T-START: Time, Status and Region Aware Taxi Mobility Model for Metropolis

Haiquan Wang[1], Shuo Lei[1], Binglin Wu[1], Yilin Li[1], Bowen Du[2],
[1] School of Software, Beihang University
Beijing, 100191- China
[e-mail: whq@buaa.edu.cn]
[2] State Key Lab of Software Development Environment, Beihang University
Beijing, 100191- China
[e-mail: dubowen@buaa.edu.cn]
*Corresponding author: Bowen Du

Received March 26, 2017; revised November 6, 2017; accepted February 1, 2018; published July 31, 2018

Abstract

The mobility model is one of the most important factors that impacts the evaluation of any transportation vehicular networking protocols via simulations. However, to obtain a realistic mobility model in the dynamic urban environment is a very challenging task. Several studies extract mobility models from large-scale real data sets (mostly taxi GPS data) in recent years, but they do not consider the statuses of taxi, which is an important factor affected taxi's mobility. In this paper, we discover three simple observations related to the taxi statuses via mining of real taxi trajectories: (1) the behavior of taxi will be influenced by the statuses, (2) the macroscopic movement is related with different geographic features in corresponding status, and (3) the taxi load/drop events are varied with time period. Based on these three observations, a novel taxi mobility model (T-START) is proposed with respect to taxi statuses, geographic region and time period. The simulation results illustrate that proposed mobility model has a good approximation with reality in trajectory samples and distribution of nodes in four typical time periods.

Keywords: Mobility model, taxi status, region transition probability, urban vehicular networks

A preliminary version of this paper appeared in IEEE INFOCOM 2015, April 26-May 1, Hong Kong. This version includes a concrete analysis and adds time dimension to the START. This research was supported by the National Key R&D program under Grant No.2016YFC0801700, Beijing Municipal Science and Technology Project No Z171100000917016,the National Natural Science Foundation Project under Grant No. U1636208.

http://doi.org/10.3837/tiis.2018.07.004 **ISSN : 1976-7277**

1. Introduction

Transportation plays a critical role in building smart cities and supporting comprehensive urban informatics [1], and the use of Intelligent Transportation Systems is one of the key technologies for improving the safety, efficiency, and environmental friendliness of the transport industry [25], [28]. Specifically, more and more vehicles are connected to the Internet through vehicle-to-anything (V2X) communication technologies, changing the automotive industry and the transportation system [22]. Validation of mobile ad hoc network protocols relies almost exclusively on simulation [2], [3], [19], [21]. The value of the validation is, therefore, highly dependent on how realistic the mobility model (i.e., mobility pattern of vehicles, including speed and direction) used in the simulations are [23]. Thus, it is necessary to have a significant effort in increasing the realism of mobility models used by network simulators. In addition, realistic mobility model can be used for city planning, traffic control, and other important tasks of smart cities. For instance, Kong et al. [29] proposed a time-location-relationship (TLR) combined taxi service recommendation model to improve taxi drivers' profit.

In recent years, mobility models [4], [5], [7] have been well-studied, and these works can be classified into free space and constrained models based on the degree of randomness.

For the free space scenario, the Random Way Point (RWP) model [6] is the most commonly used in simulations of Vehicle Network. An early study [10] shows that RWP in many cases is a good approximation of the vehicular mobility model based on real street maps. However, compared with the free space scenario, constrained mobility models [11], [12] are much closer to the realistic mobility by taking the geographic structure (such as the street layout, traffic rules, and multi-lane roads) into consideration, which will reduce the accuracy of simulation of Vehicle Network. Recently, there is also a new trend to extract the vehicular mobility model from real vehicular trace data (mainly taxi GPS trace data) [8], [13]. For example, Huang et al. proposed mobility models by estimating three parameters (turn probability, road section speed and travel pattern) from Shanghai taxi trace data [8]. However, all these existing constrained mobility models are too complicated to implement and strongly related to the simplified maps, and that will reduce the time efficiency of simulation. Also, the existing taxi-based mobility models ignore the statuses of taxi (vacant or occupied), which has an important influence on the performance of taxi mobility models.

In this paper, we propose a *Time, STAtus and Region Aware Taxi mobility* model (T-START) from both macroscope and microscope aspects. In the macroscope, T-START can divide the area into two sets of regions according to the density of passenger load or drop events in different time periods instead of simply dividing the area into coarse-grain regions. When a taxi takes a passenger, the current location is selected from the set of load-event

regions at that time. And the destination region, where the drop event happens, is selected in the set of drop-event regions. For microscope, the speed of the taxi is generated based on its status at corresponding time, which is learned via statistical analysis. Extensive simulations are carried out to verify the effectiveness of T-START from three aspects: traces and node distributions, in/out-degree distributions and contacts characteristics. The results show that T-START model has a good approximation of the real scenario in trace samples.

Our contribution can be summarized as follows. First, we find that taxis' behaviors and geographic features are strongly related to the status of the taxi. In addition, time is another factor affecting the taxi behavior because the passenger flow volume, origins and destinations vary with time. Meanwhile, the demand of passengers affects the quantity of working taxies. We validate such claims by statistical analysis over a large-scale Beijing taxi trace data. Secondly, we propose T-START model based on above findings, which is much easier to implement and more accuracy in simulation of Vehicle Network comparing with other classical mobility models. Finally, we implement a prototype system based on ONE and the results of experiments show that our model enhances the accuracy and efficiency of the performance of simulation Vehicle Network. To the best of our knowledge, our work is original to develop mobility models by investigating taxi behaviors and geographic features of different statuses and time periods.

The rest of this paper is organized as follows. Section 2 summarizes the related work. And Section 3 provides the statistical results from real data to validate three important assumptions for T-START model. Section 4 presents the detail of T-START model. Simulation results are reported in Section 5. Finally, Section 6 concludes this paper.

2. Related Work

In this section, we summarize some reported studies about mobility models. They can be classified into free space and constrained models based on the degree of randomness.

2.1 Mobility model in free space scenario

Random Walk mobility model (RW) [24], Random Way Point mobility model (RWP) [6] and Random Direction mobility model (RD) [27] are three classical random mobility models. They establish their movements without prior knowledge and apply to the simplified mobile scenarios by random selection speed and direction of node movement. Among these three models, RWP is the most commonly used in many cases. The movement model identified a pause time, speed range from zero to the maximum, and movement area where the model select a random destination. Amit Kumar Saha el at. [10] found that RWP mobility model is a good approximation of the vehicular mobility model based on real street maps. Although these models defined simple mobility patterns, which is convenient for us to create mobility models and analysis, they are out of reality due to many practical factors are ignored.

2.2 Mobility model in constrained scenario

Due to the weaknesses in free scenario, many studies began to consider more restricted condition in mobility models. Most of them are mainly devided into three parts: map-based mobility model, traffic simulator-based mobility model and trace-based mobility model.

