The Journal of Transport and Land Use http://jtlu.org
Vol. 13 No. 1 [2020] pp. 429–445

Trip mode inference from mobile phone signaling data using Logarithm Gaussian Mixture Model

Xiaoxu Chen
Key Laboratory of Road and Traffic
Engineering of the Ministry of Education
Tongji University
hitcxx@163.com

Chao Yang (corresponding author)
Key Laboratory of Road and Traffic
Engineering of the Ministry of Education
Tongji University
tongjiyc@tongji.edu.cn

Xiangdong Xu
Key Laboratory of Road and Traffic
Engineering of the Ministry of Education
Tongji University
xiangdongxu@tongji.edu.cn

Abstract: Trip mode inference plays an important role in transportation planning and management. Most studies in the field have focused on the methods based on GPS data collected from mobile devices. While these methods can achieve relatively high accuracy, they also have drawbacks in data quantity, coverage, and computational complexity. This paper develops a trip mode inference method based on mobile phone signaling data. The method mainly consists of three parts: activity-nodes recognition, travel-time computation, and clustering using the Logarithm Gaussian Mixed Model. Moreover, we compare two other methods (i.e., Gaussian Mixed Model and K-Means) with the Logarithm Gaussian Mixed Model. We conduct experiments using real mobile phone signaling data in Shanghai and the results show that the proposed method can obtain acceptable accuracy overall. This study provides an important opportunity to infer trip mode from the aspect of probability using mobile phone signaling data.

Keywords: Trip mode inference, mobile phone signaling data, Logarithm Gaussian Mixed Model

Article history:
Received: March 28, 2019
Received in revised form: March 2, 2020
Accepted: March 4, 2020
Available online: November 10, 2020

1 Introduction

Trip mode inference determines the transportation mode of travelers, based on the speed, travel time and other information obtained from their trips, which is significant for transportation planning and management. Accurate trip mode inference is also important for the study of trip mode choice (Reichert & Holtz-Rau 2015; Zahabi, Miranda-Moreno, Patterson, & Barla, 2012).

http://dx.doi.org/10.5198/jtlu.2020.1554

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

Traditional methods of acquiring the trip mode are usually based on questionnaires, travel diaries and telephone interviews (Stenneth, Wolfson, Yu, & Xu, 2011). The wide spread of mobile phones has made it an effective means of collecting trip information. Existing methods of trip mode identification mainly focus on GPS data, and some involve mobile phone sensor data and call detail records (CDRs).

Mobile phone data is mainly obtained through mobile phone positioning technology and various sensors. The data collected by different technologies has different characteristics and precisions, which can be divided into coarse-grained data and fine-grained data. The coarse-grained data is mainly obtained through the cell phone tower network positioning technology, including CDRs and mobile phone signaling data. The fine-grained data is acquired through mobile terminal positioning technology and sensors, consisting of GPS data, triaxial acceleration data, angular acceleration, and gravitational acceleration, etc. In terms of fine-grained data, some studies only applied mobile phone GPS data to identify trip mode (Young-Ji, Abdulhai, & Shalaby, 2009; Bolbol, Cheng, Tsapakis, & Haworth, 2012; Gonzalez et al, 2010; Zhang, Liu, Bao, & Qiang. (2015), and some studies only utilized the mobile phone acceleration data in the trip mode identification (Nham, Siangliulue, & Yeung, 2008; Sun, Zhang, Li, Guo, &, Li, 2010). Moreover, some scholars used the mobile phone sensor data (Nick, Coersmeier, Geldmacher, & Goetze, 2010; Frendberg, 2011), and some scholars combined the mobile phone GPS data and acceleration data (Reddy, Burke, Estrin, & Hansen, 2008; Reddy et al., 2010). In terms of coarse-grained data, some scholars used GSM data to identify trip mode (Anderson & Muller, 2006; Sohn et al., 2006). Wang, Calabrese, Lorenzo, and Ratti (2010) only utilized CDRs to infer trip mode.

In the existing literature, the identification of trip mode is generally implemented by a rule-based method or a machine learning algorithm. Rule-based methods usually set thresholds for feature variables, including speed, time, acceleration and distance, etc. In specific processing, multiple thresholds are generally used to identify trip modes (Shin et al., 2015; Li, Yang, Zhang, Zhou, & He, 2015). There are also a few scholars who utilized a single travel time (Wang et al., 2010) to develop identification rules. Some scholars have combined multiple algorithms to identify trip mode. More commonly used ones in machine learning algorithms are support vector machines (Nham et al., 2008; Ashqar, Almannaa, Elhenawy, Rakha, & House, 2018; Xiao, Wang, Fu, & Wu, 2017), decision trees (Reddy et al., 2008; Nick et al., 2010; Reddy et al., 2010), random forest (Shafique & Hato, 2015; Xiao et al., 2017), Bayesian network (Reddy et al., 2008; Nick et al.,2010), neural network (Young-Ji et al., 2009; Gonzalez et al., 2010), K-nearest neighbor algorithm (Reddy et al., 2008; Reddy et al., 2010), hidden Markov model (Reddy et al., 2008; Xu et al., 2011) and so on. In summary, in the machine learning algorithm, the decision tree, random forest and support vector machine are more excellent. Since various algorithms have their own applicable limitations, the identification accuracy is also related to data type and feature variables. In addition, combinations of various algorithms and improved machine learning methods may contribute to the improvement of identification accuracy.

