

## Research on urban road travel speed extraction based on mobile phone signaling data

Jian Zhang<sup>1, a</sup>, Xiangwei Zeng<sup>1, b, \*</sup>, Jie Zhang<sup>2, c</sup> and Gang Zhong<sup>1, d</sup>

<sup>1</sup>Jiangsu Key Laboratory of Urban ITS, School of Transportation, Southeast Univ.; Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies; Jiangsu Province Collaborative Innovation Center for Technology and Application of Internet of Things; Research Center for Internet of Mobility, Si Pai Lou #2, Nanjing, Jiangsu 210096, China

<sup>2</sup>Jiangsu Micro-electromechanical Technology and Application Center, China Nanjing  
<sup>a</sup>jianzhang@seu.edu.cn, <sup>b</sup>brendaty@126.com, <sup>c</sup>a135b920@gmail.com, <sup>d</sup>anhuizhonggang@126.com

\* Corresponding author

**Keywords:** Mobile phone signaling switching data; Road matching; Travel speed; Communication network

**Abstract:** Matching mobile phone signaling switching data with road and extraction of road travel speed is the foundation of obtaining traffic information by using mobile phone signaling data, however, the existing matching method extracting a feature sequence has a large error, and results in a lot of waste of road test mobile phone signaling switching data. Using the location feature of mobile signaling switching place, put forward an urban road calibration and matching method based on the longest common subsequence dynamic programming. By pre-processing of mobile phone signaling data measured on the road, calculate the correlation threshold of urban road calibration sequence based on the longest common subsequence; by dynamic programming method, calculate the degree of association of the sequence that has not been matched and the road test sequence set and compare it with correlation threshold, and match cellular signaling data with urban road, then calculate the travel speed of the successful matched sequence. The result of example analysis shows that the result of the road matching and travel speed calculation is less error.

### 1. Introduction

With the acceleration of urbanization construction and the development of the economic level in China, traditional questionnaires, manual surveys and other research means it cannot meet the demands of traffic study. In recent years, determining the travel route by mobile location technology to estimate the road traffic status has been valued gradually in the field of traffic study. The cellphone data has many advantages, such as low cost, huge data volume and ecological economics, etc, and it can be used to track the complete travel course of users, which is very beneficial to improve the accuracy of the estimation of road traffic status. At the time of gaining the data of the road traffic status by the cellphone data, matching correctly the data generated by the on-board mobile phone with the urban road and extracting the travel speed is a key problem to be solved.

## 2. Literature Review

In some case, the originating COO location technology is applied more often. For example, we can identify the travel law of the user through the analysis of the mobile user's stop spot [1], or extract the location of the base station where mobile users stay for extraction and analysis of the trip OD data by the COO information provided by the GSM cellular system[2,3]. However, COO location can only be accurate to within the service radius of the base station, while the service radius of the base station vary from tens of meters to several hundred meters. The lower the coverage density of the base station is, the greater the service radius will be, which means COO positioning results are more inaccurate. Therefore, the discriminant accuracy of COO positioning technology for the specific path is quite low, and it hardly determines the true trip path of mobile users in staggered urban roads with complex base station distribution.

In other case, the signaling switching location technology used commonly, that is, position the switch place of the mobile user through the change of location area code (LAC) in the cellphone signaling data and cell identity (CellID) of the base station when mobile users pass through service ranges of two different base stations. Tang Yuhao[4] describes extraction, cleaning, loading, secondary-processing method of the cellphone signaling switch data, which gets trip OD matrix, the expected routines of residents when traveling and other core data. Mao Xiaowen [5] describes the preprocessing process of the cellphone signaling data and does some research into the regional travel characteristics of residents using the cellphone signaling data. Yang Fei[6,7] calibrates roads by a continuous switching sequence with the heaviest weighting as a feature sequence of a road gained by superimposing switch sequences of all the tests, and regards it as the road matching basis of the sample to be tested. But the switch sequence with the maximum weight cannot represent the real signaling switches on the road, nor can it reflect the real traffic state on the road. A large number of signaling switches are excluded from the road calibrations, resulting in a waste of test data. It can be seen that the signaling switch positioning technology is more accurate than COO positioning technology, but the data processing and the path matching process are also more complex, there is no recognized method.

## 3. Methodology

Whether the user or the time of switch is coincided or not, the location where the same cellphone signaling switch is always located in the same interweaved region of two base station service scopes. By testing the switching sequence that occurs on a certain road and calculating the correlation degree of the signaling switches of users to be tested, we can determine whether or not the signaling data of the users matches the road.