2.2.1 Map-based mobility model

Manhattan models [9] are a typical model which models the city as a Manhattan style grid, with a uniform block size across the simulation area, while all streets are two-way with a lane in each direction which constrained car movements [11], and nodes can move straight forward or turn direction at a cross road. Bhattacharjee, D et al. proposed a mobility model with multiple features [20]. Meanwhile, A. K. Saha and D. B. Johnson [10] model the vehicle networks based on the real roads. It was compared with the RWP models, a commonly used mobility model in vehicular networks, to find out the difference between the RWP and real trajectories in routing performance. D. R. Choffnes et al. [12] proposed an integrated mobility and traffic model for vehicular wireless networks. This paper simplifies the real road to evaluate the network performance in ad hoc and proposes the mobility model STRAW. It verified the RWP can not exhibit the characteristics of urban vehicle network.

2.2.2 Real trace-based mobility model

Some researches focus on the microscopic characteristics of mobility. They introduce the transportation features into mobility, such as the traffic lights, multi-channels and intersections. These geographical information can make the model more available. Atulya Mahajan, et al. [16] accounted for the street layout, traffic rules, multilane roads, acceleration-deceleration, and radio frequent (RF) attenuation due to obstacles, and further evaluated the synthetic maps by comparing with real maps. David R. Choffnes et al.[12] developed their movement model based on a realistic vehicular traffic model on road defined by real street map data. SAME [33] is a mobility model of daily activities which is based on the analysis and conclusion of students' habits and customs in campus environments. In addtition, Huang H et al. [8] proposed mobility models based on taxi trajectory data in Shanghai, China. They designed three parameters : transition probability, traffic speed in each section and travel pattern, which can be estimated by analyzing the data statistically. But these models are too complicated to re-implement this model, for the model is strongly related to the map they simplified from the real road.

2.2.3 Sociological behavior-based mobility model

In recent years, vehicular sensors or handheld devices spread rapidly, that makes it possible to collect and analyze the real trajectories of large amount of nodes. It helps us to improve the traffic and network macroscopically. Besides, Gao et al. [26] put forward a model based on the similarity of the user's interest. They abstracted the node's mobility patterns into three states, such as the main community, the other communities, and the path to which they are linked and then considered social relationship and the driving function of

the social activities of the nodes in the real life. Recently, Musolesi put forward a kind of community based mobility model combined with social network theory [31]. The model will be distributed in multiple communities in different regions according to the degree of closeness between nodes. Also, Social, sPatial, and Temporal mobility framework (SPoT) takes a social graph as input and the spatial and temporal dimensions of mobility are added [32].

3. Statistical Analysis of Taxi Traces

In this section, we focus on the statistical analysis on the speed, duration, and taxi event characteristics of the Beijing taxi data set, which is a large-scale urban vehicular trace data.

3.1 Trace Dataset: Beijing Taxi Traces

A real-world GPS data set was used for our analysis, which was collected from Beijing taxi companies. After removing replicates and wrong records caused by machine error, we are left with about 91 million records taken by 12,455 taxis within seven days from November 1 to November 7, 2011. In the data set, each record includes a base station ID, company name, taxi ID, timestamp, current location (including longitude and latitude), speed, event, status, et al. Besides, each taxi uploads the record in every 60 seconds. Of all the fields in the record, we extract the information that later study used as a tuple (*taxi ID*, *time stamp*, *longitude*, *latitude*, *status*, *event*). There are five types of events and four types of statuses of records in the data set, which are summarized in **Table 1**. Due to the rest of other statuses are not meaningful to our work, we only focus on the vacant and occupied status (corresponding load and drop event) in this paper. Note that GPS traces from taxis have been used recently for inferring human mobility [14] and modeling city-scale traffics [15]. Therefore, we believe that they are also suitable to be used to build mobility models in large-scale urban scenario.

Table 1. Event and status in Beijing taxi traces

Category	Code	Explanation
Event	0 (drop)	a taxi's status changes to vacant.
	1 (load)	a taxi's status changes to occupied.
	2	set up defense.
	3	cancel defense.
	4	no event happened.
Status	0 (vacant)	a taxi is vacant.
	1 (occupied)	a taxi is occupied.
	2	a taxi is setting up defense.
	3	stop running.

3.2 Three Claims on Taxi Behaviors

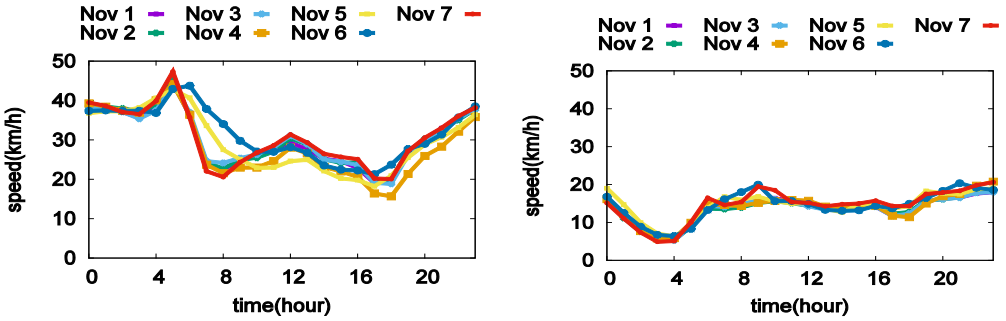
First, we proposed three claims based on the experience in our daily life, which are foundations of the taxi mobility model:

- **Claim 1:** The behavior of a taxi changes when its status updates. When a taxi is occupied, its destination is certain, and the vehicular speed of an occupied taxi accelerates relatively. In contrast, when a taxi is vacant, it slows down or even stops to search for potential passengers along the road. Therefore, taxi behavior characteristics, such as speed and status duration, vary consequently.
- **Claim 2:** Taxi has the different behavior in different time periods. One of most intuitive reflections of taxi behavior is a series of consecutive trips, where the trip is extracted by the load/drop event. Indeed, the quantities of load/drop events may vary with time conforming to certain rules. For example, the quantity of passengers late in the night is relatively fewer than that of passengers during the daytime. The correlation between taxi behavior and time may be reflected to following aspects:
  1) The hotspots of load/drop events vary with time.
  2) For the same time period during a day, the load/drop events distribute similar.
- **Claim 3:** The mobility behavior of taxis associates with geographic features. When a taxi is occupied, the destination may be tended to certain geographic places, such as the airport. Meanwhile, when a taxi is vacant, its driver tends to look for some hot spots, where more people want to take a taxi, such as downtown areas. Therefore,
  1) The destination selection of a taxi is influenced by different regions.
  2) Events occur in different regions un-evenly, passenger drop and load events are distinct.

Next, we analyze the speed, duration and passenger load/drop events distribution over the Beijing taxi trajectories to validate the three claims above.