Most studies utilized GPS data to identify trip mode. However, GPS information is not available in shielded areas (e.g., tunnels) and the GPS sensor consumes significant power so that sometimes users turn it off to save the battery (Young-Ji et al., 2009). On the other hand, large volume of data about the position of mobile phones can be collected from signaling data. When people are in motion, the geographic information will be collected by mobile base station that are nearby (Xu et al., 2011), which construct the signaling data. To acquire signaling data and obtain significant information bring no extra overhead for mobile phone users and telecom operators. Wang et al. (2010) extracted trip information (user id, origin, destination, start time, end time) form CDRs. Based on the trip data collected from CDRs, travel time can be obtained. Thus, they infer trip mode based on travel time. The problem can be stated as follows: given an origin and a destination, as well as the travel time of travelers who move from the origin to the destination, identify trip mode for each traveler. Based on the Wang et al. (2010) method, we propose the trip mode identification method for mobile phone signaling data. Travel times are not evenly distributed, and the travelers can be clustered into subgroups according to their travel time.

In this paper, we first describe the dataset and complete the data preprocessing, then conduct trip mode identification method, which includes activity nodes recognition, travel time computation, and clustering with Logarithm Gaussian Mixed Model (Log-GMM). Activity nodes recognition focuses on finding the origin and the destination for trips. Considering the recorded time of signaling data is the handover time, travel time can be obtained through travel time computation method that we designed. Gaussian Mixed Model (GMM) clustering is a popular method, which has the advantage of applicability to large sample with unknown overall distribution. Based on GMM, we replace Gaussian distribution with Lognormal distribution and the Log-GMM is developed. Finally, the experiment is designed and conducted, and the results show that the method is promising.

The remainder of the paper is organized in the following four sections. Section 2 describes the dataset consisting of mobile phone signaling data and cell phone tower data. In Section 3, the trip mode inference method is developed. Subsequently, the results of the study are presented and finally the conclusions are presented.

2 Data set

Data set used in this study consists of mobile phone signaling data and cell phone tower data in Shanghai for the period from May 4th to May 17th of 2015. The mobile phone signaling data set consists of about 390 million records of 3 million users, and the main information used includes User ID, Timestamp and NID. Moreover, NID represents ID of cell phone tower. Examples of mobile phone signaling data are shown as Table 1. There are 33,118 cell phone towers in Shanghai. Cell phone tower data includes NID, longitude, and latitude and some examples are listed in Table 2. After matching the mobile phone signaling data to cell phone tower data by NID, mobile phone user's location at the time can be determined.

Table 1. Examples of mobile phone signaling data

User ID	Timestamp	NID
00033C76D635266D926BCE5D91B51700	20150504160620	16452
00033C76D635266D926BCE5D91B51700	20150504162005	5802
00033C76D635266D926BCE5D91B51700	20150504162215	4537
00033C76D635266D926BCE5D91B51700	20150504162413	5152

Table 2. Examples of mobile phone tower data

NID	longitude	latitude
11678	120.6480	31.33301
34038	120.6778	31.11788
5413	120.7962	30.68210
17235	120.8347	31.13114

3 Trip mode inference method

The trip mode identification method proposed in this paper includes activity nodes recognition, travel time computation, and clustering.

3.1 Activity nodes recognition

Due to the fluctuation of signal, cell phone positioning results may jump at several nearby cell phone towers. This situation is called ping-pong phenomenon or ping-pong handover (Vandenbroucke, Bucher, & Crompvoets, 2013). Most existing solutions to address this issue focus on increasing hysteresis threshold (used as a spatial constraint to merge close cell phone towers) to reduce the positioning error (Vandenbroucke et al., 2013). However, simply increasing the hysteresis threshold may result in dropping useful information from the spatial-temporal trajectories. To provide a better solution, we utilize two parameters to filter the phone signaling data. The first parameter is a spatial staying threshold δ to constrain fluctuation in spatial dimension and the second parameter is the temporal staying threshold τ used to limit the temporal dimension. When a user's position is fluctuating within a circle with radius less than δ during time period τ, this user is regarded as staying in this circle, and when a user's signal jump to a remote location and back to the previous location during time period τ, this user is regarded as motionless. The rules to recognize activity nodes are illustrated in Figure 1. The value of these two parameters are determined based on domain knowledge. The spatial staying threshold is selected as δ=400 meters considering lower bound of walking trip distance is 500 meters. τ is assigned to 30 minutes based on the analysis of household travel survey data of Shanghai in 2009, which is the 5% quantile of activity duration distribution and shown in Figure 2. Based on this observation, staying at a place more than 30 minutes is recognized as an activity node.

