is acquired. LAC: location area code, location area; CELLID: cellular (also known as sector) unique code; TYPE: service type; USER: encrypted user flag, some code bits are replaced with * to protect user privacy.

### 3.2 Data processing flow

The data is filtered before modeling. For raw data, one file per hour, a total of 672 files, 972 GB. The number of records of all users in 1 month is counted, and the user records with only one location or users with more than 10,000 records in a day are removed. In this way, more than two million users are removed and 431 GB available data is remained, saving 55.66% of storage space. Then the User field is converted to a unique 64-bit integer to save storage space and speed up analysis. Again, in order to parallelize the trajectory on the big data platform, all trajectories of a single user are generated. The method is as follows: all data is grouped by user, and then sorted in ascending order of time after LAC and Type fields are deleted. Timestamp field is divided into two fields: firstTime and lastTime fields. For intra-cell handover and handover of different cells in the same BSC, a unique MergeID is specified due to the same location, and intermediate records are removed from continuous records with the same location, and the starting time and end time are recorded as the firstTime and lastTime. If there is only one record for this location, the firstTime and lastTime are the same. The number of records decreases from 9,926,992,567 to 4,941,805,282, a decrease of 50.22%. See Table 2 for comparison with the raw data.

**Table 1** Structure of CDRs data

| Timestamp | LAC | Cell id | Type | User |
|---|---|---|---|---|
| 1422756000 | 4526 | 39687 | 128 | 065e* |
| 1422756000 | 4140 | 57188 | 160 | 0ad3* |
| 1422756000 | 4132 | 29486 | 160 | f311* |

**Table 2** Comparison of data before and after pre-processing

| | Total amount (row) | User (person) | Size (GB) |
|---|---|---|---|
| Raw data | 17,908,663,174 | 29,861,129 | 972 |
| Processed data | 27,642,018 | 27,642,018 | 154 |

The data in this paper is stored on the Hadoop platform, and Scala language is adopted for data processing and analysis on the Spark platform (Version 1.6), and desktop GIS software for display. The data analysis process is shown in Fig. 1.

### 3.3 Model definition

The basic definition of user location discrimination is as follows:

$T^i_{(k)\text{firstTime}}$: the first recording time for user $k$ at base station $i$.

$T^i_{(k)\text{lastTime}}$: the last recording time for user $k$ at base station $i$.

$T^i_{(k)\text{stay}} \coloneqq T^i_{(k)\text{lastTime}} - T^i_{(k)\text{firstTime}}$ : the stay time of recording for user $k$ at base station $i$.

$T^i_{(k)\text{transfer}} = T^{i+1}_{(k)\text{firstTime}} - T^i_{(k)\text{lastTime}}$: the recording time for user $k$ for from base station $i$ to the next station $i+1$.

$D_{i,\,i+1}$: space distance between base station $i$ and $i+1$.

$v^k_{i,i+1}$: the average transfer speed of user $k$ between based station $i$ and base station $i+1$.

$T^i_{(k)d} \coloneqq t* + D_{i,i+1}/v_{i,i+1}$ : estimated transfer time between based station $i$ and base station $i+1$, $t*$ is a time compensation parameter.

$T^i_{(k)\text{remain}} \coloneqq T^i_{(k)\text{transfer}} - T^i_{(k)d} \;\; if\,(T^i_{(k)\text{transfer}} > T^i_{(k)d})$ : transfer remaining time.

$T^i_{(k)\text{speculate}} \coloneqq \beta T^i_{(k)\text{remain}}$: speculative stay time at base station $i$, $\beta$ is a scale parameter[0, 1].

$T^{i+1}_{(k)\text{speculate}} = (1-\beta) T^i_{(k)\text{remain}}$: speculative stay time at base station $i+1$.