To match the signaling switch data generated by the mobile phone users with the urban road, a lot of road tests are needed to extract "the signaling switch label" of the road. For the signaling data to be calibrated, it is necessary to be integrated into an orderly switch sequence of cellphone signaling, and then, compare data with "the signaling switch label" of the road to judge whether the match succeeds through matching algorithm. For one signaling switch data matched successfully, we can calculate the average travel speed of the user according to its starting and ending time and the matched road length. The algorithm flow is shown in Figure 1.

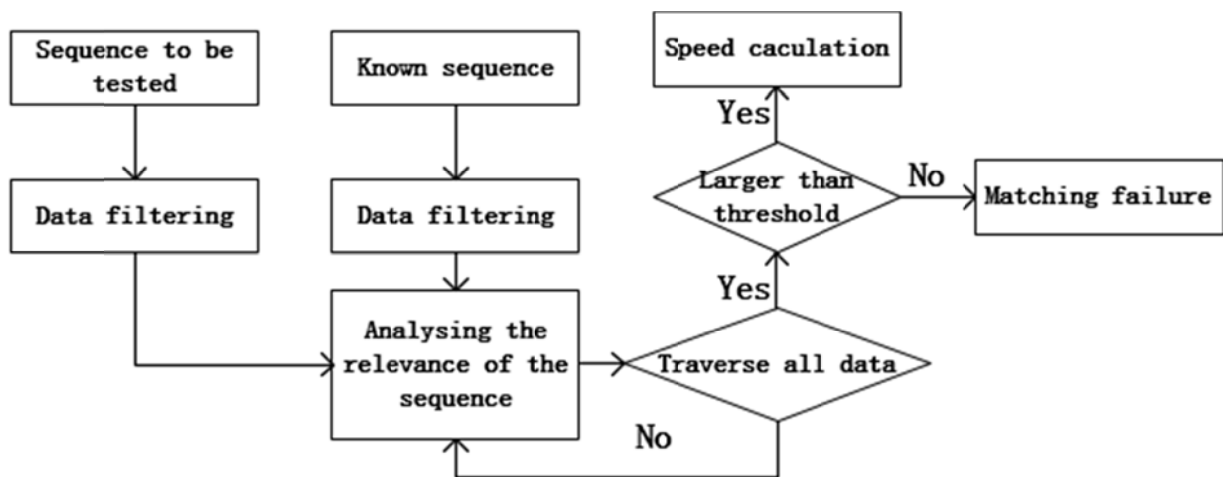


Fig. 1 The flow-process diagram of algorithm

### 3.1 The switch and positioning technology of mobile phone signaling

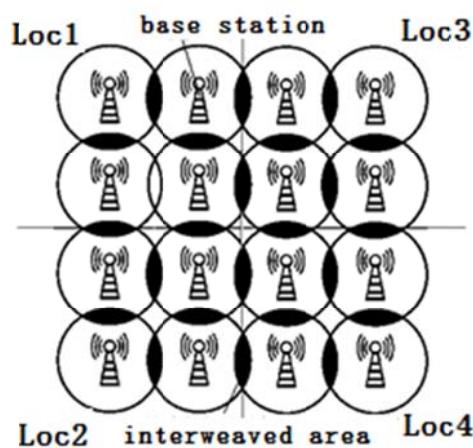


Fig.2 The work principle of GSM

GSM (GSM, Global System for Mobile Communications) is divided into MSC area, location area, base station area and cell [8], among which, the signal coverage area of the cell site is similar to a regular hexagon, several base station areas are arranged in a similar "honeycomb" structure.

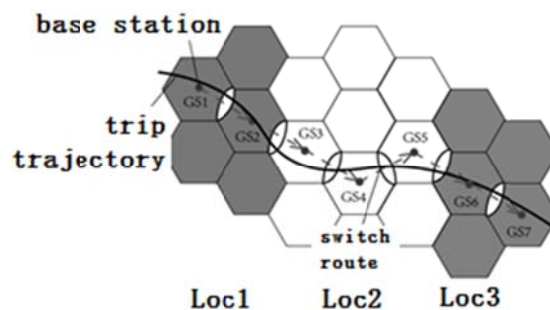


Fig.3 The mobile phone signaling data

The cellphone signaling switch refers to a process in which the mobile phone signaling switches from the base station with weak signals to a base station with strong signals due to a low communication signal strength between the current cellphone and the mobile phone base station.

The original cellphone signaling switching data mainly includes MSID (the only user ID with anonymous encryption), TIME (timestamp), LAC (location area code), CELLID (cell identity), EVENT (time types: text messages, MOC-MTC, surf the Internet, location update, cell switch) and other fields. When the signaling switch occurs, due to the mobile phone passes through different location areas, cell site, LAC code and CELLID change, we can locate mobile phone users approximately at that time by positioning the switch area as known as the fusiform region near the edge of "honeycomb" units.