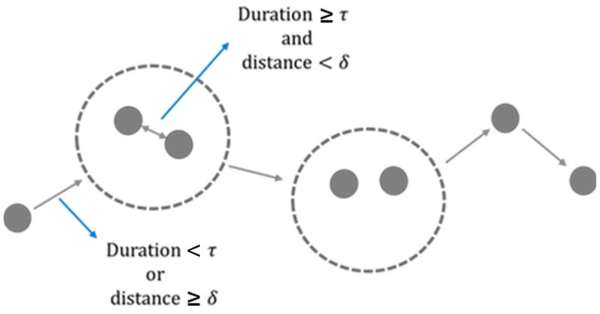


Figure 1. The rules to recognize activity nodes

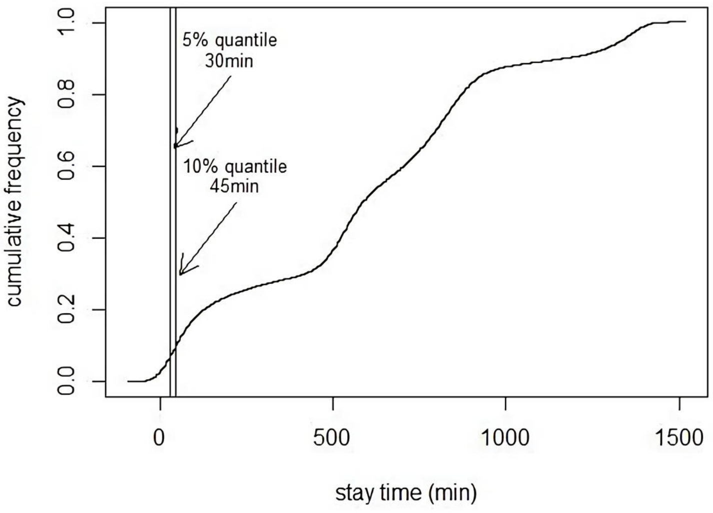


Figure 2. Cumulative frequency of activity duration

3.2 Travel time computation

Mobile phone signaling data has the trajectory of users. Once we annotate the activity nodes in trajectory, we can calculate trip travel time as following.

(1) For a user i, the series of records denoted as:

$$R_i = \{(t_{i,1}, n_{i,1}), (t_{i,2}, n_{i,2}), \dots, (t_{i,j}, n_{i,j})\} \quad j \in Z^+ \tag{1}$$

where t_{ij} and n_{ij} respectively correspond to the timestamp and cell phone tower location number of the jth record of user i.

(2) Compute travel time according to activity nodes obtained from Section 3.1. We assume the first activity node of user i is the kth record, denoted as the origin of trip m, and the second activity node is the sth record, denoted as the destination of trip m, which is also the origin of trip m+1 of user i. The OD pair of the trip m is (n_{ik}, $n_{i,s}$). $t_{i,j}$ represents the moment of arriving at cell phone tower $_{n\_}$(i,j) of user i, which also means the moment of departing from last phone tower $n_{i,j-1}$, thus the travel time T_i^m of trip m of user i can be computed by:

$$T_i^m = t_{i,s} - t_{i,k+1} \tag{2}$$

(3) Compute travel time of trips for all users.

3.3 Clustering of trip travel time

We develop Logarithm Gaussian Mixture Model based on GMM. GMM could be used to cluster travel time. GMM is a probabilistic model for representing normally distributed subpopulations within an overall population (Rasmussen, 1999). GMM refers to the estimation of the probability density distribution of a sample, and the model is a weighted sum of several Gaussian models and each Gaussian model represents a class. The data in the sample are projected on several Gaussian models, we can get the probability of each category, and then we can choose the most probable class as the result of the decision.

GMM is parameterized by two types of values, the mixture component weights and the component mean and variances. For a GMM with K components, the kth component has a mean of μ_k and variance of σ_k for the univariate case. The mixture component weights are defined as φ_k for component C_k, with the constraint that $\sum_{i=1}^{k} \varphi_i = 1$ so that the total probability distribution normalizes to 1. The mathematical form of GMM is as follows:

$$p(x) = \sum_{i=1}^{K} \varphi_i \, N(x|\mu_i, \sigma_i) \tag{3}$$

$$N(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{4}$$

where $\varphi_i \geq 0$, $N(x \mid \mu_i, \sigma_i)$ is normal distribution.

If the number of components K is known, expectation maximization (EM) is the technique most commonly used to estimate the mixture model's parameters. EM is a numerical technique for maximum likelihood estimation and is usually used when closed form expressions for updating the model parameters can be calculated. In frequentist probability theory, models are typically learned by using maximum likelihood estimation techniques, which seek to maximize the probability, or likelihood, of the observed data given the model parameters. Unfortunately, finding the maximum likelihood solution for mixture models by differentiating the log likelihood and solving for 0 is usually analytically impossible. EM algorithm (Dempster, Laird, & Rubin, 1977) is an iterative algorithm and has the convenient property that the maximum likelihood of the data strictly increases with each subsequent iteration, meaning it is guaranteed to approach a local maximum or saddle point.

Expectation maximization for mixture models consists of two steps. The first step, known as the expectation step or E step, consists of calculating the expectation of the component assignments C_k for each data point $x_i \in X$ given the model parameters φ_k, μ_k, *and* σ_k. The second step is known as the maximization step or M step, which consists of maximizing the expectations calculated in the E step with respect to the model parameters. This step consists of updating the values φ_k, μ_k, *and* σ_k. The entire iterative process repeats until the algorithm converges, giving a maximum likelihood estimate. Intuitively, the algorithm works because knowing the component assignment C_k for each x_i makes solving for φ_k, μ_k, *and* σ_k. easy, while knowing φ_k, μ_k, *and* σ_k. makes inferring $p\,(C_k | x_i)$ easy. The expectation step corresponds to the latter case while the maximization step corresponds to the former. Thus, by alternating between which values are assumed fixed, or known, maximum likelihood estimates of the non-fixed values can be calculated in an efficient manner.