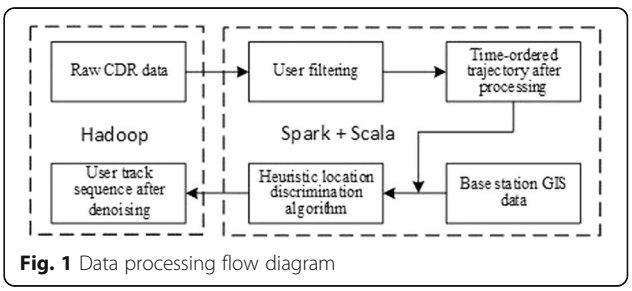


**Fig. 1** Data processing flow diagram

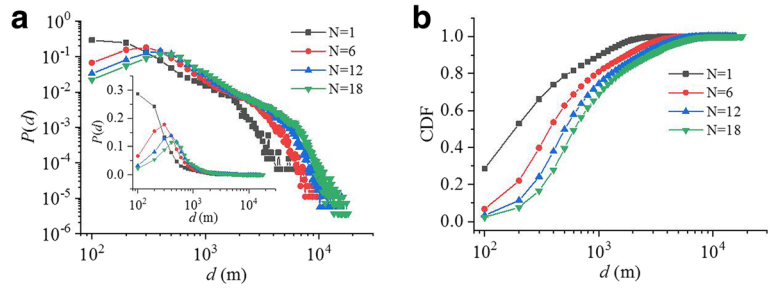


**Fig. 2** Distance distribution of base station and its nearest N neighbors. **a** Distance distribution. **b** Cumulative distribution

Stay point: If $T^{i}_{(k)\text{stay}} \geq \delta_{\text{duration}}$, the current location is an obvious stay point; otherwise, it is not an obvious stay point; when $T^{i}_{(k)\text{speculate}} \geq \delta_{\text{transfer}}$, the point with the longest stay time is selected as the speculative stay point, and the current point is taken as the speculative stay point by default. $\delta_{\text{transfer}}$ and $\delta_{\text{duration}}$ are time thresholds.

Note: stay in this paper is defined as the situation where the mobile phone does not switch to other locations within the specified time. Time is defined as recording time, and does not represent the actual arrival or departure time of user. Widhalm et al. [8] conducted related study on this issue.

### 3.4 Model heuristic strategies

The following heuristic strategies are proposed for different types of handover.

1. Ping-Pong handover processing: judgment of location relationship of pre- and post-sequence. If the sequence is a ABAB... type Ping-Pong handover sequence, the stay duration of Ping-Pong handover between two base stations and the number of communications are counted, and the station with obvious longer stay time was selected as the main base station of Ping-Pong handover sequence, and the firstTime and lastTime of main base station are changed to be the starting and end times of Ping-Pong handover sequence. If the stay time is the same, the base station with the most communication times is selected. If the handover is an ABA type handover, whether one of the two as is the stay point is judged (Wu et al. [13] required that both pre-A and post B are stay points). If $D_{\delta 1} < D_{ab} < D_{\delta 2}$ and $D_{\delta 1} < D_{\delta 2}$, $V_{ab}$, or $V_{ba}$ is larger than $V_{\delta}$, then B is an oscillation point, and ABA is merged into record of one A.
2. Back handover processing. The situation where the second A is an oscillation point in ABAC type sequence always occurs during user movement. The normal movement path of users is ABC, but the trajectory fails back to A from B to C. Since it is not as obvious as "Ping-Pong handover," it is easy to remove point B as a Ping-Pong handover, which is a difficult part to handle. With this algorithm, when no A is a stay point (including the speculative stay point), if $D_{\delta 2} > D_{ac} > D_{bc}$ and the time is less than $\delta_{\text{transfer}}$, the second A is an oscillation point.
3. The processing of moving sequence and Ping-Pong handover last sequence as ABA based on a triangle inequality with parameter $\lambda$: if $D^2_{ac} < \lambda * (D^2_{ab} + D^2_{bc})$ and $T_{ab} < T_{\delta}$ or $T_{bc} < T_{\delta}$, then point B may be an oscillation point, where $\lambda$ is the parameter with range (0, 2). $\lambda$ is set to 0.8 according to the road network and experience.
4. The non-Ping-Pong handover stay point is not an oscillating point.
5. Correct the triangle relationship to determine oscillation point. The sequence ABCD is taken as an example, if the i-1 and i-2 base stations are oscillation points in two consecutive judgments, let $L_{abd} = L_{ab} + L_{bd}$ and $L_{acd} = L_{ac} + L_{cd}$. If $L_{abd} >= L_{acd}$, B is the oscillation point; otherwise, C is an oscillation point.
6. Processing when the handover distance is greater than the threshold. ABC sequence is taken as an example. If B is not a stay point, the handover distance $D_{AB} > D_{\delta 2}$ and $v_{AB} > \delta_{v\max}$, then B is an oscillation point based on strategy (3).
7. Transfer time estimation. In general, the speed of movement is proportional to the distance, thus, we define a step-function which returns the transfer time given the distance $d$ to be covered:

**Table 3** Distance between base station and its *N* nearest neighbors

| N | Min (m) | Max (m) | Mean (m) | Standard deviation |
|---|---|---|---|---|
| 1 | 0.01 | 6118.64 | 375.85 | 505.31 |
| 6 | 0.01 | 12,391.39 | 692.06 | 861.70 |
| 12 | 0.01 | 15,823.22 | 940.52 | 1161.21 |
| 18 | 0.01 | 18,194.96 | 1138.45 | 1399.50 |

**Table 4** Distance distribution comparison of handover and Ping-Pang handover

| Distance | 1 km | 2 km | 3 km | 4 km | 5 km | 6 km | 7 km | 8 km |
|---|---|---|---|---|---|---|---|---|
| Handover | 46.75% | 71.37% | 82.09% | 87.38% | 90.15% | 92.0% | 93.1% | 94.1% |
| Ping-Pong | 61.95% | 83.35% | 90.59% | 93.57% | 94.90% | 95.7% | 96.2% | 96.5% |

$$T^{i}_{(k)d} = \begin{cases} t_0 & d \le 0.1\,\mathrm{km} \\ t_1 + d*60/v_1 & 0.1\,\mathrm{km} < d \le 1\,\mathrm{km} \\ t_2 + d*60/v_2 & 1\,\mathrm{km} < d \le 5\,\mathrm{km} \\ t_3 + d*60/v_3 & 5\,\mathrm{km} < d \le 10\,\mathrm{km} \\ t_4 + d*60/v_4 & 10\,\mathrm{km} < d \end{cases} \quad (1)$$

where $t_0$, $t_1$, $t_2$, $t_3$, $t_4$ are time offsets that could be set according to different scenarios; $v_1$, $v_2$, $v_3$, $v_4$are transfer speed according to different transfer distance.

The mobile phone location determined by the model strategy is the primary home base station. In practice, it can be processed in depth with algorithms such as centroid, weight, and clustering, which can be more realistic in some cases.

### 3.5 Model parameter determination

#### *3.5.1 Base station spatial distribution*

According to the theory that cellular layout is featured with regular hexagon [12], $N$ neighboring base stations that are closest to each base station are found, and $N$ is 1, 6, 12, or 18. Statistics is made to distance distribution $i$ at intervals of 100 m to the neighboring base stations, as shown in Fig. 2 ($P(d)$-PDF, probability distribution function, probability distribution, CDF-cumulative distribution function, cumulative distribution). They all have two power-law distributions with negative slopes in Fig. 2. The difference is that the $N = 1$ do not show positive slope of the first power law. The statistical results of range, mean, and standard variance of distance between the base station and its neighboring base stations are shown in Table 3.

#### *3.5.2 Handover distance threshold*

**3.5.2.1 Handover distance** Calculate all handover distances and perform statistics to the distances at an interval of 100 (excluding handover in the same location). Distance distribution is obtained as shown in Fig. 3 whose distribution is significantly different from the distribution of $N = 1$ in Fig. 2 and is significantly similar to the distribution of $N = 6$ or 12 in Fig. 2, proving that handover not always switches to the nearest base station.

**3.5.2.2 Ping-Pong handover distance** The distribution of Ping-Pong handover distance is obtained through the same handover distance method as shown in Fig. 3. Comparison of the handover distance cumulative distribution at 1 km, 2 km, 3 km, ... , 8km is shown in Table 4, which shows that Ping-Pong handover distance is closer than general handover. The handover and Ping-Pong handover thresholds are 5 km and 3 km, $D_{\delta 2}$ and $D_{\delta 1}$ are 5 km and 3 km respectively.

#### *3.5.3 Stay time threshold*

Determination of stay time threshold $\delta_{\mathrm{duration}}$ of $T^{i}_{(k)\mathrm{stay}}$: the stay time of the user at each base station is counted, wherein the time of 81.40% of location points is less than 1 min, and the time of only 18.60% of the location points is 2 min and above. The stay time distribution is shown in Fig. 4 (CCDF-complementary cumulative distribution function), an inflection point occurs near 10 min, so the significant stay time threshold is set to 10 min.