3.3 Taxi Behavior Varied with Status

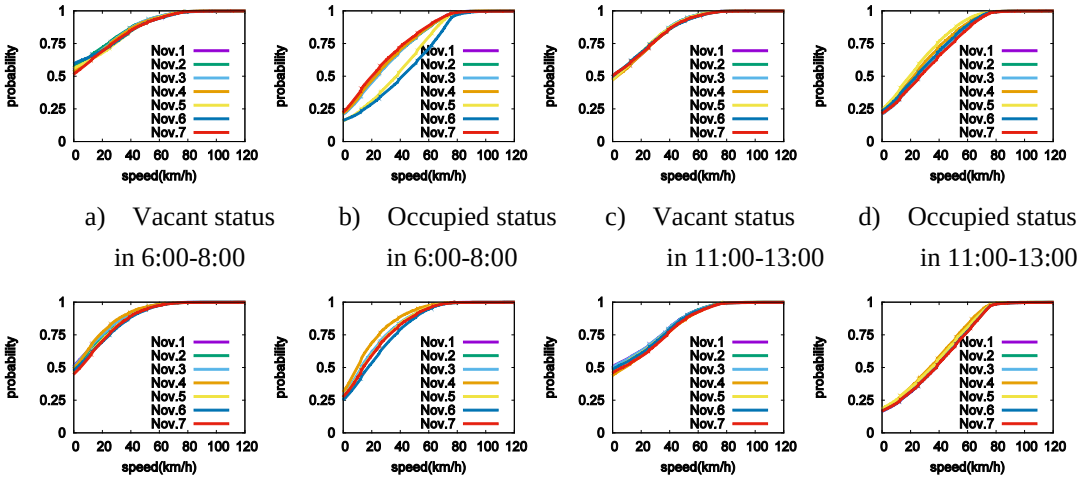
The average of instantaneous speed distributions for the two statuses for different time periods are explored.



a) Occupied status b) Vacant status

Fig. 1. Average of instantaneous speed of two statuses on each hour

We calculate the average of instantaneous speed on each hour for vacant and occupied status from November 1 to 7, 2011. As shown in **Fig. 1**, occupied taxis drives much faster than vacant one. And the average speed is affected by time, especially for the occupied status. To further investigate the cumulative speed distribution, we calculated and plotted the proportion for every speed section is in **Fig. 2**.



a) Vacant status in 6:00-8:00 b) Occupied status in 6:00-8:00 c) Vacant status in 11:00-13:00 d) Occupied status in 11:00-13:00

e) Vacant status in 17:00-19:00 f) Occupied status in 17:00-19:00 g) Vacant status in 22:00-24:00 h) Occupied status in 22:00-24:00

Fig. 2. Speed distributions for vacant and occupied statuses.

Specifically, the x-axis and y-axis represent the speed range of the car and the cumulative probability, respectively. For example, a point at (5, 0.2) presents 20% records fall in the speed range [0,5) km/h. We also fit the speed to model the microscope behavior (will be discussed in Section 4). **Fig. 2** shows that speed distribution differs for each status and with strong regularity for each status at corresponding time.

In **Fig. 2(b)**, from 6:00 to 8:00, curves of Nov. 5 and Nov. 6 are different from other curves. This may because Nov. 5 and Nov. 6 are weekend and more workers will get up late at weekend. So the vehicle volume in the weekend morning will decrease so that the speed will increase.

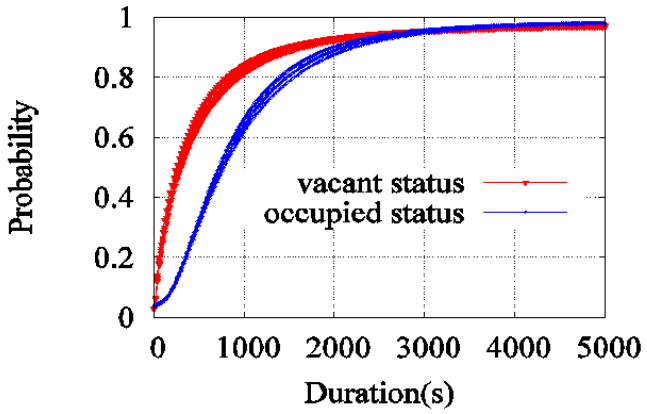


Fig. 3. Status duration distributions.

The duration distribution for each status is shown in **Fig. 3**. Status duration represents the time length of a taxi staying in a certain status. The red line presents the duration time distribution for vacant status, and the blue one is for occupied status. We can find that the red line (vacant status) approaches to 1 earlier than the blue line (occupied status). And the value of vacant duration is smaller than the value of occupied duration. This is reasonable since drivers tend to shorten the waiting time to raise their incomes.

Overall, the statistical results for both speed and status duration are consistent with **Claim 1**, that is, the behaviors of taxis are similar within each status while differ between the two statuses.

3.4 Taxi Behavior Varied with Time

In this section, we analyze the number of load (and drop) event happened on each hour. **Table 2** presents the results that the total volumes of load and drop events for a week are similar, close to 2.7million. And the maximum event number is much larger than the minimum event number.

From **Fig. 4**, we can find that the event quantity varied with time shows strong regularity and the curves of the load and drop events follows parallel rules. In addition, ranges of two types of events quantities at the same time are similar. Which is consistent with our experiences, due to the load and drop quantity should be in balance.

Table 2. Events quantity varied with time

Item	drop event quantity	load event quantity
Total quantity for a week	2,679,385	2,707,290
maximum of an hour	28,583	28,130
minimum of an hour	861	918
time of the peak value	Nov 4, 19:00-20:00	Nov 4, 19:00-20:00
time of the valley	Nov 3, 4:00-5:00	Nov 3, 4:00-5:00

The analysis results validate **Claim 2** and fit with our daily experience: the load event quantity equilibrates with the drop event quantity, and the event quantity at certain time presents certain regularity.

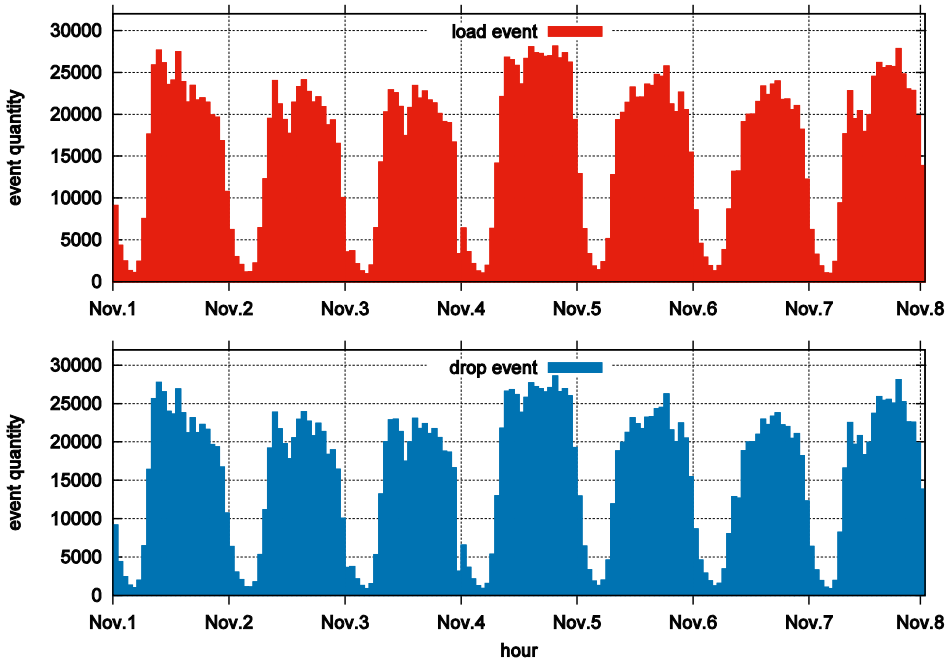


Fig. 4. Two events distributions of a week.