The EM algorithm for GMM starts with an initialization step, which assigns model parameters to reasonable values based on the data. Then, the model iterates over the expectation (E) and maximization (M) steps until the parameters' estimates converge, i.e., for all parameters θ_t at iteration t, $|\theta_t - \theta_{t\text{-}1}| \leq \epsilon$ for some user-defined tolerance ϵ. The EM algorithm for GMM with K components can be described as follows:

Initialization Step: Randomly selected samples without replacement from the dataset
$X = \{x\_{1},\ldots, x_N\}$ as the component mean estimates $\hat{\mu}_1,\ldots, \hat{\mu}_K$. Set all component variance estimates to the sample variance $\hat{\sigma}_1^2,\ldots, \hat{\sigma}_K^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$, where $\bar{x}$ is the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Set all component distribution prior estimates to the uniform distribution $\hat{\varphi}_1, \ldots, \hat{\varphi}_K = \frac{1}{k}$.

E-step: Calculate ∀ *i,k*

$$\hat{\gamma}_{ik} = \frac{\hat{\varphi}_k N(x_i|\hat{\mu}_k, \hat{\sigma}_k)}{\sum_{j=1}^{K} \hat{\varphi}_j \, N(x_i|\hat{\mu}_j, \hat{\sigma}_j)} \tag{5}$$

where $\hat{\gamma}_{ik}$ is the probability that x_i is generated by component C_k.

M-step: Using the $\hat{\gamma}_{ik}$ calculated in the E-step, calculate the following in that order ∀ k :

$$\hat{\varphi}_k = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N} \tag{6}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} x_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{7}$$

$$\hat{\sigma}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} (x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{8}$$

Most travel times represent skewed distribution (Fosgerau & Fukuda, 2012; Guessous, Aron, Bhouri, & Cohen, 2014; Rahman, Wirasinghe, & Kattan, 2018). Sometimes, travel time can represent Gaussian distribution. In this paper, we assume travel times obey Lognormal distribution. Based on GMM, we develop Logarithm Gaussian Mixture Model as follows:

$$p(x) = \sum_{i=1}^{K} \varphi_i \cdot logN(x|\mu_i, \sigma_i) \tag{9}$$

$$logN(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(lnx - \mu_i)^2}{2\sigma_i^2}\right) \tag{10}$$

The EM algorithm also can be used for Log-GMM. The dataset $X = \{x_1 ,..., x_N\}$ can be transformed to $logX = \{ lnx_1 ,..., lnx_N\}$. Then the dataset $\{ lnx_1 ,..., lnx_N\}$ is used to estimate parameters of Log-GMM through the EM algorithm for GMM.

By estimating the parameters of Log-GMM, clustering can be achieved through probability computation. Figure 4 illustrates how trip mode identification is accomplished using Log-GMM. There are three components in Figure 4, and each component has its own parameters. For a sample, probabilities in components can be calculated and the component with the largest probability is regard as the trip mode of the sample. The thresholds can be obtained after parameters estimation. When travel time t of a trip is less than t_1, the mode of the trip is identified as Mode 1. When travel time t of a trip is larger than t_2, the mode of the trip is identified as Mode 3. While travel time t of a trip is between t_1 and t_2, the mode of the trip is identified as Mode 2. In this paper, we mainly study the long-distance trips and the modes to identify include car, subway and bus. In Figure 3, Mode 1, Mode 2 and Mode 3 respectively correspond to car, subway and bus.

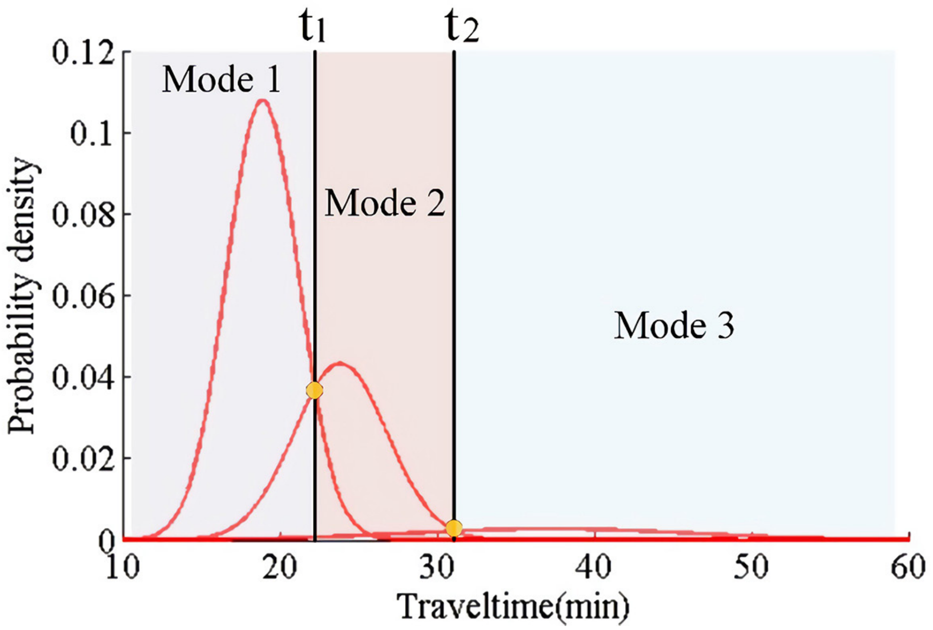


Figure 3. Illustration of trip mode identification

4 Results

Travel time in given two activity nodes is mostly fluctuating at a certain value for each trip mode, thus we select two determined activity nodes after all travel time have been computed. In the experiment, we select Pinglu Rd. and Beijing West Rd., and the detailed information is shown in Table 3. In order to validate the travel time computed from mobile phone signaling data, we obtain the travel time reported from Auto Navi Map (https://www.amap.com/), which are shown in Table 4.