#### *3.5.4 Transfer parameter*

Handover of a mobile phone from one cellular to another cellular involves two important parameters: transfer speed and transfer time threshold. The speed limit of China's expressways is 120 km/h, and in the city, due to intersection and road speed limit, $V_{\delta}^{\max} = 80\,\mathrm{km/h}$ and the minimum speed is calculated based on walking, $V_{\delta}^{\min}$

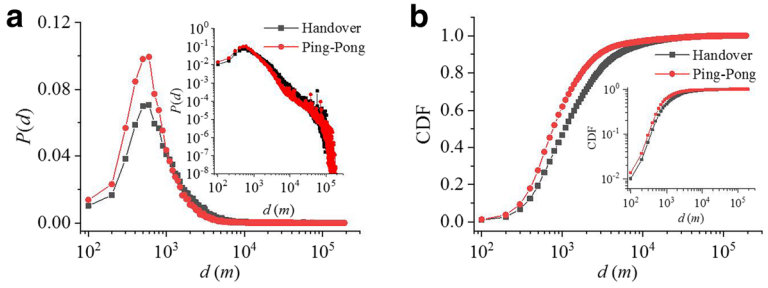


**Fig. 3** Distance distribution of handover. **a** Distance distribution. **b** Cumulative distribution

**Table 5** Records of one user trajectory

| ID | Merge ID | First time | Last time | Inter distance |
|---|---|---|---|---|
| 1 | 1825 | 17:49:40 | 19:28:32 | 838.26 |
| 2 | 1591 | 19:31:13 | 19:31:13 | 1714.03 |
| **3** | **1212** | 19:41:07 | 20:15:47 | 4602.64 |
| 4 | **1524** | 20:26:18 | 20:26:18 | 4048.61 |
| 5 | **1212** | 20:28:08 | 20:28:08 | |
| 6 | **1524** | 20:54:35 | **20:54:35** | |
| 7 | **1212** | **20:54:56** | **20:55:00** | |
| 8 | **1524** | **21:05:09** | **21:05:09** | |
| 9 | **1212** | **21:05:33** | 21:05:33 | |
| 10 | **1524** | 21:06:40 | 21:06:40 | |
| 11 | **1212** | 21:26:49 | **22:10:14** | |
| 12 | **1524** | **22:10:22** | **22:16:39** | |
| 13 | **1212** | **22:16:54** | 22:24:52 | |
| 14 | **1524** | 22:51:20 | 22:51:20 | |
| 15 | **1212** | 22:54:23 | **23:01:46** | |
| 16 | **1524** | **23:03:00** | 23:03:10 | |
| 17 | **1212** | 23:04:24 | 23:04:29 | |
| 18 | **1524** | 23:05:44 | **23:06:10** | |
| 19 | **1212** | **23:07:24** | 23:33:07 | |
| 20 | **1524** | 23:33:53 | **23:41:07** | |
| **21** | **1212** | **23:41:46** | 1:13:58 | |
| 22 | 1142 | 1:32:31 | 1:32:31 | 3080.62 |
| 23 | 1524 | 1:33:50 | 1:33:50 | 3951.39 |
| 24 | 1825 | 1:46:46 | 11:07:04 | 2376.10 |

Bold data records are tend to be misclassified by others methods

$= 3.6\,\mathrm{km/h}$ and $\delta_{\mathrm{transfer}} = \delta_{\mathrm{duration}}$. $t_0$, $t_1$, $t_2$, $t_3$, $t_4$ might resemble the average waiting time for different traffic mode and $v_1$, $v_2$, $v_3$, $v_4$ are average speed of pedestrian, bike, busses, cars, or taxi; respectively. We have chosen: $t_0 = 4$ min, $t_1 = 2.5$ min; $v_1 = 3.6$ km/h; $t_2 = 5$ min; $v_2 = 10$ km/h, $t_3 = 8$ min; $v_3 = 15$ km/h; $t_4 = 12$ min; $v_4 =$