## 3.2 The algorithm of road matching and speed extraction

### 3.2.1 Basic idea of the algorithm

Assuming that there is a unknown test sequence  $\mathbf{S} = \{a_1, a_2, a_3 \dots a_n\}$  and a signaling switch label sequence  $\mathbf{T} = \{b_1, b_2, b_3 \dots b_m\}$  obtained on a known road, either of which is a particular switch. We can figure out the matching result by extracting the common part between the test sequence S and the signaling switch label sequence T of this road, and calculating the ratio between the common part and the sequence length.

Assuming C is not only a subsequence of S, but a subsequence of T, then C is a common subsequence of S and T. If C is the longest among common subsequence of S and T, C is the longest common subsequence (LCS) of S and T [9]. Obviously, the simplest way to compare the relevance of two sequences is to compare the length of the longest common subsequence.

Once the longest common subsequence of the two switching sequences is obtained, the relevance of the two switching sequences can be estimated easily. We can calculate the LCS of the switching sequences (sequence S and sequence T) and compare with the relevance thresholds. While sequences are matched, the travel speed can be easily got based on the length of the urban road and the start and end time of the switch.

### 3.2.2 Data preprocessing

When performing road calibration and match, the valid fields in the signaling data include LAC (Location Area Code), CELLID (cell indenty), and TIME (Time Stamp) detection occurrence time. On the one hand, the mobile phone not only reports the position information to the network but also records the signaling data when crossing the cell site, producing redundancy record at the same time. On the other hand, the ping pong effect is easy to happen due to the switch disturbance caused by the unstable signal, which means the mobile phone signal switches back and forth among several base stations. These phenomena not only consume storing resources, but dramatically reduce the matching accuracy. Therefore, we preprocess the signaling data in the following ways to extract a simple and effective switching sequence:

1) Comparing LAC of the two piece of adjacent data and the cell identity CELLID, when some of them change, the program will affirm that a switch happens  $\{(Lac_1 \text{ CellID}_1), (Lac_2 \text{ CellID}_2)\}$ .

2) After Step 1 finished, retrieve the previous switch. If the sequence of switch has the shape of  $\{(Lac_1 \text{ CellID}_1), (Lac_2 \text{ CellID}_2), (Lac_1 \text{ CellID}_1)\}$ , the time interval of switch will be calculated. The program will only keep  $\{(Lac_1 \text{ CellID}_1), (Lac_2 \text{ CellID}_2)\}$  when the time interval is too short.

### 3.2.3 The relevance calculation of the switching sequence

With the dynamic programming method, we calculate the longest common subsequence of the test sequence  $S = \{a_1, a_2, a_3 \dots a_n\}$  and the calibration sequence  $T = \{b_1, b_2, b_3 \dots b_m\}$  gained by the road test. Specific steps of relevance calculation are shown as follows:

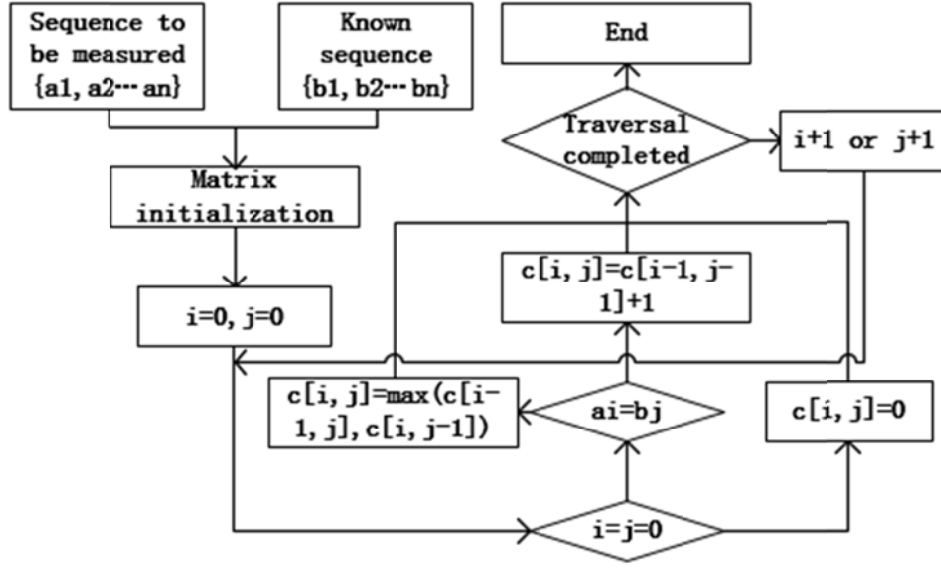


Fig.4 The solution diagram of the longest common subsequence

1) Calculate the length of the longest common subsequence of the test sequence and the calibration sequence  $LCS$ , with the method shown in Figure 4.