3.5 Taxi Behavior Varied with Geographic

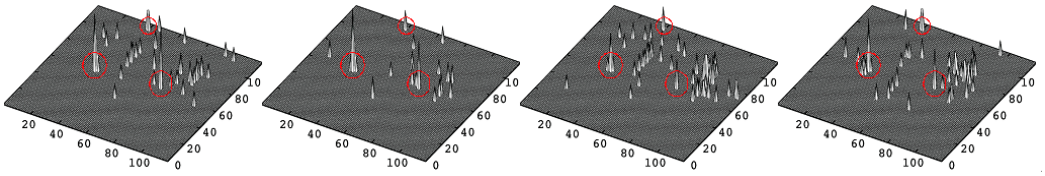
To capture the characteristics of events distributions with geographical preference, we divide regions into $200\text{m} \times 200\text{m}$ grids, and count the load and drop events happened in each hour. By filtering the cells whose event quantities are lower than 5 per hour, we found that the load and drop events tend to happen in different places, even though the event quantities and time periods are similar.

To further investigate features of event distributions, we select two days in workdays and weekends, respectively. **Fig. 5** and **Fig. 6** shows the hotspots (more than 20 events happened in one hour) of load/drop events at the rush hour (i.e., from 19:00 to 20:00), where each bar represents the number of happened events in the grid. Comparing the load and drop event hotspots, we can find the load event distributes much evenly than drop one. And some places are the hotspots of both load and drop events, as highlighted in the red circles.

Although the amounts of events are different from the workdays and weekends, the position of those hotspots are still similar. Because the load-event spots are mainly at homes

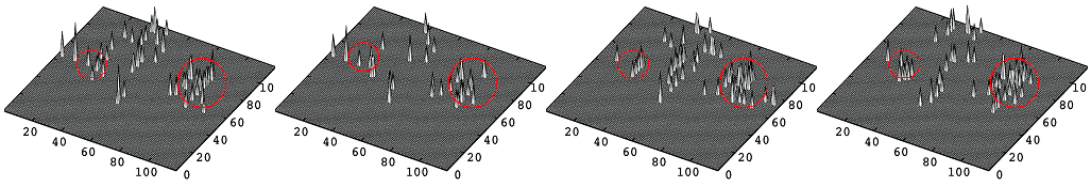
of the residents, while the drop-event spots tend to gather at workplaces, shopping malls, railway stations or scenic spots.

Overall, amounts of loading/dropping passengers in each cell shows geographic features: the distribution is uneven, and the difference between load/drop-event distributions illustrates the load/drop-event regions are different. All of these support **Claim 3** we given in the Section 3.2.



(a) Nov.1th, Tue. (b) Nov.3th, Thur. (c) Nov.5th, Sat. (d) Nov.6th, Sun.

Fig. 5. Hotspots of load events from 19:00 to 20:00



(a) Nov.1th, Tue. (b) Nov.3th, Thur. (c) Nov.5th, Sat. (d) Nov.6th, Sun.

Fig. 6. Hotspots of drop events from 19:00 to 20:00.

4. T-START Mobility Model

In this section, we provide technical details of modeling. Based on the features of taxis moving we extracted in the Section 3, we construct a Time, STAtus and Region Aware Taxi mobility model (T-START). There are two main tasks of T-START: destination selection and moving process.

4.1 Motivation

Movement model defines the mobility pattern of nodes, which can be represented as a collection of path segments denoted as $Paths :< p_1, p_2, ..., p_n >$. Therefore, generating p_i precisely becomes the key process of a good movement model. To generate a p_i, T-START takes two steps: destination selection and moving process.

Destination Selection: In T-START, Besides the influence of time, the selection of a node's destination is closely related to not only its current location but also its current status. A travel path of a taxi can be simplified as a multi-hop process, in which a hop indicates a load/drop event happened. Considering that, we first divide the whole area into regions by the density of passenger load/drop events at different time, respectively. This step will help us recognize load/drop region more scientifically and reduce the calculation capacity.

Second, based on the region which are recognized in the last step, we define a region transition probability to figure out the probability of the next hop falling in a certain region from the current region in the specified time. Therefore, we can construct the transition probability matrix in the specified time to select the next hop of the node. The specific process will be introduced in Section 4.2.

Moving Process: When the source location (current location) and destination location are selected, the next step is to find a path to connect them and simulate the speed of vehicles. First, although there are some approaches of dynamic path planning [30], which are very complicate to complement, due to main purpose of this paper is to establish a model which are more realistic but also much easier to complement, we adopt the Dijkstra algorithm as the path selection method of our model to simplify the process, which will find a shortest path from the source location to the destination based on the map. More specifically, this will not reduce the accurate of mobility model as moving features like direction and time costing have been already considered in extraction of taxi behaviors. Next, the speed of the path is assigned to *speed* based on current statuses. Here, the value of *speed* is drawn from historical speed distribution. Specifically, we fit the cumulative instantaneous speed distribution to get the cumulative probability distribution function of corresponding status, which will be introduced in the last subsection of this section.

4.2 Region Transition Probability

Due to the event distributions of load and drop events are different with each other and varied with time, region $R_{i,t}^{load}$ and $R_{j,t}^{drop}$ are recognized by different metrics, that is, drop or load event distribution during each time period. For instance, if the taxi is currently occupied, then the next hop event is the drop one. Hence, choosing a target region from a region set obtained based on drop event distribution is more logical. Here, we provide the following definitions.

Definition 1. *A cell $C_{x,y}$ is a set of consecutive geographic points, where x,y denotes the cell identifier; len_x and len_y are side length of the cell*; *lon and lat present longitude and latitude,respectively;*

$$C_{x,y} ::= \left\{ (lon, lat) \mid x \le \frac{lon}{len_x} < x+1, y \le \frac{lat}{len_y} < y+1 \right\}$$

Definition 2. *We consider a region R_m as a union of adjacent cells, and R_m is the smallest unit of transition probability, where m denotes the region identifier.*

$$R_m ::= \left\{ C_{i,j} \mid \exists C_{x,y} \in R_m \Rightarrow \|x-i\| \le 1, \|y-j\| \le 1 \right\}$$

The main idea of clustering cells to regions is merging adjacent cells whose event density is larger than an event threshold η into a same region. To avoid the size of a region become too large or too small, we set a limitation on the size of a region, which is $\|R_i\| \le \phi_{size}$, and also the number of final regions need to be less than or equal to ϕ_{top}.

