Federation University

Federation University ResearchOnline

https://researchonline.federation.edu.au

Copyright Notice

This paper is a postprint of a paper submitted to and accepted for publication in IEEE Transactions on Vehicular Technology and is subject to Institution of Engineering and Technology Copyright. The copy of record is available at the IET Digital Library.

Kong, X., Chen, Q., Hou, M., Rahim, A., Ma, K., & Xia, F. (2022). RMGen: A Tri-Layer Vehicular Trajectory Data Generation Model Exploring Urban Region Division and Mobility Pattern. *IEEE Transactions on Vehicular Technology, 71*(9), 1–1.

Avalable at: https://doi.org/10.1109/TVT.2022.3176243

See this record in Federation ResearchOnline at:

http://researchonline.federation.edu.au/vital/access/HandleResolver/1959.17/188467

RMGen: A Tri-Layer Vehicular Trajectory Data Generation Model Exploring Urban Region Division and Mobility Pattern

Xiangjie Kong, *Senior Member, IEEE,* Qiao Chen, Mingliang Hou, Azizur Rahim, Kai Ma, and Feng Xia, *Senior Member, IEEE*

***Abstract*—As an important branch of the Internet of Things (IoT), the Internet of Vehicles (IoV) has attracted extensive attention in the research field. To deeply study the IoV and build a vehicle spatiotemporal interaction network, it is necessary to use the trajectory data of private cars. However, due to privacy and security protection policies and other reasons, the data set of private cars cannot be obtained, which hinders the research on the social attributes of vehicles in the IoV. Most of the previous work generated the same type of data, and how to generate private car data sets from various existing data sets is a huge challenge. In this paper, we propose a tri-layer framework to solve this problem. First, we propose a novel region division scheme that considers detailed inter-region relations connected by traffic flux. Second, a new spatial-temporal interaction model is developed to estimate the traffic flow between two regions. Third, we devise an evaluation pipeline to validate generation results from microscopic and macroscopic perspectives. Qualitative and quantitative results demonstrate that the data generated in heavy density scenarios can provide strong data support for downstream IoV and mobility research tasks.**

***Index Terms*—Trajectory data, dataset generation, region division, spatial-temporal interaction, mobility pattern.**

I. Introduction

With more vehicles getting connected to the web of things, traditional vehicle autonomous network transitions to the Internet of Vehicles (IoV) [1]. IoV takes the moving vehicle as the information perception object and uses a new generation of information and communication technology to realize the network connection between vehicles and vehicles, vehicles and people, vehicles and roads, and vehicles and service platforms [2]. IoV is a core component of future intelligent transportation systems and will provide new research directions for smart cities, intelligent transportation, and social networks [3] [4] [5].

The vehicular mobility datasets which include the private car data and floating car data as a cornerstone of IoV, make various kinds of research conceptions possible. Making it possible to respond to traffic conditions timely and rapidly. For example, when an emergency occurs, drones can be used to communicate with ground vehicles to ensure smooth rescue by rescue teams [6]. Making it possible to travel orderly and safely. For example, based on IoV-related technologies, collecting information related to vehicles and driving environments, making traffic flow predictions can help plan public travel routes [7] [8]. Making it possible to evaluate new communication protocols for vehicular networks. For example, the construction of traffic scenarios using real information from various data sources to generate traffic demand enables the evaluation and testing of new network protocols [9] [10]. However, due to privacy protection policies and security restrictions, the GPS data of private cars cannot be obtained, which hinders the development of related work.

Although private car trajectory data is difficult to obtain, taxi data can be easily obtained from the Internet. Online taxi-hailing applications such as Didi and Uber have had a huge impact on human travel, attracting a large number of commuters to choose taxi-hailing. Taxi travel from a specific destination to a specific destination reflects people's movement patterns. Imagine this situation. When the taxi is empty, it will wander aimlessly in a busy area, waiting for passengers, or just park in some densely populated areas. In this case, taxi travel is meaningless. When a taxi is carrying passengers, its driving route is similar to that of an ordinary private car. For example, on weekends, people driving to attractions or shopping centers will have the same starting point and ending point as tourists taking taxis. The taxi data set contains the moving trajectories of empty cars or passengers. Because taxis use a similar movement mode as private cars when carrying passengers, this article is only interested in the trajectory of taxis when carrying passengers.

This paper proposes a tri-layer private vehicular trajectory data generation model based on urban region division and car mobility pattern (RMGen). This model consists of the preparation layer, generation layer, and verification layer, generating private cars' trajectory data based on the taxi GPS data and

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was supported in part by the National Natural Science Foundation of China under Grant 62072409, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR21F020003, and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-B2020001. (Corresponding author: Feng Xia.)

Xiangjie Kong is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China, and also with the School of Software, Dalian University of Technology, Dalian 116620, China (e-mail: xjkong@ieee.org).

Qiao Chen, Mingliang Hou, and Kai Ma are with the School of Software, Dalian University of Technology, Dalian 116620, China (e-mail: qiaochen2020@outlook.com; teemohold@outlook.com; makai1163801756@outlook.com).

Azizur Rahim is with the Department of Computer Systems Engineering, the University of Engineering and Applied Sciences, Swat, Pakistan (aziz.rahim@seecs.edu.pk).

Feng Xia is with Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, VIC 3353, Australia (e-mail: f.xia@ieee.org).

urban vehicular network information. We do the trajectory data preprocessing and region division in the preparation layer. In the generation layer, we construct the spatial-temporal interaction model based on the mobility pattern and generate private car trajectory data by SUMO. In the verification layer, we validate the accuracy of the generated data from the macroscopic and microscopic perspectives.