2) The length of the real longest common subsequence gained by calculation varies greatly since the lengths of both test sequence and the road test calibration sequence are not fixed. The relevancy  $\rho$  is introduced to calculate the relevance.

$$\rho = \frac{LCS}{\min(m,n)} \quad (1)$$

### 3.2.4 Calculation of road matching and travel speed

The relevance between the test sequence  $S$  and the calibration sequence  $T$  can be found, according to the relevance calculation method in the previous section. Obviously, a calibration sequence cannot fully reflect the switch sequence characteristics of this road due to the randomness of the mobile phone signaling switch position. It is necessary to invoke a large number of calibration sequences on this road for the relevance calculation and calculate average values of its arithmetic to improve the matching accuracy and make use of all the reasonable road test data effectively. For the test sequence which successfully matches with a road, the average travel speed of the users on the road is obtained.

1) Make a statistic of the percentage of different sequences  $P(T_i)$ , based on massive calibrated sets  $\{T_1, T_2 \dots T_k\}$  gained on a particular road.

2) For all calibrated sequences  $T_i$ , we can calculate the relevance  $\rho_i$  of the test Sequence  $S$  using the method described above and obtain it by invoking the result of step 1.

3) We can finally calculate the weighted relevance of the test sequence  $S$  based on the results above.

$$\rho = \sum_{i=1}^n \rho_i * P(S_i) \quad (2)$$

4) The result of weighted relevance can be tested by the preset threshold  $Th$ , at the time when  $\rho > Th$ , the test Sequence  $S$  can be matched to the road where the sequence set  $\{T_1, T_2 \dots T_k\}$  is located. At that time when  $\rho < Th$ , the test Sequence  $S$  cannot be matched to the road.

5) For the tested the test Sequence  $S$  which has been successfully matched, the start and end

switching time  $t_1, t_2$  of the sequence can be obtained. With the given total length  $l$  of the road, the average travel speed of the user is:

$$v = \frac{l}{t_2 - t_1} \quad (3)$$

#### 4. Case Study

Table 1 The test sample of mobile phone signaling data in Taiping North Road

LAC1	CELLID1	LAC2	CELLID2
53504	100687959	53504	100668757
53504	100668757	53504	100667467
53504	100667467	53504	100667468
53504	100667468	53504	100666408
53504	100666408	53504	100666407
53504	100666407	53504	100667468
53504	100666408	53504	100668759
53504	100668759	53504	100673237
53504	100673237	53504	100666968
53504	100666968	53504	100667127
53504	100667127	53504	100667128
53504	100667128	53504	100667129

Real field testing is required. The authors designed a real life experiment, which test sample is shown in Table 1, and the diagram of measured data is shown in Fig. 5. The threshold of the weighted relevance judgment results in this experiment is 2.5. Given an extra user test Sequence A, we can obtain the maximum relevance between each sequence sample and Sequence A, then calculate the weighted relevance between Sequence A and all the sequence samples which is 2.7414 through the algorithm discussed above. The weighted relevance is greater than the threshold, so it can be sure that the user's signaling data matches with the road successfully.

Table 2 The result of path matching and travel speed gathering

Actual travel time [h]	Estimated travel time[h]	Actual speed [km/h]	Estimated speed [km/h]	Speed error[%]	Whether the match is successful
0.0439	0.0497	34.1772	30.1676	11.7318%	Yes
0.0733	0.0667	20.4545	22.5000	10.0000%	Yes
0.0850	0.0917	17.6471	16.3577	7.3064%	Yes
0.0922	0.0850	16.2686	17.6471	8.4733%	Yes
0.0854	0.0783	17.5671	19.1489	9.0045%	Yes
0.0755	0.0817	19.8582	18.3673	7.5075%	Yes
0.0571	0.0650	26.2910	23.0769	12.2250%	Yes
0.0781	0.0717	19.1971	20.9302	9.0282%	Yes
0.1072	0.0983	13.9923	15.2542	9.0192%	Yes
0.1154	0.1017	13.0033	14.7541	13.4641%	Yes
0.0989	0.0887	15.1711	16.9131	11.4822%	Yes
0.1067	0.0972	14.0572	15.4316	9.7771%	Yes

Getting the time when the mobile phone signaling is switched to the base station near the road and the time when it leaves the base station near the road, we can calculate the time  $t=t_2-t_1$ , when the vehicle is running on the road. The length  $l$  of the road section is gained from the database. In

this experiment, the length of Taiping North Road  $l = 1.5\text{km}$ , while  $t = 0.0917\text{ h}$ . Then the travel speed is

$$u = \frac{l}{t} = \frac{1.5}{0.0917} = 16.4\text{km/h} \quad (4)$$

Which is equal to the measured average speed  $16.7\text{km/h}$  basically.