Table 3. Information of two activity nodes

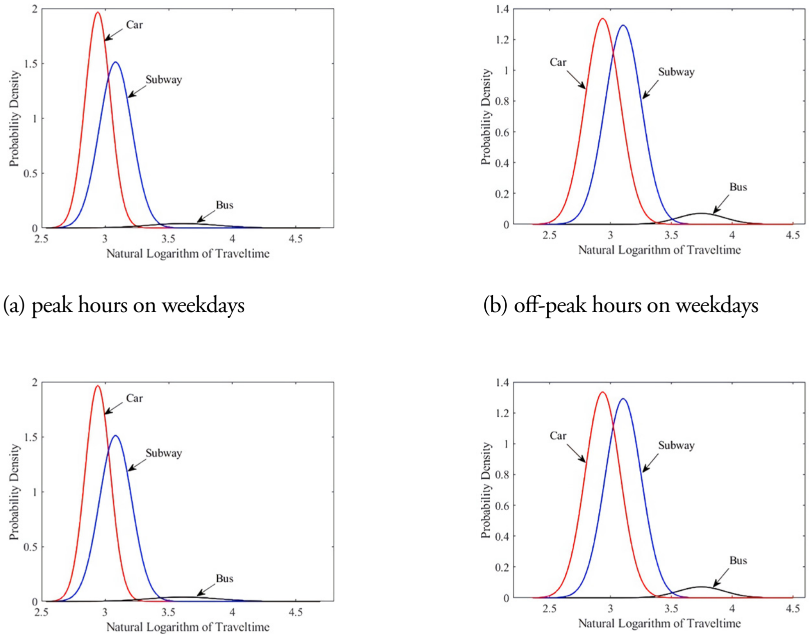
	NID	Longitude	Latitude	Location
Activity Node 1	1098	121.214973	31.304360	Pinglu Rd.
Activity Node 2	5716	121.252368	31.328304	Beijing West Rd.

Table 4. Travel time reported from Auto Navi Map

Trip Mode		Car	Subway	Bus
Distance (km)		7.3	8	8.6
Travel Time (min)	Weekdays	14-17	31	43-50
	Weekends	14-16	32	44-50

Considering that traffic condition varies with time, travel time clustering can be conducted under four situations, which are peak hours on weekdays, off-peak hours on weekdays, peak hours on weekends, and off-peak hours on weekends. Since it is a rare case for a traveler to walk more than 7 km, we hypothesize that the records with computed travel time larger than 65 minutes are noise in the data and they are removed from the dataset. Therefore, three trip modes including subway, car and bus are identified by Log-GMM clustering, and we define K=3.

The parameters of Log-GMM are estimated in four different cases and the results are listed in Table 5. Figure 4 depicts the distribution of Log-GMM under different conditions. One can find the most trip mode between Pinglu Rd. and Beijing West Rd. is car. Subsequently, the subway ranked second, with the fewest number of activities on the bus. The clustering results using Log-GMM are listed in Table 6. Comparing the ratios of the three trip modes on weekdays and weekends, it can be seen that the number of car users on weekdays is less than the number of car users on weekends, while the number of users by public transit on weekdays is larger than the number of users by public transit on weekends. On working days, people may be more willing to take public transport due to the requirement to work on time and the shortage of parkinglots. However, on non-working days, most people do not have too many restrictions for entertainment, thus they would like to travel by car. As for the ratio of trip modes during peak hours and off-peak hours, it can be concluded that there is no much difference between them.



(a) peak hours on weekdays

(b) off-peak hours on weekdays

(c) peak hours on weekends

(d) off-peak hours on weekends

Figure 4. Travel time clustering with Log-GMM under four situations

Table 5. The estimated parameters of Log-GMM under four situations

		Weekdays			Weekends		
		k=1	***k=2***	***k=3***	***k=1***	***k=2***	***k=3***
Peak Hours	φ_k	0.64	0.34	0.02	0.61	0.34	0.05
	μ_k	2.96	3.12	3.75	3.04	2.91	3.65
	σ_k	0.01	0.02	0.03	0.01	0.03	0.10
Off-peak Hours	φ_k	0.49	0.48	0.03	0.70	0.26	0.04
	μ_k	2.94	3.10	3.75	2.92	3.05	3.80
	σ_k	0.02	0.02	0.04	0.01	0.02	0.05

Table 6. The clustering results using Log-GMM

Average Travel Time (min)	**Weekdays**			**Weekends**		
	Car	**Subway**	**Bus**	**Car**	**Subway**	**Bus**
Peak Hours	17.24	26.68	44.94	19.67	34.25	48.38
Off-peak Hours	16.06	25.71	44.26	19.15	31.27	47.34

The similar method was also proposed by Wang et al. (2010). They considered the probability to identify trip mode based on travel time collected from CDRs, and they utilized K-means unsupervised clustering algorithm to classify the samples. In this case, the results from the K-means clustering are listed in Table 7. To evaluate the performance of our method, we compare the travel time results with the travel time reported from Auto Navi Map. We define the error of trip mode identification as the differences between average time acquired from the clustering with Log-GMM and the travel time reported from Auto Navi Map. The errors are calculated for three methods (i.e., Log-GMM, GMM, K-means) as listed in Table 8, and one can see that clustering using Log-GMM and GMM perform better than using K-means algorithm. Moreover, Log-GMM performs better than GMM on weekdays while GMM slightly outperforms Log-GMM on weekends. Overall, Log-GMM is more suitable for trip mode inference.