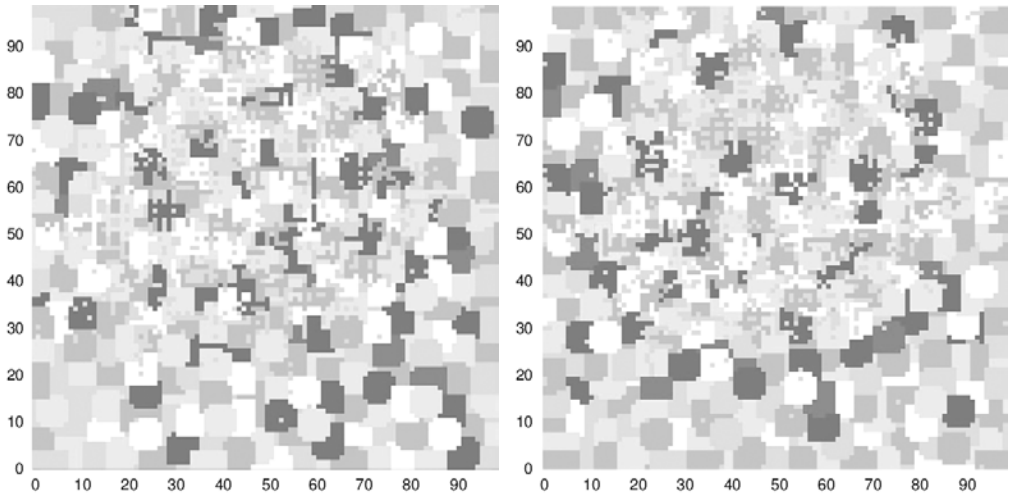
We first divide the whole area (within fourth ring roads in Beijing) into 100×100 cells, then sort all the cells by event density in descending order, and begin with the first cell to search its neighbors whether to join the same region or not using breadth traversal. After the top regions are formed, the other cells which do not belong to the top ϕ_{top} regions will also be clustered into regions, whose size should still be smaller than ϕ_{size}. Consequently, each cell will be clustered into regions and the size of each region are not larger than ϕ_{size}.

By clustering cells into regions, two region sets, $\mathbf{R}_t^{load}$ and $\mathbf{R}_t^{drop}$, can be recognized from the data set. For each time period, we set different threshold by its average events number in each cell at that time, that is, η equals to twice the average event number. ϕ_{top} is 200 and ϕ_{size} is 500 in all time periods, and the rest parameters settings of region recognition in each time period are showed in **Table 3**.

Table 3. Region recognition parameters

Item	0:00-8:59	9:00-12:59	13:00-20:59	21:00-23:59
n_{drop}	56	84	180	51
n_{load}	58	84	182	51

One of the region recognition results for load/drop events are shown in **Fig. 7**, which are the clustering regions from 9:00 to 12:59. In this figure, every colored block presents a region.



a) load event regions b) drop event regions

Fig. 7. Region recognition from 9:00 to 12:59.

Calculation of region transition probability:

We propose a region transition probability to figure out the probability of the next hop falling in a certain region from the current region.

Definition 3. *A transition probability from a load region i to a drop region j in time t is denoted as* $p^t_{load_i \to drop_j}$ *; Similarly, A transition probability from a drop region j to a load region i in time t is denoted as* $p^t_{drop_j \to load_i}$.

Since both transition probability can be calculated similarly, we only introduce the detailed one of $p^t_{load_i \to drop_j}$.

$$p^t_{load_i \to drop_j} = \frac{\left|\left\{taxi^{t'}_{load_i} \mid t \le t' < t + \Delta t\right\} \bigcap \left\{taxi^{t''}_{drop_j} \mid t' < t''\right\}\right|}{\left|\left\{taxi^{t'}_{load_i} \mid t \le t' < t + \Delta t\right\}\right|} \tag{1}$$

Where $taxi^t_{load_i}$ presents the taxi, which has the load event in region *i* in the time period *t*.

We restrict the time from current drop event record to next load event cannot be across more than one hour, that is the region *i* belongs to the region set of time t or time *t+1* and we ignore the records whose hour of timestamp is more than *t+1* hour. For example, for *t=7*, the record with timestamp 9:00:00 is invalid, while the record whose timestamp is 8:59:59 is valid.

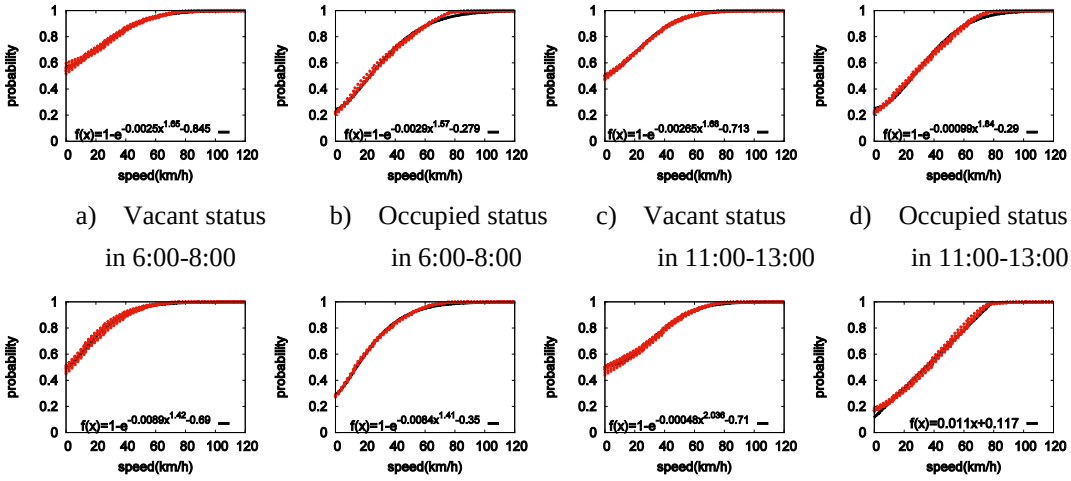
Then we can construct a region transition probability matrix during time period *t*, which is denoted as $P_{load \to drop}(t)$.

$$P_{load\to drop}(t)=\begin{pmatrix} p^t_{load_0\to drop_0} & p^t_{load_0\to drop_1} & \cdots & p^t_{load_0\to drop_m} \\ p^t_{load_1\to drop_0} & p^t_{load_1\to drop_1} & \cdots & p^t_{load_1\to drop_m} \\ \cdots & \cdots & \cdots & \cdots \\ p^t_{load_n\to drop_0} & p^t_{load_n\to drop_1} & \cdots & p^t_{load_n\to drop_m} \end{pmatrix} \tag{2}$$

4.3 Speed Distribution

To obtain the speed distribution of each status, we fit the cumulative instantaneous speed distribution to get the cumulative probability distribution function, and then take a derivative with it to obtain the speed probability distribution. From **Fig. 8**, the instantaneous speed distribution shows exponential law except the one that is occupied status from 22:00 to 24:00. Considering that, we fit the speed distribution by an exponential function $f_1(x)$, and fit the cumulative speed distribution of occupied status from 22:00 to 24:00 by a linear function $f_2(x)$, presented in Equation (3). In order to eliminate the influence caused by the weekend, we remove the speed distribution data such as the data of occupied status from 6:00 to 8:00, to generalize the fitting results.

$$\begin{cases} f_1(x)=1-1/\exp\left(-ax^b-c\right) \\ f_2(x)=ax+b \end{cases} \tag{3}$$



a) Vacant status in 6:00-8:00 b) Occupied status in 6:00-8:00 c) Vacant status in 11:00-13:00 d) Occupied status in 11:00-13:00

e) Vacant status in 17:00-19:00 f) Occupied status in 17:00-19:00 g) Vacant status in 22:00-24:00 h) Occupied status in 22:00-24:00

Fig. 8. fit result for taxi speed distribution

Here, $f_i(x)$ is the function form for the instantaneous speed distribution. The root mean square (rms) of residuals for each fit are reported in **Table 4**. The smaller rms of residuals

means the better fitting. In this table, the values are all less than 0.025, reflecting a good similarity.