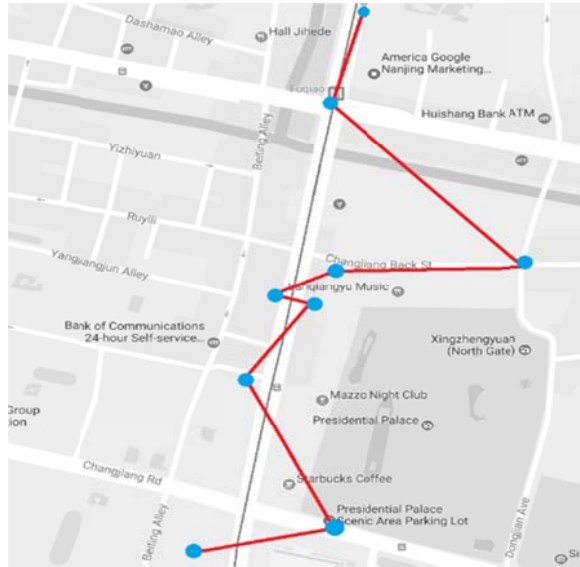


Fig.5 The diagram of measured mobile phone signaling data

## 5. Conclusions

The mobile phone signaling switch data has a broad application prospect in the field of traffic engineering due to its low cost of data sources, a wide coverage, etc. The road match of mobile phone signaling data and the travel speed of mobile phone user are the foundations of the estimation of road traffic state and the research on residents' travel behavior. In light of insufficient research methods for the road match, this paper proposes a new extraction model of the urban road travel speed based on the mobile phone signaling switching data, which makes use of a large number of road test switches, and improves the accuracy and matching efficiency of matching the mobile phone signaling data with the urban road. This model presents a completely new road matching method, which provides data support for the estimation of the road traffic state and the study of the resident travel path. But the model also has its limitations, for example, a more complex road environment with sideways, crossroads and so on is not studied. In practical applications, the privacy concerns of mobile phone users, the distinction between non-motor vehicles and motor vehicles, how to identify the number of mobile phones carried on the same vehicle and other issues also need to be considered. Only an in-depth study can make the application of mobile data truly contribute a lot to the intellectualization, high-speed and simplicity for the acquisition of urban traffic conditions.

## Acknowledgements

This study is partially supported by the Information Technology Research Project of Ministry of Transport of China (No. 2015364X16030), and the Fundamental Research Funds for the Central Universities (2242015K42132).

## References

- [1] Zhang Yongchuan. The Analysis of User Travel Law Based on the Mobile Phone Location Data [D]: [Master's thesis]. Kunming University of Science and Technology (2013)
- [2] Li Jia. The Study of Dynamic OD Matrix Acquisition Method Based on the Mobile Location Data. [D]: [Master's thesis]. Kunming University of Science and Technology (2013)
- [3] Wang Ruipeng. OD Data Acquisition and Flow Pattern Extraction of Mobile Phone User [J]. Fujian Computer, Vol. 1, (2016), p. 92-94.
- [4] Tang Yuhao. A New Method to Obtain OD Characteristics of Urban Residents Based on Large Mobile Phone Signaling Data [J]. Information Communication. Is. 2, (2016), p. 267-269.
- [5] Mao Xiaowen. Research on the Trip Characteristics of Regional Transportation Based on Mobile Signaling Technology [D]: [Master's Thesis]. Chongqing Jiaotong University, 2014.
- [6] Yang Fei. Road Matching Method Based on Mobile Switching Mode [J]. Systematic Engineering, Is. 11, (2013), p. 9-12.
- [7] Yang Fei, Hui Ying, Yang Dongyuan. Traffic Network Calibration Method Based on Mobile Handoff Location [J]. Tongji University, Is. 1, (2009), p. 67-72
- [8] Li Jinming. The Overview of GSM Mobile Communication System [J]. Lanzhou Petrochemical College of Vocational Technology Journal, Vol. 1 (2001), p. 21-24.
- [9] Zheng Cui-ling. Analysis Algorithm and Implementation of Longest Common Subsequence [J]. Journal of Wuyi University, Is. 2, (2010), p. 44-48.