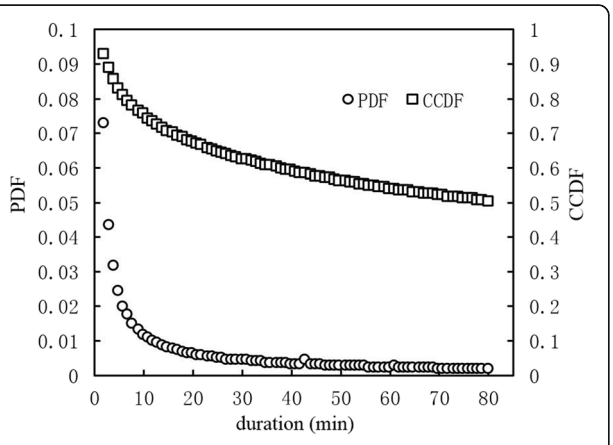


**Fig. 4** Distribution of stay time


**Fig. 5** Trajectory of user's records in Table 5

20 km/h accordingly to data of public transportation in Beijing city. $\beta$ is determined by the stay time $i$ and $i + 1$ base station or the number of visits.

## 4 Result and discussions

### 4.1 Individual level

#### *4.1.1 Ping-Pong handover*

The trajectory of the user shown in Table 5 by the dotted arrow in Fig. 5 is taken as an example to illustrate the advantages of heuristic rules. Location 1825 is taken as the starting and end point of a trajectory of the user. According to record 3 to record 21, Ping-Pong handover occurs continuously between 1212 and 1524 of this trajectory. According to common sense, it is less likely that users continuously visit two locations in a short period of time. According to the heuristic strategy (1), the base station 1212 with a long stay time is selected, and there is no trip between 1212 and 1524. Record 6 to record 9 shows that handover above 4 km is completed in a few seconds and the oscillation is obvious, which can be recognized based on speed threshold [8, 13]. When the distance threshold is set to be less than the Ping-Pong handover distance, the Ping-Pong handover will be incorrectly judged as trip. With 5 min and 3 km as the stay time and trip distance threshold, record 11 to record 13, record 15, and record 19 to record 21 are all stay points, so multiple trip behaviors between 1212 and 1524 are obviously not realistic [13]. If 10 min [6] and

**Table 6** Trajectory records of one user

| ID | MergeID | firstTime | lastTime | interDistance |
|---|---|---|---|---|
| 1 | 6393 | 12:59:56 | 12:59:56 | 583.53 |
| 2 | 5821 | 13:01:15 | 13:01:15 | 1745.15 |
| 3 | 6642 | 13:02:06 | 13:02:06 | 1364.90 |
| 4 | 6514 | 13:04:09 | 13:04:09 | 874.85 |
| 5 | 6611 | 13:08:55 | 13:08:55 | 619.66 |
| 6 | 6373 | 13:09:02 | 13:09:02 | 536.13 |
| 7 | 6640 | 13:09:32 | 13:09:32 | 453.66 |
| 8 | 6467 | 13:11:16 | 13:11:16 | 585.13 |
| 9 | 6357 | 13:11:57 | 13:11:57 | 566.37 |

15 min [8] are taken as thresholds, there is no stay between record 4 and record 10, and according to the strategy (1), the recording time is attributed to one site to generate a stay point. A method with a cluster radius of 1 km and a stay of more than 15 min will also generate multiple trips between such handovers [8]. The method of counting the most access point as the stay point in the time window will lose this trip [9].

To judge record 22, only considering the transfer speed, $v_{1142-1524} = 50.01$ m/s, namely 180.06 km/h, so there is not enough evidence to exclude this point. But based on the heuristic strategy (3), $L^2_{1212-1524} < (L^2_{1212-1142} + L^2_{1142-1524}) * 0.8$ (the choice of this parameter is somewhat subjective. The Pareto principle is referred to based on the road network layout and some actual conditions), and point 1142 is not a trajectory point.