Table 4. Parameters and rms of residuals of fitting curves

Time period	Vacant status	Occupied status
06:00-08:00	0.0129207	0.0198180
11:00-13:00	0.0086617	0.0204889
17:00-19:00	0.0176578	0.0105868
22:00-24:00	0.0154822	0.0240426

5. Model Verification

In this section, T-START mobility model is validated on the aspects of node distribution compared with existing mobility models and the real traces.

5.1 Experiment Settings

In order to confirm the effectiveness of our models, we picked the following basic simple mobility models for comparison:

- **Real trace:** the real taxi moving trajectory data (Nov. 1, 2011 to Nov. 7, 2011).
- **Random Way Point (RWP) model:** a classical mobility model is commonly used in simulation of as hoc.
- **Shortest Path (SP) model:** a mobility model based on the underlying map of Beijing where vehicles move along the map roads by Dijkstra algorithm to random destinations.

To evaluate our model from different aspects, we adopt the following three features:

- **Trace and node distribution:** Trace and their node distribution snapshots are the most intuitive display for demonstrating the efficiency of the mobility model.
- **In-degree and out-degree:** The in-degree (out-degree) figures out the number of taxies moving in (out) from a region during a time period. It can reflect dynamic node distributions and evaluate the model in the dynamic aspect [18].
- **Contacts Characteristics:** Contact is a concept used in Delay Tolerant Network (DTN), ad hoc networks, and can be defined as a communication opportunity. Therefore, the contact time and inter contact time among vehicles are also evaluated as the indicators to validate the similarity.

All mobility models are implemented on Opportunistic Networking Environment (ONE) [17]. And other related settings are showed in **Table 5**.

Table 5. Simulation parameters of ONE

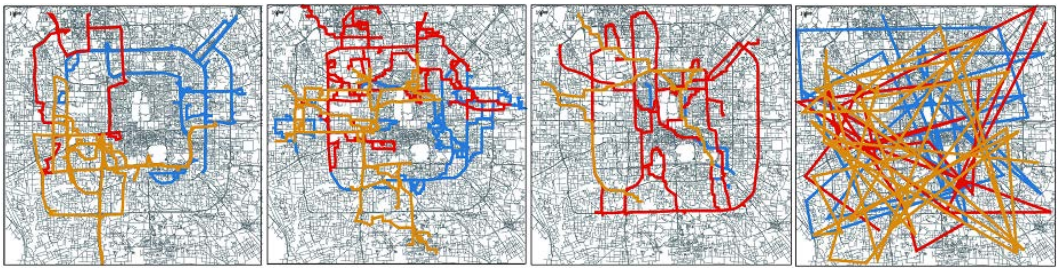
Map size	24×24 km^2	
Simulation time	2 h	
Number of vehicles	3000	
Transmit range	50 m	
Speed range	06:00-08:00	[1.8, 31.644] km/h
	11:00-13:00	[1.8, 36.864] km/h
	17:00-19:00	[1.8, 30.492] km/h
	22:00-24:00	[0.5, 40.921] km/h

5.2 Performance Comparison

5.2.1. Traces and node distributions

Trace samples and their node distribution snapshots from different mobility models are reported in **Fig. 9** and **Fig. 10**. From **Fig. 9** we can find that the traces of the real data and T-START only cover some parts of the area, while the traces of SP and RWP almost go through the whole area. Recall that SP and RWP select a destination randomly in the area, while T-START takes the associations between current region and destinations into consideration (which satisfies the movement rules of taxis).

In **Fig. 10**, real trace, T-START and SP exhibit the road structures, while the node distribution of RWP is much uniform. As to T-START, the destination section process decides that it tends to select a destination in the regions with higher load/drop event probability. Therefore, with the decline of the randomness, the snapshot of T-START becomes much clear and centralized on the main roads, which matches real traces very well.



(a) Real Trace (b) T-START (c) SP (d) RWP

Fig. 9. Trace samples.

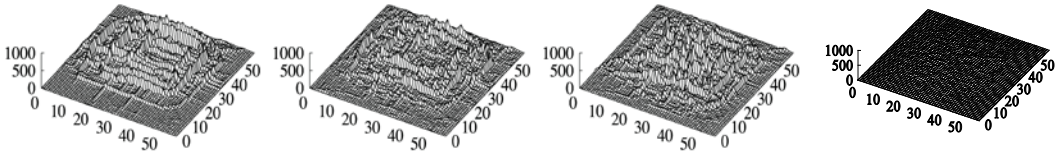
(a) Real Trace (b) T-START (c) SP (d) RWP

Fig. 10. Nodes distribution snapshots.

5.2.2. In-degree and out-degree

Since the node distribution has a great impact on the transport and network performance, a good understanding of it can help to route and control. However, nodes are dynamic leading to a dynamic node distribution. In order to quantify the changing node distribution, we introduce the in/out degree. The in/out degree figures out how many taxies moving in or out from a region in a time period. In/out degree defines how many nodes moving in or out an area during a period of time.

We divide the simulation scenario into grids of $400m \times 400m$ to investigate the in/out degree, and the time period to measure the in/out degree is as two hours according to the simulation time. **Fig. 11** shows the in-degree distributions for the real trace, T-START, SP and RWP.



(a) Real Trace (b) T-START (c) SP (d) RWP

Fig. 11. In-degree distributions for the real trace, T-START, SP and RWP.

As shown in **Fig. 11**, the hotspots of the real traces and T-START are concentrated on the main roads. As for the result of SP and RWP, both of them chose the destination in a random way, the difference between them is that peaks of SP gather in the central city and RWP has unobvious visiting hotspots. Because SP will choose the shortest way to a destination using the Dijkstra algorithm, while RWP choose the route randomly.

Moreover, we adopt the error rate (ER) to measure the performance. Specifically,

$$ER = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{\left| \hat{d}_i - d_i \right|}{d_i} \tag{4}$$

where $\hat{d}_i$ is the simulate value while d_i is the real trace value. **Table 6** shows the result of three models in different time period. The ER of T-START is about 0.48, while that of SP is about 0.65 and RWP is more than 0.8 for every time period. T-START has the best performance comparing with other two models in all time periods.