The contributions of this paper are summarized as follows:

- We study a novel trajectory data generation problem for private cars. One type of dataset is utilized to generate another type of dataset by considering their social connections.
- A tri-layer trajectory dataset generation framework is presented which consists of three components: a novel urban functional area division part by considering detailed inter-region correlations connected by traffic flow, a universal mobility model based on analyzing of urban vehicle mobility pattern, an evaluation part to validate the generation from macroscopic and macroscopic perspectives.
- We demonstrate the effectiveness of the proposed tri-layer framework. Qualitative and quantitative experimental results show the generated trajectory dataset of private cars can support downstream research tasks.

The rest of this paper is organized as follows. Related work is summarized in Section II. We describe the framework of the RMGen model in Section III. Section IV presents the preparation layer of our vehicular trajectory dataset generation model in detail. In Section V, we point out the generation layer of our model. Section VI illustrates the verification layer and the experimental result based on our generation datasets. In Section VII, we conclude our work and point out the future direction.

II. Related Work

In this section, we give a brief review of the related work, including the urban region division method, various models of human mobility, the technology we used in the simulation, and some trajectory data generation works which are useful for our study.

A. Urban Region Division

With the gradual emergence of urban functional areas. How to divide these urban functional areas is the basic premise of policy formulation, resource allocation, and social recommendation research.

Some studies show that human mobility can describe the functions of regions. Qi *et al.* [11] find that get-on/off amount in a region can depict the social activity dynamics in that area. In their study, regions are rigid squares that may not represent an intact region in the city, and they consider regions with pure social function in this paper. In contrast, regions are more complicated in reality. In [12], authors present a novel method for delineating urban functional areas based on building-level social media data. A dynamic time warping (DTW) distance-based k-medoids method is subsequently applied to group buildings with similar social media activities into functional areas. But human mobility is not taken into consideration. Wang *et al.* [13] propose a framework for automated urban function zoning, which is based on VGI geo-tagged photos and OpenStreetMap (OSM) data. However, the above method cannot describe composite functional areas. Yuan *et al.* [14] propose discovering regions of different functions (DRoF) using both human mobility among regions and points of interest (POI) located in a region. The framework inferred the functions of each region using a topic-based inference model witch regards a region as a document, a function as a topic, categories of POIs as metadata, and human mobility patterns as words; it reduces the data sparseness problem before clustering regions.

However, in the present study, the division of functional areas is not detailed enough. Due to the large area divided, the inter-regional traffic flow rules are not reasonable when studying these areas. To obtain more exact regional division results, we propose the ARS regional division method.

B. Vehicle Mobility Models

Mobility modeling is a challenging issue in IoV [15]. For instance, analysis of human mobility patterns based on multi-source large-scale datasets plays a vital role in understanding the formation of social-economic phenomena in smart cities [16] [17]. It is essential to consider the possible social features for mobility modeling, such as population density, vehicular connectivity, and traffic volume.

The gravity model is the typical traditional model originating from physics, which describes mobility fluxes. Despite its extensive use to predict mobility patterns at different spatial scales [18] [19], the gravity model relies on specific parameters fitted from systematic collections of traffic data. The formula is as follows:

$$T_{ij} = \alpha \frac{m_i m_j}{r_{ij}^{\beta}}, \tag{1}$$

where T_{ij} is the travels departed from location i to location j, m_i and m_j are the populations of origin and destination and r_{ij} is the distance between i and j.

Another kind of trip distribution model is the radiation model [20]. Nevertheless, some evidence demonstrates that the radiation model maybe not apply to predict human mobility at the city scale. Some studies suggest the diversity of human mobility at different spatial scales. Therefore, the models that succeed in predicting mobility patterns at large spatial scales, such as countries, are inappropriate at the city scale because of the underestimation of human mobility [21]. The population-weight opportunities (PWO) model without any adjustable parameters provides a new approach to predict social mobility patterns at the city scale. The PWO model enlarges the possible chosen area of individuals into the city regarding the relatively high mobility at the city scale.

Either the gravity model or the PWO model, the areas they apply to are small and regular shapes, such as regular areas of the same size divided by latitude and longitude or land area. Most of the regions are irregular. Thus, based on the PWO model, we propose the RPWO model to solve the problem of modeling the vehicle mobility behavior between uneven shape areas.

C. Trajectory Data Generation

For a long time, the study on vehicle trajectory generation has never been stopped. In 1985, Simon *et al.* [22] used on-off data to obtain the origin-destination matrix of bus routes, to provide better and more efficient services for people to travel. Friedrich *et al.* [23] used mobile phone data to generate origin-destination matrices for traffic state prediction or planning. Caiati *et al.* [24] used an open-source dataset to generate an origin-destination matrix to estimate and verify the daily travel demand of a city. Harri *et al.* [25] are the first to put forward the guidelines of the vehicle trajectory generation model and provide an overview and comparison of different mobility models proposed for VANETs. Pigné *et al.* [26] generate movement trajectories of vehicles in Luxembourg using microscopic simulation tools based on the traffic flows obtained from induction, which is only about the main roads in the city or highways out of the city rather than the whole city. Pappalardo *et al.* [27] propose the DITRAS framework to simulate the patterns of human mobility. The proposed framework first generates a mobility diary and then translates it into a mobility trajectory. They also consider the possibility of individuals breaking the routine. Recently, Kang *et al.* [28] propose a data-driven trajectory generation method that can capture the context and statistical mobility features.