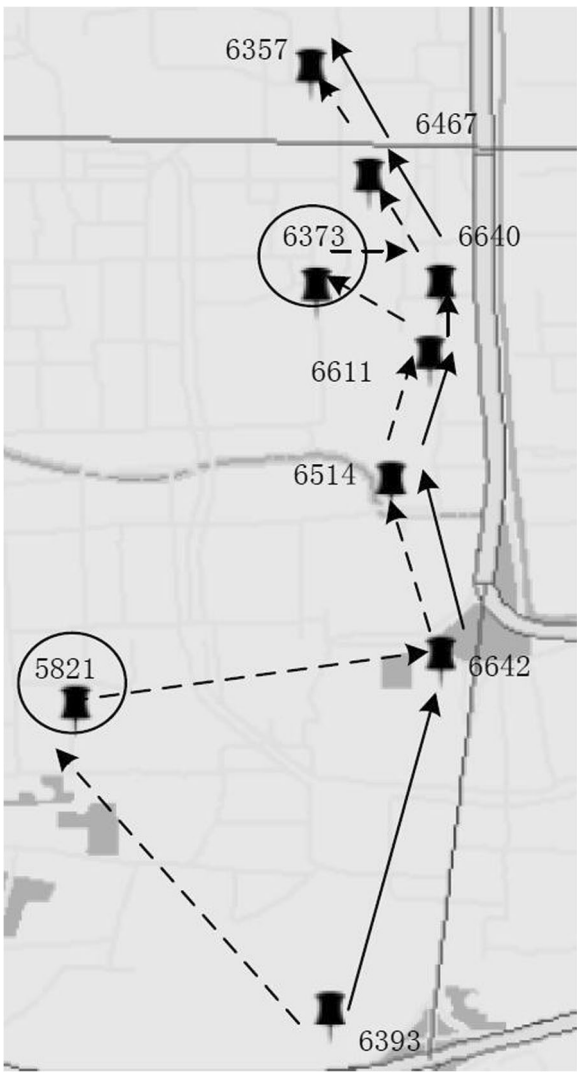


**Fig. 6** Trajectory of user's records in Table 6

#### 4.1.2 Mobility trajectory recognition

A trajectory of the user in Table 6 is shown by the dotted arrow in Fig. 6. According to the heuristic strategy (6) and speed threshold, trajectory point 5821 is excluded. However, the user path from 6611 can be trajectory 1: 6611 → 6640 → 6467, or trajectory 2: 6611 → 6373 → 6467. It is a discrimination difficulty to select a path. When the trajectory jump to 6640, 6373 is judged as an oscillating point, but when it is judged to 6467, 6640 is also judged to be an oscillating point. According to strategy (5), trajectory 1 is selected. The solid lines shown in Fig. 6 are the final trajectory obtained by the model. According to the clustering algorithm [8], 6611, 6373, and 6640 are generated as cluster centers, and other points within the cluster radius are merged, and many trajectory points cannot beretained.

### 4.2 Aggregate level

#### 4.2.1 Average daily trip times

A trip with distance between two successive stay locations exceeding the distance threshold is considered as a valid trip. The time threshold is set from 2 min, and counted to be 80 min at an interval of 1 min. The distance threshold is 1 km, and the transfer speed is 3.6 km/h. Then the average daily trip times are shown in Fig. 7 according to stay time threshold. Figure 7 shows that the threshold method combined with time or distance or time and distance under appropriate threshold can also be adopted to obtain average daily trip times in accordance with the traffic survey, but it is not applicable to some individuals. For example, Iqbal et al. [10] treated the oscillations with a threshold of only 10 min, which is somewhat simple. We also get speculated stay locations except obvious stay locations.

#### 4.2.2 Stay time characteristics

Stay time characteristics are analyzed by mainly studying the statistical characteristics of the user's stay duration,

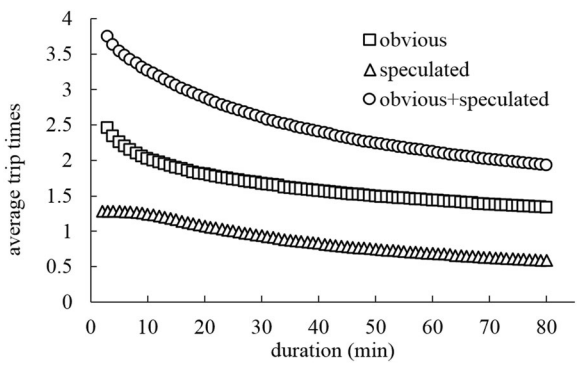


**Fig. 7** Average daily trip times

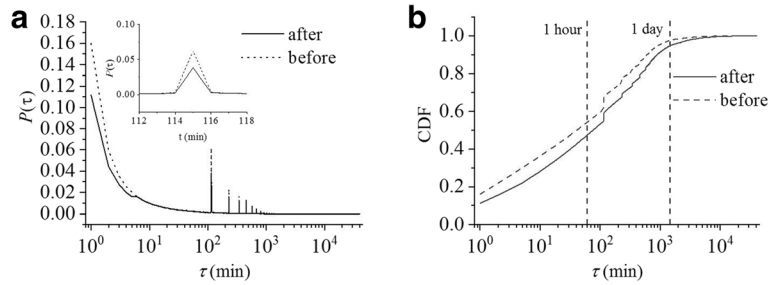


**Fig. 8** Stay time distribution comparison between before and after using the heuristic strategies. **a** Probability distribution. **b** Cumulative distribution