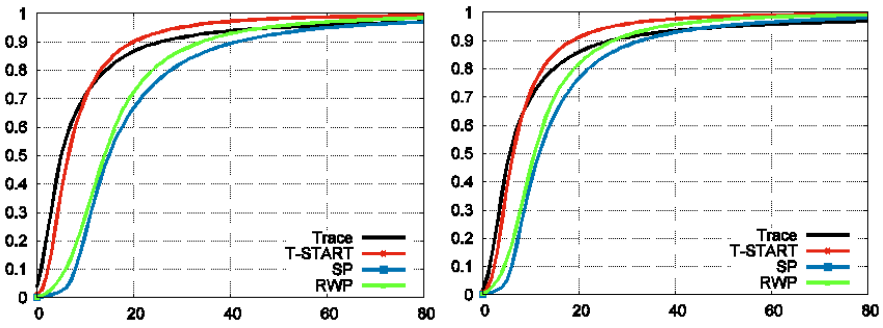
Table 6. Performance comparison of in-degree and out-degree

Model	In-degree				Out-degree			
	06:00-08:00	11:00-13:00	17:00-19:00	22:00-24:00	06:00-08:00	11:00-13:00	17:00-19:00	22:00-24:00
TSTART	**0.4927**	**0.4888**	**0.4694**	**0.4812**	**0.4952**	**0.4908**	**0.4730**	**0.4842**
SP	0.6923	0.6783	0.6533	0.6840	0.6903	0.6766	0.6515	0.6821
RWP	0.8037	0.8360	0.8382	0.8177	0.8030	0.8371	0.8380	0.8179

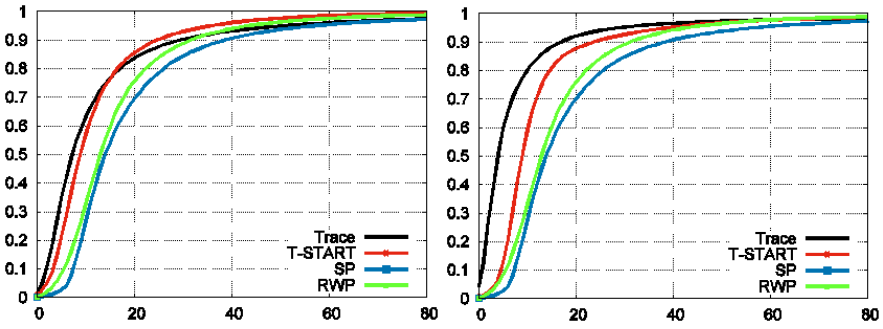
5.2.3. Contacts characteristics

The contact time and inter contact time among vehicles are also evaluated as the indicators to validate the similarity. **Fig. 12** and **Fig. 13** report the cumulative contact and inter-contact time distributions, respectively. In these figures, the x-axis and y-axis represent the time period(s) and the cumulative probability, respectively. Clearly, T-START matches the real traces best among three mobility models in all time periods. The performance of SP and RWP are similar may be caused by random destination selections.

From contact time and inter contact time, we can find that T-START simulates actual trajectories better. Mainly due to it choose the destination based on the transition probability matrix, which is constructed from historical trip data. Although SP used the real map and the speed of vehicles, its characteristics of contact are restrained by random destination selections.

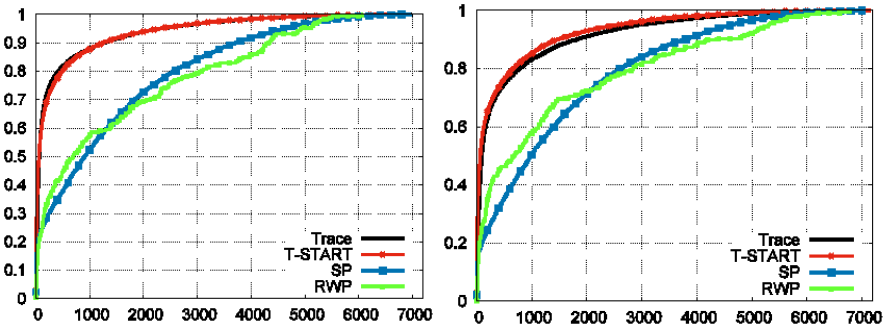


a) 6:00-8:00 b) 11:00-13:00

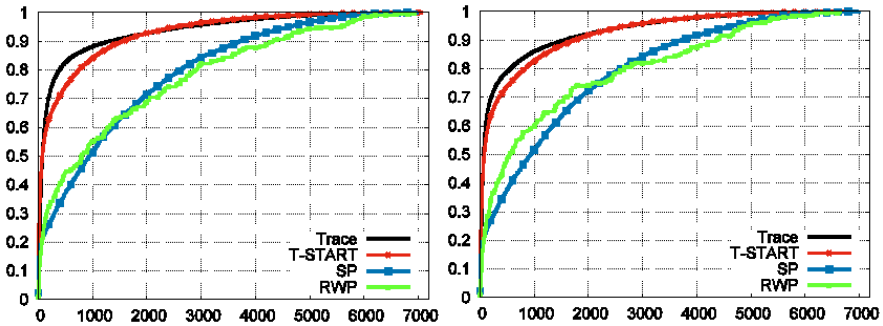


c) 17:00-19:00 d) 22:00-24:00

Fig. 12. Cumulative contact time distribution in different time period



a) 6:00-8:00 b) 11:00-13:00



c) 17:00-19:00 d) 22:00-24:00

Fig. 13. Cumulative inter contact time distribution in different time period

6.Conclusion

In this paper, we proposed a new mobility model T-START based on real taxi GPS data. By assuming the taxi behavior is related with its statuses, time and geographic features, statistical experiments are conducted to demonstrate those assumptions using the real trace data. With carefully estimations of the speed distribution of each status for different time

periods and the region transition probability between drop and load event regions, T-START considers both macroscopic and microscopic movements. For the macroscopic movements, a node moves and switches between load-event regions and drop-event regions. Then the microscopic movements (such as the speed for each status in the corresponding time period) can be applied. T-START is implemented and evaluated in ONE simulator by comparing with the real trace, RWP and SP mobility models. For node distribution, in/out-degree and contact features, T-START shows better performance than the other two mobility models. This demonstrates that T-START has a good approximation with reality and can be used for urban vehicular network research and applications.

References

[1] M. Michael D., and E J. Miller,"Urban transportation planning: a decision-oriented approach," 2001. Article (CrossRef Link)

[2] H. Wang,X. Ma,C. Xia, et al."A Modeling Approach of Mobile Ad Hoc Networks Survivability Model," in *Proc. of Information Science and Engineering (ICISE), 2009 1st International Conference on*. IEEE, 2456-2460, 2009. Article (CrossRef Link)

[3] C. Xia, D. Liang, H. Wang, et al."Characterization and modeling in large-scale urban DTNs," in *Proc. of Local Computer Networks (LCN), 2012 IEEE 37th Conference on*. IEEE, 352-359, 2012. Article (CrossRef Link)

[4] X. Lu, Y.-c. Chen, I. Leung, Z. Xiong, and P. Lio, "A novel mobility model from a heterogeneous military MANET trace," in *Proc. of 7th International Conference on Ad-hoc, Mobile and Wireless Networks (ADHOC-NOW)*, 2008. Article (CrossRef Link)

[5] S. Ahmed, G. C. Karmakar, and J. Kamruzzaman, "An environmentaware mobility model for wireless ad hoc network," *Computer Networks*, vol. 54, no. 9, pp. 1470–1489, 2010. Article (CrossRef Link)

[6] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proc. of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, 1998. Article (CrossRef Link)

[7] H. Wang, W. Yang, J. Zhang, et al. START: Status and Region Aware Taxi Mobility Model for Urban Vehicular Networks," in *Proc. of The First International Workshop on Smart Cities and Urban Informatics 2015*, 2015. Article (CrossRef Link)