To study the reliability of network simulation, Bedogni *et al.* in [29] developed reliable and publicly available mobile tracking by road information. The authors apply SUMO that allows for the import of OSM data in a clean and automated manner. They then generate a raw dataset of road traffic in Bologna, Italy. Uppoor *et al.* [30] incorporates report information and demographic information to synthesize vehicle trajectory for the city of Köln. The traffic demand in both studies were constructed as OD matrices, which include individual trips. However, their research is to generate a similar dataset, and we utilize taxis to generate the trajectory of private cars.

Ketabi *et al.* [31] design a scenario generation framework which can be adjusted for different ideas and models so that to know the real world human mobility. The framework includes data-driven components for performance of various calculations, such as traffic density, flow, road occupancy. Compared to previous work, based on the data of taxis, Kong *et al.* [32] propose a method to generate private car traffic trajectory data by using taxi trajectory data in 2018. The problem with this method is that the gravity model's important parameters are too dependent on different urban attributes, and the model is inaccurate.

Through the work of the dataset generation of others, we can find that the generation of private car trajectory data is basically in a new stage. The lack of original data makes it challenging to generate datasets. Meanwhile, there is no uniform standard for the verification of trajectory data. To overcome the challenges mentioned above, this paper proposes a three-layer dataset generation model with detailed implementation plans from data generation to data verification.

D. Simulation of Urban Mobility

Traffic simulation is an important method to solve complex traffic problems. Traffic simulation can analyze and predict the location and cause of traffic jams, compare and evaluate urban planning, traffic engineering, and traffic management related programs, and avoid before the problem becomes a reality.

Simulation of Urban Mobility (SUMO) is an open-source tool to simulate traffic conditions, used to simulate vehicular mobility in the city [33]. Viewed from the simulation content, SUMO is a space-continuous, discrete-time microscopic simulation package, including road network import and demand modeling components [32]. SUMO is so powerful that it can help to study urban traffic conditions more deeply. One of the most common applications in it is the OD2TRIPS, which converts O/D matrices to single-vehicle trips. As the traffic simulation sumo requires the representation of road networks and traffic demand to simulate in a specific format, both have to be imported or generated using different sources [33]. Instead of focusing on the traffic flows like macroscopic traffic simulators, SUMO pays attention to the behavior of a single vehicle in the traffic flow as typical of microscopic simulators. For the VoI, analysis of a single vehicles trajectories is vital. Therefore, we choose OD2TRIPS to get all trajectories of every single vehicle.

Compared with previous work, the method proposed in this paper is a three-layer private vehicle trajectory data generation model based on urban area division and vehicle movement patterns. Taking the points of interest and road network data into consideration, the area is divided reasonably and the Adjacent Road Segmentation (ARS) method is proposed. The proposed RPWO model is used to generate inter-regional vehicle movement behavior by combining traffic development annual reports and taxi datasets. Using the simulation tool SUMO, simulate the behavior of each vehicle and complete the task of trajectory generation. The generated trajectory data are verified from both macroscopic and microscopic aspects.

III. Framework

In this section, we describe the RMGen model framework in detail. The structure of this model is shown in Fig. 1.

The core of our RMGen dataset generation model is to generate and predict private car traffic volume between regions. We can use simulation tools, such as SUMO, to create vehicle trajectory data for a certain period based on traffic volume. To calculate the amount of traffic, we need to start from two aspects. On the one hand, the urban functional area division makes it easy for us to study the city's movement laws. This part of the work is described in Section IV. On the other hand, after the functional area is divided, we can predict the traffic volume between each functional area by using the existing vehicle data and geographic information by constructing the regional population-weight opportunities (RP-WO) model. Finally, using inter-regional traffic as a parameter, we can generate detailed trajectory data for the vehicle using the SUMO simulation tool expounded in Section V.

Specific to our RMGen model, the first layer of the RMGen model is the preparation layer. In this layer, we first do the

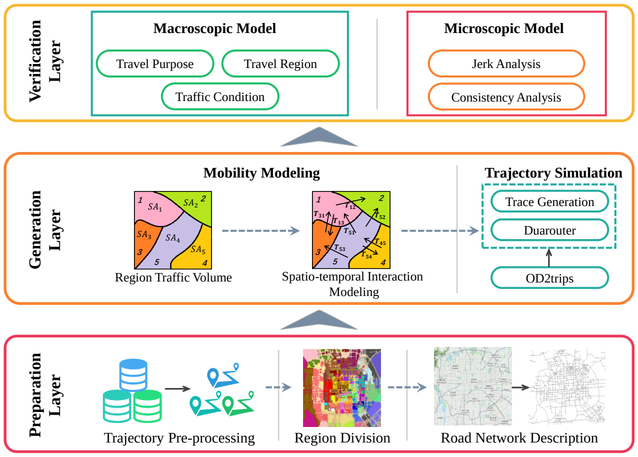


Fig. 1. The structure of RMGen model.

data pre-processing to get the available taxi trajectory data and remove the error data. We propose a new method of regional division named Adjacent Road Segmentation (ARS) method to divide Beijing into several areas to simplify the travels of vehicles to the trips between different regions. After that, we process the urban road network. Through the preparation layer, we lay the foundation for the following work.

The generation layer mainly involves the generation of private car trajectory datasets. Combined with the functional area obtained by the region division, we calculate the traffic volume between the regions. Then, based on the RPWO model, we get the traffic volume origin-destination (OD) matrix. After that, we use the SUMO tools to generate the travel start point file for the single-vehicle through the O/D matrix using the OD2TRIPS function. The Duarouter function is then used to generate the details of the travel path for each vehicle. Finally, using the Tutorials/Trace File Generation function, we get the travel trajectory data with an interval of one second.