Table 7. The clustering results using K-means

Average Travel Time (min)	**Weekdays**			**Weekends**		
	Car	**Subway**	**Bus**	**Car**	**Subway**	**Bus**
Peak Hours	19.12	27.31	42.93	20.22	30.01	46.19
Off-peak Hours	18.34	25.18	41.56	20.04	29.23	45.62

Table 8. The errors of different methods in Travel Time (min)

Weekdays	Trip mode			Error
	Car	Subway	Bus	
Auto Navi Map	15.5	31	46.5	/
Log-GMM	16.65	26.20	44.60	2.62
GMM	18.18	26.72	41.96	3.83
K-means	18.63	26.23	42.09	4.10
Weekends	**Trip mode**			**Error**
	Car	**Subway**	**Bus**	
Auto Navi Map	15	32	47	/
Log-GMM	19.41	32.76	47.86	2.01
GMM	19.29	30.57	46.89	1.94
K-means	20.11	29.64	45.97	2.83

In order to avoid fortuity, we make two more experiments, in which the information is listed in Table 9 and Table 10. The results of the experiments are shown in Table 11. We can find on average the error of Log-GMM is smaller than that of GMM and K-means, which means Log-GMM performs better. The largest error of Log-GMM is 6.32 min, which seems acceptable compared to the large variation of observed travel time.

Table 9. Information of two activity nodes of case 2 and case 3

	Pair of Nodes	NID	Longitude	Latitude	Location
Case 2	Activity Node 1	3817	121.312549	31.1926370	518, Xianxia Rd.
	Activity Node 2	1645	121.214973	31.3043600	Hongqiao Rd.
Case 3	Activity Node 1	164	121.605248	31.1428438	1926, Xiuya Rd.
	Activity Node 2	8562	121.571744	31.2519540	685, Deping Rd.

Table 10. Travel time reported from Auto Navi Map of case 2 and case 3

Travel Time (min)		Car	Subway	Bus
Case 2	Weekdays	12	23	27
	Weekends	11	25	30
Case 3	Weekdays	21	44	88
	Weekends	19	45	85

Table 11. Errors of different methods in Travel Time (min)

	Weekdays	**Trip mode**			**Error**
		Car	**Subway**	**Bus**	
Case 2	Auto Navi Map	12	23	27	/
	Log-GMM	13.41	22.14	25.73	1.18
	GMM	16.23	20.15	23.54	3.51
	K-means	16.64	21.53	22.84	3.42
Case 3	Auto Navi Map	21	44	88	/
	Log-GMM	21.87	42.54	71.35	6.32
	GMM	23.54	40.21	68.21	8.70
	K-means	22.58	39.12	64.25	10.07
	Weekends	**Trip mode**			**Error**
		Car	**Subway**	**Bus**	
Case 2	Auto Navi Map	11	25	30	/
	Log-GMM	18.34	26.12	32.58	3.68
	GMM	17.25	23.87	27.21	3.39
	K-means	16.58	21.26	26.14	4.39
Case 3	Auto Navi Map	19	45	85	/
	Log-GMM	24.67	44.25	75.19	5.41
	GMM	24.13	42.51	71.20	7.14
	K-means	23.51	40.15	68.11	8.75

Though the validations from Auto Navi Map present acceptable results overall, the actual inference prediction cannot be acquired. To assess the inference performance, another experiment is conducted and more data are acquired. In the experiment, two activity nodes (i.e., Jing'an Temple and Longyang Rd Motorway Interchange) are selected and the location information as well as travel time information is shown in Figure 5. Subway has its own special cell phone base stations. In the case, the cell phone base station information of subway is obtained, which means the subway mode can be identified through base stations. However, the modes of car and bus cannot be distinguished. Therefore, inference performance of subway can be assessed. The parameters of Log-GMM are estimated and the results are listed in Table 12. Figure 6 depicts the distribution of Log-GMM. The errors are also calculated for three methods as listed in Table 13, and clustering using Log-GMM performs better than using K-means algorithm and GMM. Moreover, the confusion matrices of trip mode inference using different methods are listed in Table 14. One can find that clustering using Log-GMM could infer subway mode with the recall of 53.77% and the precision of 84.25%, which outperforms GMM and K-means. The precision of inference using Log-GMM is acceptable. However, the recall of that is a bit low, which means that the method cannot identify most users by subway very well. Overall the trip mode inference using Log-GMM can achieve the accuracy of 72.68%, which is acceptable considering the sparseness and quality of mobile phone signaling data.