[8] H. Huang, Y. Zhu, X. Li, M. Li, and M.-Y. Wu, "Meta: A mobility model of metropolitan taxis extracted from gps traces," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, 2010. Article (CrossRef Link)

[9] F. Bai, N. Sadagopan, and A. Helmy."Important: a framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networksi," in *Proc. of Proceedings of INFOCOM 2003*, San Francisco, CA, April 2003. Article (CrossRef Link)

[10] A. K. Saha and D. B. Johnson, “Modeling mobility for vehicular ad-hoc networks,” in *Proc. of the 1st ACM International Workshop on Vehicular Ad Hoc Networks*, 2004. Article (CrossRef Link)

[11] F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, “Citymob: a mobility model pattern generator for vanets,” in *Proc. of IEEE International Conf. on Communications Workshops*, 2008. Article (CrossRef Link)

[12] D. R. Choffnes and F. A. N. E. Bustamante, “An integrated mobility and traffic model for vehicular wireless networks,” in *Proc. of the 2nd ACM international workshop on Vehicular ad hoc networks*, 2005. Article (CrossRef Link)

[13] M. Kim, D. Kotz, and S. Kim, “Extracting a mobility model from real user traces.” in *Proc. of 25th IEEE International Conference on Computer Communications (INFOCOM)*, 2006. Article (CrossRef Link)

[14] R. Ganti, M. Srivatsa, A. Ranganathan, and J. Han, “Inferring human mobility patterns from taxicab location traces,” in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013. Article (CrossRef Link)

[15] J. Aslam, S. Lim, X. Pan, and D. Rus, “City-scale traffic estimation from a roving sensor network,” in *Proc. of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys)*, 2012. Article (CrossRef Link)

[16] Mahajan A, Potnis N, Gopalan K, et al.“Modeling vanet deployment in urban settings,”in *Proc. of Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*. ACM, 151-158, 2007. Article (CrossRef Link)

[17] A. Keraen, J. Ott, and T. Karkkainen, “The ONE simulator for DTN protocol evaluation,” in *Proc. of the 2nd International Conference on Simulation Tools and Techniques*, 2009. Article (CrossRef Link)

[18] P. Basu, N. Khan and T. D. Little, "A mobility based metric for clustering in mobile ad hoc networks,": IEEE, pp. 413--418, 2001. Article (CrossRef Link)

[19] D. L O Pez and A. E. L. Lozano, "Techniques in Multimodal Shortest Path in Public Transport Systems," *Transportation Research Procedia*, vol. 3, pp. 886--894, 2014. Article (CrossRef Link)

[20] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, "Combining multiple feature extraction techniques for handwritten devnagari character recognition,": IEEE, pp. 1--6, 2008. Article (CrossRef Link)

[21] Xie L F, Chong P H J. “Performance Improvement of Delay-Tolerant Networks with Mobility Control under Group Mobility.[J],” TIIS,9(6):2180-2200, 2015. Article (CrossRef Link)

[22] Silva C M, Masini B M, Ferrari G, et al. “A Survey on Infrastructure-Based Vehicular Networks[J],” *Mobile Information Systems*, (2017-8-6), 2017, 2017(6) , 2017. Article (CrossRef Link)

[23] Gramaglia, Marco, Fiore, Marco. “Highway Road Traffic Modeling for ITS Simulation[M],” *Networking Simulation for Intelligent Transportation Systems: High Mobile Wireless Nodes*. John Wiley & Sons, Inc. 2017. Article (CrossRef Link)

[24] Johnson D B, Maltz D A. “Dynamic Source Routing in Ad Hoc Wireless Networks[C],” *Mobile Computing*. 153-181, 1996. Article (CrossRef Link)

[25] Bazzi A, Masini B M, Pasolini G, et al. “Telecommunication systems enabling real time navigation[C],” in *Proc. of International IEEE Conference on Intelligent Transportation Systems.* IEEE, 1057-1064, 2010. Article (CrossRef Link)

[26] Gao, Y., Wang, S., Sun, J.: “Node mobility model based on user interest similarity,” *Journal of Computer Applications*, 35(9): 2457-2460(in Chinese) (2015) Article (CrossRef Link)

[27] Royer E M, Melliarsmith P M, Moser L E. “An analysis of the optimum node density for ad hoc mobile networks[C],” in *Proc. of IEEE International Conference on Communications. IEEE*, 857-861 vol.3, 2001. Article (CrossRef Link)

[28] Bazzi A, Masini B M, Zanella A. “Immediate feedback to increase the throughput of full duplex networks based on IEEE 802.11p[C],” in *Proc. of International Conference on ITS Telecommunications. IEEE*, 1-5, 2017. Article (CrossRef Link)

[29] Kong X, Xia F, Wang J, et al. “Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data[J],” *IEEE Transactions on Industrial Informatics*, PP(99):1-1, 2017. Article (CrossRef Link)

[30] Castro C D, Leonardi G, Masini B M, et al. “An Integrated Architecture for Infomobility Services - Advantages of Genetic Algorithms in Real-time Route Planning.[C],” in *Proc. of Icec 2010 - Proceedings of the International Conference on Evolutionary Computation. DBLP*, 300-305, 2010. Article (CrossRef Link)

[31] Zhang S, Yao M H, Wang X, et al. “Survey on Mobility Model of Opportunistic Networks[C],” in *Proc. of International Conference on Wireless Communication and Sensor Networks*. 2017. Article (CrossRef Link)

[32] Karamshuk D, Boldrini C, Conti M, et al. “SPoT: Representing the social, spatial, and temporal dimensions of human mobility with a unifying framework[J],” *Pervasive & Mobile Computing* 11(6):19-40, 2014. Article (CrossRef Link)

[33] Zhu X, Bai Y, Yang W, et al. “SAME: A students' daily activity mobility model for campus delay-tolerant networks[C],” *Communications*. IEEE, 528-533, 2012. Article (CrossRef Link)

Haiquan Wang received the PhD degree in computer science from the Beihang University in 2013. Now, he is an associate professor of Beihang University, Beijing, China. His research interests focus on Intelligent Transport System and Software Engineering. He has been conducting researches on Intelligent Transport System in recent years, hosting or participating many national projects including National Nature Science Foundation of China, National High Technology Research and Development Program of China (863 Program).

Shuo Lei received the bachelor's degree from Beihang University, Beijing, China, in 2015. She is working toward the master's degree in software engineering at Beihang University. Her research interests include spatial data mining and machine learning.

Binglin Wu received the bachelor's degree from Beihang University, Beijing, China, in 2015.And he also got the master's degree in software engineering at Beihang University. His research interests mainly focus on big data and machine learning.

Yilin Li, received the bachelor's degree from Beihang University, Beijing, China, in 2016. She is working toward the master's degree in software engineering at Beihang University. Her research focuses on traffic data analysis.

Bowen Du received his B.S. degree from Shijiazhuang Tiedao University, China, and his Ph.D. degree in computer science and engineering from Beihang University, Beijing, China, in 2005 and 2013, respectively. Currently, he is an assistant professor with the State Key Laboratory of Software Development Environment, Beihang University. His research interests include smart city technology, traffic data mining, and data service.