At the verification level, we present a validation model based on macroscopic and microscopic levels to validate the authenticity and accuracy of the generated data set. From a macroscopic perspective, we refer to the Beijing Traffic Development Annual Report [1] published by the Beijing government in 2012, and compare the generated datasets with the actual traffic conditions in Beijing shown in the report to validate the accuracy of the generated data. At the microscopic level, we mainly consider whether the dataset itself is contrary to reality. On the one hand, we conduct acceleration and jerk analysis. On the other hand, we do the consistency analysis, randomly extract vehicle pairs to analyze the relative distance between two vehicles.

IV. Preparation Layer

A. Dataset Pre-processing

We obtain the taxi trajectory data from Beijing, China, in November 2012, which contains more than 10 billion GPS records by about 27,000 taxis. GPS updates the location information of taxi devices at a frequency of 11 seconds. The origin data files are stored in the text documents named after the storage time. Taxis have two mobility patterns, carrying passengers or not. Thus we delete the useless trajectory data when taxis have no passengers. Then we process the data to acquire every vehicle trip. We extract the same vehicle ID into one file and sort them by trip time. Thus we get the single taxis to travel trajectory.

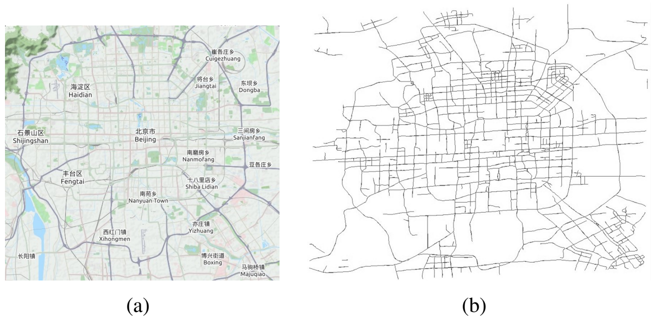

(a) (b)

Fig. 2. The map and road network of Beijing. (a) The fifth Ring Road of Beijing. (b) The processed Beijing road network.

It is worth mentioning that we focus our analysis on the vehicle mobility pattern within the fifth Ring Road in Beijing ($[116.1970°E, 116.5425°E]$ and $[39.7775°N, 40.0335°N]$), which is shown in Fig. 2(a). Therefore, we remove the useless vehicle trajectory data where the latitude and longitude are not within the five-ring road range.

B. Region Division

In this section, we start from the purpose of travel and describe people's travel behavior according to the categories of origin point and destination point. That is, travel behavior is regarded as a behavior based on a certain purpose (such as going home, going to school, going to work, etc.) from one functional area (school) to another functional area (residential area). Therefore, it is first necessary to determine the different functional areas in the city, and then perform functional characterization, and finally to divide the different functional areas of the city. The region division process is shown in Fig. 3.

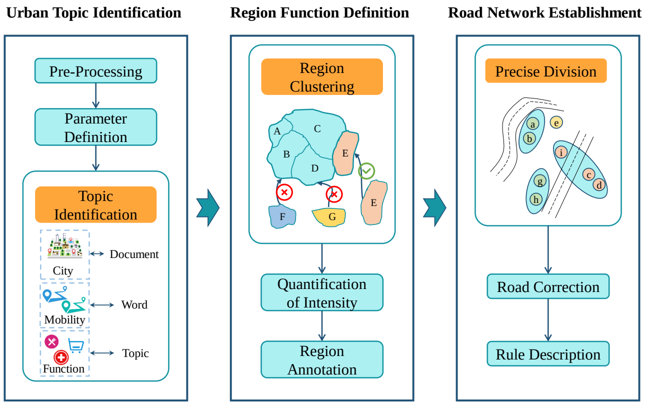


Fig. 3. The detailed process of region division includes topic identification, function definition and road network establishment.

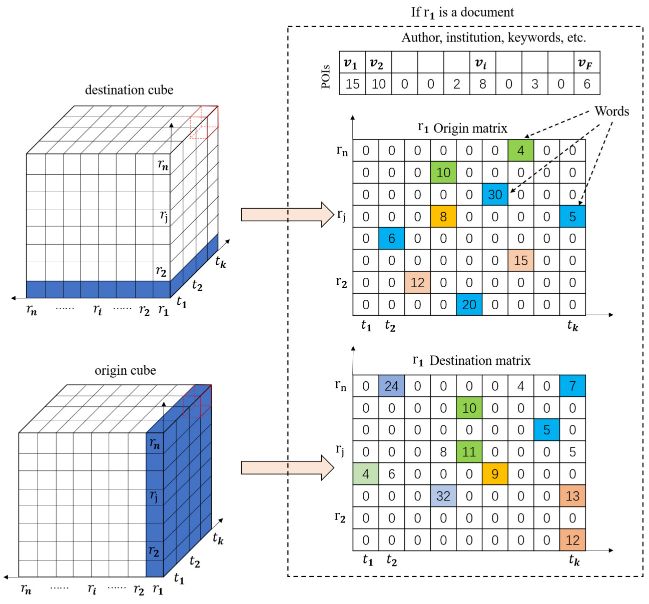
[1] Beijing Traffic Development Annual Report, http://tjj.beijing.gov.cn

TABLE I
ANALOGY FROM REGION-FUNCTIONS TO DOCUMENT-TOPICS

Region function contains elements	Corresponding elements of the document theme
Migration cube	Vocabulary
Collection of research areas	Document collection
Region	Documentation
The function of the area	The subject of the document
Move sample	Word
POI feature vector	Document metadata

1) Urban Toptic Identification: In 2013, Blei *et al.* [34] proposed a topic model for studying text processing, the Latent Dirichlet Allocation (LDA) model, which can determine the probability of multiple topics in each article in a corpus. This has many similarities with the recognition of urban functions. See TABLE I for the analogy.