as shown in Fig. 8, where τ represents the stay time (stay duration τ). It can be seem that after user's trajectory is processed by our model, the number of stay duration is 72.6% of that before processing, which reduces the small stay duration data and increases the long-term stay duration, and is also the result of eliminating some oscillation points. It makes the user's stay time distribution more reasonable. At the same time, the jump growth occurring near 115 min due to the mobile system setting during the stay duration is also greatly reduced, as shown in the illustration of Fig. 8a.

#### *4.2.3 Time interval characteristics*

The study of time intervals is generally expressed in terms of a distribution, such as the distribution of time intervals between consecutive communication records. Such intervals aim at user's call behaviors. Many studies have pointed out that the communication behavior has power-law characteristics, which is not described in detail in this paper. However, it should be pointed out that the time interval here is closer to the trip time and is the time interval between two stay locations, as shown in Fig. 9. In this figure, the differences between the distribution stay time intervals before and after processing with trajectory processing algorithm are compared. The distributions are similar, but the number of time intervals after processing is 63.7% of that not being processed, which is obviously different with many power law distributions and Poisson distributions pointed out in many references.

## 5 Conclusion

Location discrimination of mobile phone CDRs data is the starting point and difficulty of research in this field. In order to overcome these problems, in this paper, the heuristic strategies in different situations are introduced into the location discrimination model, and the transfer time allocation heuristic strategy is added for the first time. The model parameter selection is elaborated by taking about 1 TB data as an example. The results show that this method can discriminate false location better, and it can be adopted to obtain more practical results at the individual level than the threshold and clustering methods, also can retain more track points to improve the accuracy of CDRs data in trajectory recognition. The model also performs well at the group level. In a word, the results provided remarkable improvement over existing techniques on real data. The increased accuracy can make mobile positioning methods more useful in mobility modeling. In addition, data structures and algorithms suitable for big data parallel architecture are proposed, which provide a solution for large-scale analysis of mobile phone data.

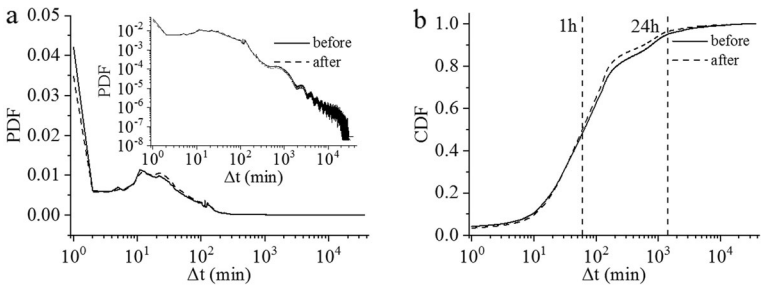


**Fig. 9** Time interval between stays distribution before and after using the heuristic strategies. **a** Probability distribution. **b** Cumulative distribution

### Abbreviations
CDR: Call detail record; OD: Origination destination; BSC: Base station controller; MSC: Mobile switching center; GPS: Global position system; PDF: Probability distribution function; CDF: Cumulative distribution function; CCDF: Complementary cumulative distribution function

### Acknowledgments
The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions. I would like to acknowledge all our team members. These authors contributed equally to this work.

### About the authors
Qingchao Shan is a doctoral student from the School of Traffic and Transportation, Beijing Jiaotong University. His research interests include traffic big data mining, intelligent traffic monitoring system and machine learning.
Limin Jia graduated from the China Academy of Railway Sciences, Beijing, China, in 1991. He is a Ph.D. supervisor, works in state key lab of rail traffic control & safety of Beijing Jiao Tong University. His research field is in intelligent control and intelligent theory, rail traffic intelligent control and safety system key technology development, and new energy system theory and technology research.
Honghui Dong received his doctor degree in pattern recognition and intelligent systems form Institute of automation, Chinese academy of sciences. His research interests include pattern recognition, machine learning and intelligent traffic system.
Hua Yuan graduated from Beijing Technology and Business University in 2001. She is a software engineer, works in Beijing MainSoft Technology Corporation Limited.
Hui Zhang received the doctor's degree from Beijing Jiaotong University, Beijing, China, in 2018. He works at Transportation and Economics Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing from 2017. His current research interests include automatic drive, railway transportation and transport of dangerous goods.