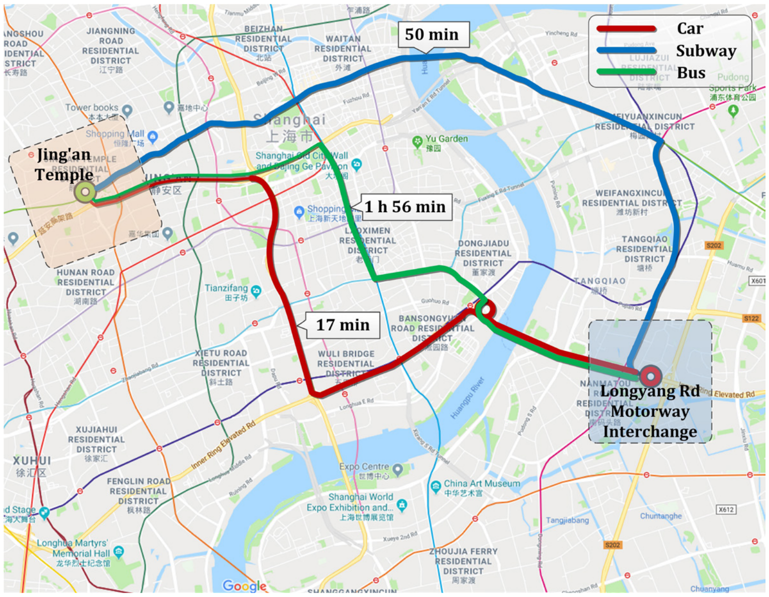


Figure 5. Detailed information of activity nodes

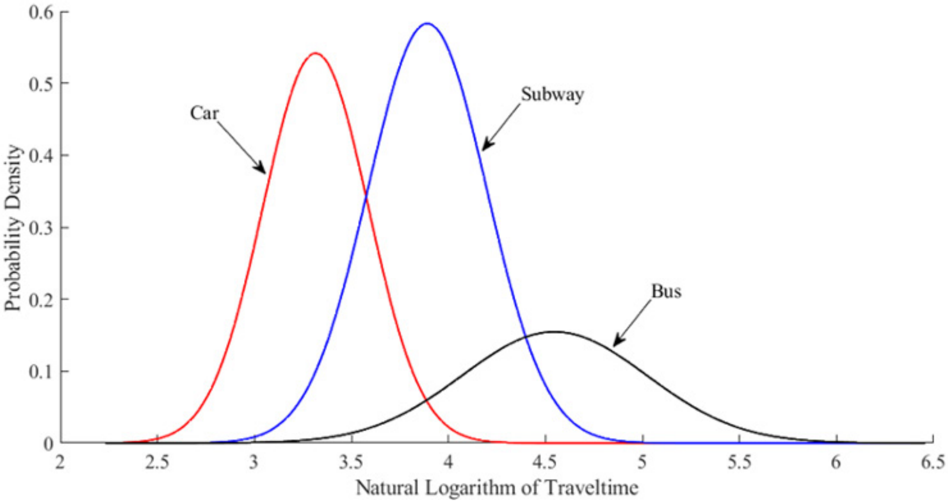


Figure 6. Travel time clustering with Log-GMM

Table 12. The estimated parameters of Log-GMM

	k=1	*k=2*	*k=3*
φ_k	0.37	0.45	0.18
μ_k	3.32	3.89	4.55
σ_k	0.07	0.09	0.23

Table 13. Errors of different methods in travel time (min)

Clustering Method	Car	Subway	Bus	Error
Auto Navi Map	17	50	116	/
Log-GMM	20.64	45.17	97.63	8.95
GMM	18.81	42.54	94.76	10.17
K-means	16.28	41.65	92.10	10.99

Table 14. Confusion matrices of trip mode inference using different methods

Actual modes	Inferred modes using Log-GMM		Recall
	Car and Bus	Subway	
Car and Bus	191	20	90.52%
Subway	92	107	53.77%
Precision	67.49%	84.25%	72.68%
Actual modes	**Inferred modes using GMM**		**Recall**
	Car and Bus	**Subway**	
Car and Bus	178	33	84.36%
Subway	105	94	52.76%
Precision	62.89%	74.02%	66.34%
Actual modes	**Inferred modes using K-means**		**Recall**
	Car and Bus	**Subway**	
Car and Bus	182	29	86.26%
Subway	102	97	51.26%
Precision	64.08%	76.98%	68.05%

5 Conclusion

The main goal of this paper is to infer trip mode from mobile phone signaling data. The proposed method includes activity nodes recognition, travel time computation, and clustering with Log-GMM. Then we select four different OD pairs to conduct experiments. In order to validate the accuracy of results, we firstly compare the computed travel time with travel time from Auto Navi Map. The largest error is 6.32 min, which is acceptable compared to the large variation of observed travel time. We also obtain the cell phone base station information of subway in one case, which is utilized to assess the trip mode inference performance. Moreover, we compare our method (Log-GMM) with the previous methods (GMM, K-means), and results indicate that Log-GMM performs better than GMM and K-means. The precision of inference using Log-GMM is acceptable. However, the recall of that is a bit low, which means that the method cannot identify most users by subway very well. Overall, trip mode inference using Log-GMM can be acceptable considering the sparseness and quality of mobile phone signaling data.

Trip mode identification is of importance for transportation planning and management. The proposed method has advantages of easy calculation and realization and has better performance in identifying bus, car, and subway, which can provide significant data reference for city transportation planning, construction, and operation.