Regarding the researched city as a collection of documents, each area in the city can be regarded as the documents in the collection, and the different functions of each area can be regarded as the different themes of the documents. At the same time, the movement trajectory data is constructed as a migration cube, including the departure cube and the arrival cube, which are analogous to the vocabulary in the document. On this basis, the movement samples between regions (spatial-temporal trajectory) can be regarded as words in the document, and the POI feature vector can be regarded as the metadata of the document. In this paper, the Dirichlet Multinomial Regression (DMR) model [35] based on the improvement of LDA is used. Its advantage is that it can combine the POI features with the mobile mode and the experimental results are more in line with the real situation.



v_1	v_2				v_i				v_F
15	10	0	0	2	8	0	3	0	6

r_1 Origin matrix

r_n	0	0	0	0	0	0	4	0	0
	0	0	0	10	0	0	0	0	0
	0	0	0	0	0	30	0	0	0
r_j	0	0	0	8	0	0	0	0	5
	0	6	0	0	0	0	0	0	0
	0	0	0	0	0	0	15	0	0
r_2	0	0	12	0	0	0	0	0	0
	0	0	0	0	20	0	0	0	0
	t_1	t_2							t_k

r_1 Destination matrix

r_n	0	24	0	0	0	0	4	0	7
	0	0	0	0	10	0	0	0	0
	0	0	0	0	0	0	0	5	0
r_j	0	0	0	8	11	0	0	0	5
	4	6	0	0	0	9	0	0	0
	0	0	0	32	0	0	0	0	13
r_2	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	12
	t_1	t_2							t_k

Fig. 4. Analogy between mobility patterns and words based on transition cuboids.

For the migration cube, this paper uses Fig. 4 to compare it with a vocabulary in a document. As shown on the left side of Fig. 4, a migration cube is defined as a cube with a size of $R \times R \times T$. Among them, R is the total number of areas involved in the study, and T is the number of fixed time intervals. Since actual travel is continuous in time and space, to facilitate research, a day is divided into a fixed number of time intervals at the time level; at the space level, the start and end points of the travel are divided according to regions.

Thereby, the spatio-temporal data of travel can be discretized. In the research, the movement pattern of an object is defined as a group of elements containing the start and end location and time. There are two types of movement patterns, namely the departure movement pattern $M_L = (r_O, r_D, t_L)$ and the arrival movement pattern $M_A = (r_O, r_D, t_A)$, where r_O is the departure location, r_D is the arrival location, t_L is the departure time, and t_A is the arrival time. Since there are two types of movement modes, there are also two types of corresponding migration cubes, namely the departure cube C_L and the arrival cube C_A. An element (i, j, k) in the departure cube C_L represents the number of trips from the area r_i to the area r_j in the time period t_k. The formula is as follows:

$$C_L(i,j,k) = || \{M_L = (x,y,z) \mid x = r_i, y = r_j, z = t_k\} || \tag{2}$$

Similarly, the reaching cube C_A can be described as:

$$C_A(i,j,k) = || \{M_A = (x,y,z) \mid x = r_i, y = r_j, z = t_k\} || \tag{3}$$

As shown on the right side of Fig. 4, the area r_i is regarded as a document, and each cell in the matrix represents a specific movement pattern, and the number in the cell represents the number of occurrences of the pattern. Just as the metadata information of a document includes the author, address, keywords, etc., POI is used in the area to represent its metadata information. The POI is recorded by a tuple (in the POI database), and the tuple consists of the POI category and name. And geographic location (latitude, longitude). For each region r, the number of different types of POI in the region can be obtained by statistics. The formula for calculating the frequency density $\mathrm{v}_{i,r}$ of the i-th POI in the region r is:

$$v_{i,r} = \frac{Num_i}{S_r} \tag{4}$$

where Num_i represents the number of the i-th type of POI in the area r, and S_r represents the area of the area r. In addition, for region r, its POI feature vector can be written as $x_r = (v_{1,r}, v_{2,r}, \ldots, v_{F,r}, 1)$, represents the metadata of the region r, F is the number of POI types in r, and the last vector 1 is the default feature.

In this section, using the DMR topic model in unsupervised learning, by combining the feature vector of POIs and moving samples, the function of the region is comprehensively explored from two aspects. By applying the DMR model, given the movement pattern and POI characteristics, the topic assignment of each area and the movement pattern distribution of each topic are obtained.

After the parameter estimation using the DMR based topic model, for the region r, the topic distribution is a K-dimensional vector $\theta_r = (\theta_{r,1}, \theta_{r,2}, ..., \theta_{r,K})$, where $\theta_{r,K}$ represents the proportion of topic K in the region r.

2) Urban Function Recognition: Due to LDA being an unsupervised learning model, it is not used for classification itself, and it needs to be embedded with a suitable clustering algorithm to identify city functions. Therefore, we chose the classic clustering algorithm K-Means [36] with fast running speed, simple calculation and low complexity for optimization, and proposed a clustering algorithm based on connected components. Algorithm 1 presents the pseudocode of the proposed cluster method.