### Authors' contributions
All authors take part in the discussion of the work described in this paper. These authors contributed equally to this work and should be considered co-first authors. All authors read and approved the final manuscript.

### Funding
This work was financially supported through grants from the National Science and Technology Support Plan Project (2014BAG01B02). The authors thank the anonymous reviewers for their helpful suggestions.

### Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

### Competing interests
The authors declare no conflict of interest. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

### Author details
[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China. [2]Beijing MainSoft Technology Corporation Limited, Beijing 100041, China. [3]Transportation and Economics Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China.

Received: 27 May 2019 Accepted: 7 August 2019
Published online: 02 September 2019

### References
1. F. Yang, *Link travel speed data capture technology based on cellular handoff information: method, algorithm and evaluation* (Science Press, 2013)
2. Z. Wang, S.Y. He, Y. Leung, Applying mobile phone data to travel behaviour research: a literature review. Travel Behav. Soc. **11**, 141–155 (2018)
3. M.C. González, C.A. Hidalgo, A. Barabási, Understanding individual human mobility patterns. Nature **453**, 779–782 2008-06-05 2008
4. C. Song, Z. Qu, N. Blumm, A. Barabasi, Limits of predictability in human mobility. Science **327**, 1018–1021 (2010)
5. K. Keramat Jahromi, M. Zignani, S. Gaito, G.P. Rossi, Simulating human mobility patterns in urban areas. Simul. Model. Pract. Theory **62**, 137–156 2016-01-01 2016
6. S. Jiang, J. Ferreira, M.C. Gonzalez, Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Transactions on Big Data **3**, 208–219 (2017)
7. C. Song, T. Koren, P. Wang, A. Barabási, Modelling the scaling properties of human mobility. Nat. Phys. **6**, 818–823 (2010)
8. P. Widhalm, Y. Yang, M. Ulm, S. Athavale, M.C. González, Discovering urban activity patterns in cell phone data. Transportation **42**, 597–623 (2015)
9. D. Gundlegård, C. Rydergren, N. Breyer, B. Rajna, Travel demand estimation and network assignment based on cellular network data. Comput. Commun. **95**, 29–42 2016-12-01 2016
10. M.S. Iqbal, C.F. Choudhury, P. Wang, M.C. González, Development of origin–destination matrices using mobile phone call data. Transportation Res. Part C: Emerg. Technol. **40**, 63–74 (2014)
11. Z. Fan, T. Pei, T. Ma, Y. Du, C. Song, Z. Liu, C. Zhou, Estimation of urban crowd flux based on mobile phone location data: a case study of Beijing, China. Comput. Environ. Urban Syst. **69**, 114–123, 2018-01-01 (2018)
12. Y. Cai, W. Qihui, H. Tian, *Modern Mobile Communications* (China Machine Press, 2013)
13. W. Wu, S. Krishnaswamy, J. Decraene, A.S. Nash, Y. Wang, J.B. Gomes, T.A. Dang, S. Antonatos, M. Xue, P. Yang, in *2014 IEEE 15th International Conference on Mobile Data Management*. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling (IEEE Computer Society, Brisbane, 2014), pp. 321–328
14. M.A. Bayir, M. Demirbas, N. Eagle, Mobility profiler: a framework for discovering mobility profiles of cell phone users. Pervasive Mob. Comput. **6**, 435–454, 2010-08-01 (2010)
15. S.A. Shad, E. Chen, Spatial outlier detection from GSM mobility data. Computer Sci. **3**(3), 68–74 (2012)
16. P. Fiadino, D. Valerio, F. Ricciato, K.A. Hummel, in *Traffic Monitoring and Analysis. vol. 7189*. Steps towards the extraction of vehicular mobility patterns from 3G signaling data (Springer Berlin Heidelberg, 2012), pp. 66–80
17. T. Vajakas, J. Vajakas, R. Lillemets, Trajectory reconstruction from mobile positioning data using cell-to-cell travel time information. Int. J. Geogr. Inf. Sci. **29**, 1941–1954 (2015)

## Publisher's Note