In future, we plan to improve our method on the following aspects. Firstly, we will consider the traffic conditions corresponding to the trip. Secondly, we can extract velocity features of trips under the consideration of geographic information, which may be useful to improve the identification performance. Finally, with the wide open of base station information of subway, the trip mode inferred may only include car and bus, which will improve the current inference performance.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities under Grant No. 22120180241; Transportation Science and Technology Planning Project of Henan Province, China under Grant No. 2019G-2-2.

References

Anderson, I., & Muller, H. (2006). *Practical activity recognition using GSM data* (Technical Report CSTR-06-016). Bristol, England: Department of Computer Science, University of Bristol.

Ashqar, H. I., Almannaa, M. H., Elhenawy, M., Rakha, H. A., & House, L. (2018). Smartphone transportation mode recognition using a hierarchical machine learning classifier and pooled features from time and frequency domains. *IEEE Transactions on Intelligent Transportation Systems, 99*, 1–9.

Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers Environment & Urban Systems, 36*(6), 526–537.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Frendberg, M. (2011). *Determining transportation mode through cellphone sensor fusion.* Cambridge, MA: Massachusetts Institute of Technology.

Fosgerau, M., & Fukuda, D. (2012). Valuing travel time variability: Characteristics of the travel time distribution on an urban road. *Transportation Research Part C: Emerging Technologies, 24,* 83–101.

Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., & Perz, R. (2010). Automating mode detection for travel behavior analysis by using global positioning systems enabled mobile phones and neural networks. *IET Intelligent Transport Systems, 4*(1), 37–49.

Guessous, Y., Aron, M., Bhouri, N., & Cohen, S. (2014). Estimating travel time distribution under different traffic conditions. *Transportation Research Procedia, 3*, 339–348.

Li, Y., Yang, J. F., Zhang, N. F., Zhou, H. D., & He, J. R. (2015). A transportation mode identification method based on mobile phone positioning. *Advances in Transportation Studies, 1,* 175–186.

Nham, B., Siangliulue, K., & Yeung, S. (2008). *Predicting mode of transport from iPhone accelerometer data* (Technical report). Stanford, CA: Stanford University.

Nick, T., Coersmeier, E., Geldmacher, J., & Goetze, J. (2010). Classifying means of transportation using mobile sensor data. *International Joint Conference on Neural Networks, 41*, 1–6.

Rahman, M. M., Wirasinghe, S. C., & Kattan, L. (2018). Analysis of bus travel time distributions for varying horizons and real-time applications. *Transportation Research Part C: Emerging Technologies, 86*, 453–466.

Rasmussen, C. E. (1999). The infinite Gaussian mixture model. *Proceedings of the 12th International Conference on Neural Information Processing Systems,12,* 554–560.

Reddy, S., Burke, J., Estrin, D., & Hansen, M. (2008). Determining transportation mode on mobile phones. IEEE International Symposium on Wearable Computers, Sept. 28–Oct. 1, Pittsburgh, PA, 25-28. https://doi.10.1109/ISWC.2008.4911579

Reddy, S., Min, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks, 6*(2), 1–27.

Reichert, A., & Holz-Rau, C. (2015). Mode use in long-distance travel. *Journal of Transport and Land Use, 8*(2), 87–105.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation, 42*(1), 163–188.

Shin, D., Aliaga, D., Tunçer, B., Arisona, S. M., Kim, S., Zünd, D., & Schmitt, G. (2015). Urban sensing: Using smartphones for transportation mode classification. *Computers Environment & Urban Systems, 53,* 76–86.

Sohn, T., Varshavsky, A., Lamarca, A., Chen, M. Y., Choudhury, T., Smith, I., … & de Eyal, L. (2006). Mobility detection using everyday GSM traces. *Proceedings of the International Conference on Ubiq-*

uitous Computing, 4206, 212–224.

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011). Transportation mode detection using mobile phones and GIS information. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,* 54–63. https://doi.10.1145/2093973.2093982

Sun, L., Zhang, D., Li, B., Guo, B., & Li, S. (2016). Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. *International Conference on Ubiquitous Intelligence and Computing, 6406,* 548–562.

Vandenbroucke, D., Bucher, B., & Crompvoets, J. (2013). *Geographic information science at the heart of Europe.* New York: Springer.

Wang, H., Calabrese, F., Lorenzo, G. D., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. *International IEEE Conference on Intelligent Transportation Systems*, Sept. 2010, 318-323. https://doi.10.1109/ITSC.2010.5625188

Xiao, Z., Wang, Y., Fu, K., & Wu, F. (2017). Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *International Journal of Geo-Information, 6*(3), 57.

Xu, D., Song, G., Gao, P., Cao, R., Nie, X., & Xie, K. (2011). Transportation modes identification from mobile phone data using probabilistic models. *Proceedings of the International Conference on Advanced Data Mining and Applications, 7121*, 359371.

Young-Ji, B., Abdulhai, B., & Shalaby, A. (2009). Real-time transportation mode detection via tracking global positioning system mobile devices. *Journal of Intelligent Transportation Systems, 13*(4), 161–170.

Zahabi, S. A., Miranda-Moreno, L., Patterson, Z., & Barla, P. (2012). Evaluating the effects of land use and strategies for parking and transit supply on mode choice of downtown commuters. *Journal of Transport and Land Use, 5*(2), 103–119.

Zhang, L., Liu, L. J., Bao, S. N., & Qiang, M. T. (2015). Transportation mode detection based on permutation entropy and extreme learning machine. *Mathematical Problems in Engineering,* (2015), 1–10.