First, we regard all topic distribution vectors in the dataset as nodes, define the reciprocal similarity as the distance between nodes, and s is the similarity threshold. Calculate the distance between the node pairs u, v of the topic distribution vector, if the distance is less than s, an edge is generated between u, v (Lines 1-5). We record the processed connected components as C and store them in the definition set T, that is, $T = \{C_1, C_2, , C_n\}$ (Lines 6). Then, for each connected component in the set T, calculate the distance between all its node pairs (Lines 7-9), and find the two nodes m, n with the farthest distance. If the distance between nodes m and n is greater than $2s$, the connected component is split with m, n as the cluster center and K-means algorithm is used to generate new clusters C' and C" and add it to the set T (Lines 10-12). Otherwise the loop ends(Lines 13). For the research area within the fifth Ring Road of Beijing, we finally cluster nine different functional regions.

Algorithm 1: Pseudocode of connected component based urban region clustering algorithm

```
   Input: the topic distribution vector $\theta_r$ and similarity
          threshold $s$
   Output: clusters of the topic distribution vector
1  assume all vectors as nodes:
   for all node pairs $u, v$ do
2      if $Distance < s$ then
3          $Connect(u, v)$
4      end
5  end
6  get the set of connected component $T$,
   $T = \{C_1, C_2, , C_n\}$
   for all $C$ in $T$ do
7      for all node pairs $u, v$ do
8          calculate $Distance(u, v)$
9      end
10     find two nodes $m, n$ in $C$ with the farthest distance
       if $Distance(m, n) > 2s$ then
11         execute $K - means(m, n)$ and generate $C', C''$
           delete $C$ from $T$
           add $C', C''$ to $T$
12     end
13 end
```

To quantify the functional area's popularity and range, we estimate the functionality intensity of each functional region, which is reflected by the human mobility pattern. We quantify the functionality intensity in the functional areas by the Kernel Density Estimation (KDE) model [37]. We assume that there are n regions $(x_1, x_2, ..., x_n)$, and use the KDE model to calculate the functionality intensity of the region s by the kernel density estimator:

$$\begin{aligned}\lambda(s) &= \sum_{i=1}^{n} \frac{1}{nr^2} K(\frac{d_{i,s}}{r}) \\ &= \sum_{i=1}^{n} \frac{1}{nr^2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{d_{i,s}^2}{2r^2}),\end{aligned} \tag{5}$$

where $d_{i,s}$ represents the distance of the region x_i to s, r represents the bandwidth, $K(\cdot)$ represents the kernel function, and the value decreases when $d_{i,s}$ increases. In our study, we use Gaussian function as the kernel function and formulate the value of r based on the Mean Integrated Square Error (MISE) criterion.

After estimating the functionality intensity, we mark the divided areas to reflect the city's actual function. We consider regional annotations in four ways. Firstly, the frequency of POIs in the functional area is sorted according to the average frequency density of the POIs feature vector of each region. The frequency sizes of all functional regions, including POIs are sorted. Secondly, we calculated the most frequent mobility patterns in each functional area. Thirdly, we use functionality intensity to explore the most representative POIs in each functional kernel and then make regional annotations. Fourthly, we carry out manual marking according to actual conditions, such as government agencies.

Finally, we preliminarily divide Beijing within 5th Ring Road into nine function areas and number them: Diplomatic/Embassy Area, Emerging commercial/ Entertainment Area, Science/Education/Technology Area, Nature Area, Historical Interests/Parks, Developed Commercial/ Entertainment Area, Developed Residential Area, Old Neighborhoods Area and Emerging Residential Area.

3) Urban Regional Division: In the previous section, the marker area we obtain can represent the city's functional area. By studying the amount of travel between functional areas, it is possible to characterize the city's travel patterns and lay the foundation for generating the travel trajectory of microscopic vehicles. In the previous work, we roughly divide Beijing into nine functional areas. However, we still need to make a more detailed division of the functional area.

In order to facilitate our next step, we propose a new method of regional division named the Adjacent Road Segmentation (ARS) method. We rasterize the functional areas and split it into a number of grid cells on the map projection according to the range of 0.001 latitudes and longitude. The relative length and width of each grid is defined as one and has a fixed ID, shown in Fig. 5(a). If a grid contains multiple different regions, the mesh is placed in the functional area with the most significant area.

For Beijing, we propose a method for dividing the functional area accurately based on important urban roads. According to the road traffic information from the Beijing Traffic Development Annual Report of 2012, we select 18 urban roads with an average daily traffic flow of more than 100,000 as of the major roads, which is displayed in TABLE II and numbered by Arabic numerals 1-18. We mark each important road in the

TABLE II
THE PROPORTION OF TAXIS AND PRIVATE CARS ON MAIN STREETS.

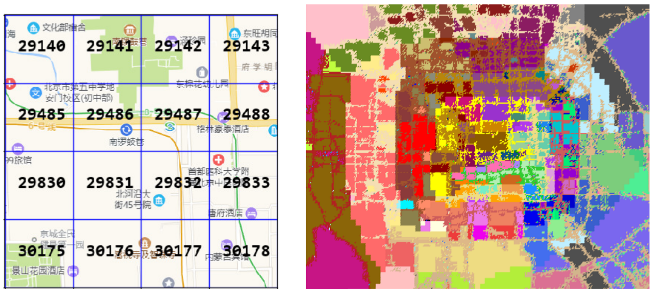
Street Name	Private Car Ratio	Taxis Ratio
East fifth ring road	59.52%	4.78%
South fifth ring road	37.24%	0.58%
West fifth ring road	68.50%	3.26%
North fifth ring road	59.97%	4.52%
East forth ring road	65.74%	12.44%
South forth ring road	71.41%	7.69%
West forth ring road	72.44%	10.84%
North forth ring road	60.50%	18.45%
East third ring road	55.38%	20.78%
South third ring road	57.88%	15.72%
West third ring road	62.19%	16.62%
North third ring road	59.03%	20.10%
East Second ring road	65.92%	19.60%
South Second ring road	52.53%	10.88%
West Second ring road	63.68%	20.02%
North Second ring road	69.63%	18.99%
Changan Avenue	65.53%	17.15%
Liangguang Avenue	45.50%	14.86%

rasterized Beijing map and treat it as a line segment. At the same time, we regard each grid as a node approximatively. Subsequently, we calculate the Euclidean distance from each node (grid) to the line segment (road) and record the nearest road ID for each grid.

After the distance calculation is completed, each grid node has two attribute values: the function area ID K_a and the nearest road ID K_r. We cluster grids in the rasterized map and treat nodes with the same K_a and k_r values as a community. Finally, we get 153 subdivisions of functional areas. The visualized display of the divided regions on the map is shown in Fig. 5(b).

C. Network Description

The urban map data can be downloaded free from the OpenStreetMap (OSM) or other open-source websites. OSM data can be uploaded by any user, so most of us can maintain and modify map data, which has both advantages and disadvantages. In our study, we download the OSM file of Beijing, including the information of roads, undergrounds, various construction facilities, which reflects the geographical information of the city.



(a)

(b)

Fig. 5. The urban regional division by ARS method. (a) Division by longitude and latitude. (b) The functional regions of Beijing.

However, because of the open-source nature, there may be some errors between the downloaded data and the actual situation. To build an accurate simulated road network, we correct the road topology. We modify the error roads using Java OpenStreetMap (JOSM) technology, which is a free editing tool for open street map geographic information. Moreover, we focus on stimulating private cars travelling in Beijing, so we deleted railways, sidewalks, and so on. The network after processing is shown in Fig. 2(b). Finally, we map the functional area generated in the previous section to the road network and mark which area each road belongs to.

V. Generation Layer

In this section, we introduce the methods to get the traffic volume of every functional region and propose a new model to calculate the traffic volumes between two traffic regions and predict the O/D matrix of the traffic volumes of private cars based on the RPOW model. After that, we use SUMO tools and Python programs to convert the matrix into single paths of private cars.

A. Demand Description

1) Region Traffic Volume: Vehicle traffic volume refers to the number of vehicles passing through a specific road section within a specified period. Traffic volume can reflect the overall traffic flow of a road and has significant research value. In our evaluation, we mainly study the traffic volume of each region which plays an important part in our prediction of private cars traces. In Section III-B, we divide Beijing into different regions. Now we calculate the amount of traffic in each area for one day. The regional traffic is the cumulative traffic of all road segments. The amount of traffic on the road is determined by both private cars and taxis, but the data we obtain is only the trajectory data of taxis. Therefore, we need to calculate the traffic volume of private cars through the traffic volume of taxis. The ratio of private cars to taxis on each road is different, and if we get this ratio, we can calculate the traffic volume of private cars easily. From the information provided by the Beijing Traffic Development Annual Report, we know the proportions of the number of taxis and private cars on the main traffic roads. Because the areas we divide are based on these major traffic roads, we can assume that all roads in the same area have the same taxi and private car ratio. We calculate the traffic volume of each functional area by the following formulas:

$$SA_i = \alpha_i \sum_{j=1}^{N_i} SG_j, \tag{6}$$

$$SG_j = \sum_{j=1}^{n_j} SR_k, \tag{7}$$

where SA_i means the total number of private cars in the functional region i, which is divided into N_i girds. And α_i represents the corresponding ratio of private cars and taxis in the region i. SG_j represents the total number of taxis in grid j which contains n roads. SR_k means the number of taxis on road k. And through formulas (6) and (7), we obtain the traffic volume of private cars in each functional region.

2) Spatio-temporal Interaction Modeling: After obtaining the traffic volume in various areas of Beijing city, we continue to study the human mobility patterns under the city scale. Despite the long history of building human mobility models, researchers still lack highly accurate methods to predict urban mobility patterns, especially if the types of data are not diverse.

In the related work, we introduce the Gravity model. Gravity models are widely used in the study of travel distribution and are recognized as models for predicting urban travel patterns. Although this model has a very similar form to Newtons gravity law, it is unconstrained, which will lead to problems in predicting traffic volume. To ensure the constraints, we employ the origin-constrained gravity model to predict mobility patterns in cities, described as:

$$T_{ij} = SA_i \frac{m_j f(i,j)}{\sum_{k \neq i}^{R} m_k f(i,k)}, \tag{8}$$

where the distance function $f(i,j)$ can be of any forms, and in our study, we define $f(i,j)$ as the relative attraction of destination j to travelers at origin i.

The PWO model is proposed by Yan *et al.* [21] to capture the potential drivers of human mobility patterns at the city scale, which does not depend on any adjustable parameters.

In [21], the authors abstract the travel origin and destination into two nodes, and simply partition all cities into $1 \times 1km^2$ square zones. Different from their study, we consider the regional factors and study the patterns of human mobility between regions instead of nodes, which we obtain through the ARS method. We named this new method as RPWO model.

The model is derived from a stochastic decision-making process of an individuals destination selection. People will weigh the benefit of each locations opportunities before choosing the travel destination. The more opportunities a location has, the higher the interest it offers, and the higher the chance of it being chosen[38]. Its population can reflect the number of a locations opportunities. As the population distribution is available, it is reasonable to assume that the number of opportunities at a location is proportional to its population.

In our RPWO model, we simply assume that the attraction of a destination is inversely proportional to the population Q_{ji} in the circle centred at the destination with radius R_{ij} (the distance between the centre of gravity of original region i and destination region j), and to make the result more in line with the real situation, minus a finite-size correction $1/M$. For a region, the centre of gravity $G(\overline{x}, \overline{y})$, the calculating methods are as follows.

$$\overline{x} = \frac{1}{L} \sum_{l=1}^{L} x_l, \overline{y} = \frac{1}{L} \sum_{l=1}^{L} y_l, \tag{9}$$

where L means the number of squares (0.001 latitude and 0.001 longitude) in the region. x_l and y_l represent the relative longitude and relative latitude of the squares node in the region respectively.

TABLE III
THE FORMAT OF THE GENERATED TRACE DATASET.

Attribute	Notes	Example
ID	The vehicle ID	3901
Depart	The time when the trip beginning	59900.64
FromTaz	The orgin region	103
ToTaz	The destination region	95
Route edges	The ID list of the road through which the trip passes in sequence	”201526561#2,238470275,...,247814291#0”

Then we can calculate the the relative attraction of destination to origin:

$$Q_{ji} = \sum_{r=1}^{N} \beta_r P_r = \sum_{r=1}^{N} \frac{\widetilde{S}_r}{S_r} P_r, \tag{10}$$

$$f(i,j) = o_j \left(\frac{1}{Q_{ji}} - \frac{1}{M} \right), \tag{11}$$

where Q_{ji} means the population in the circle centred at the destination with radius R_{ij}. N is the total number of the region contained in the circle. β_r means the ratio of the region r contained in the circle, which is obtained by dividing the area contained within the circle ($\widetilde{S}_r$) by the total area (S_r). P_r is the population of region r. $f(i,j)$ represents the relative attraction of destination j to travelers at origin i. oj is the total opportunities of destination j, and M is the total number of people in the city.

We assume that the probability of mobility from region i to j is directly proportional to the attraction of j. And referring to [38], the number of opportunities o_j is directly proportional to the population m_j. Putting formula (11) to (8), we can calculate the travel volume from region i to region j:

$$T_{ij} = SA_i \frac{m_j (1/Q_{ji} - 1/M)}{\sum_{k \neq i}^{R} m_k (1/Q_{ki} - 1/M)}, \tag{12}$$

where SA_i is the number of trips departing from i, which can be obtained by (6) and (7). R means the total number of regions in the city.

B. Trajectory Simulation

We calculate the traffic volume of all regions by the RPWO model and put them into a matrix named traffic volume O/D matrix, where the value in row i and column j means the traffic volume from i to j. After obtaining the urban traffic volume O/D matrix, we can perform dataset simulation generation work combined with the modified road network files. To achieve this goal, we use SUMO tools.

Firstly, we classify the roads of the city according to the areas that have been divided. The road network files contain the latitude and longitude of the connection points of every road. We use the latitude and longitude to calculate which area a road belongs to, and write the road ID contained in each area in the road network files.

Then, we use the OD2TRIPS plugin in the SUMO tools, import the O/D matrix, and split it into individual vehicle itineraries. According to the city's specific conditions and data, we input the O/D matrix, the road network file, and the road list included in the area, set the generation period, the travel ratio of each period, and generate vehicle type parameters. The

TABLE IV
THE FORMAT OF THE GENERATED TRACE DATASET.

Attribute Name	Notes	Example
time	Current time(s)	25200
id	The vehicle id	1604674
x	Longitude of the vehicle	116.2821
y	Latitude of the vehicle	39.82672
angle	The angle at which vehicle is traveling	300.5352
speed	Current vehicle speed(m/s)	13.30696
lane	The road ID where vehicle is currently located	139824665#1_0

XML file of series vehicle travel information can be generated, where each trip information includes a vehicle ID, a departure time, a departure place ID, and a destination ID.

With the help of OD2TRIPS, the origin and destination information for each vehicle trip is generated. However, it is not a complete trip representation. Thus, we use the D-UAROUTER plugin in the SUMO tool to make vehicular trajectories using shortest path computation. The generated trajectories consist of vehicular information such as road segment, speed, and travel time. We input the road network files and the trip information generated by OD2TRIPS, set the simulation period, shortest path calculation method, and finally create the vehicle trip trajectory information, including the vehicle ID, travel time, and the travel route. Vehicle trajectory information can help us to analyze urban road traffic, regional travel modes, and so on. The format of the trajectory data we get is shown in TABLE III, and the size of this trajectory generated dataset of one day is about 3 GB.

In addition to vehicle trajectory information, microscopic information is also important for our next study. We use the Trace File Generation plugin in the SUMO tools to generate information including the relative position, latitude and longitude of the vehicle, driving angle, road ID and instantaneous speed of the vehicle per second for a specified time interval. We enter the same information as required in the DUAROUTER function, write the corresponding configuration file at the same time, set the time interval to 15 minutes, and generate vehicular trace files for different time periods. The size of the generated dataset of one day is about 200 GB, and we can see the format of the generated trace datasets from TABLE IV.

C. Complexity Analysis

The time complexity analysis of our proposed framework considers the main four components. For LDA, its time complexity is $O(rK)$ (r is the number of regions and K is the number of topics). For Algorithm 1, its time complexity is $O(\frac{r^2+r+m^2\times n}{2})$ (r is the number of regions, m is the total number of nodes in each connected component, n is the total number of connected components). The time complexity for ARS is $O(NR)$ (N is the number of grids, R is the number of major urbanroads). For the RPWO model, since every parameter in the model can be obtained directly, the parameters can be ignored. To sum up, the overall time complexity of our model is $O(rK+\frac{r^2+r+m^2\times n}{2}+NR)$.

VI. Verification Layer

After generating the private car trajectory data, we design the verification model to validate the data's accuracy and authenticity. In our verification layer, the data can be validated in the macroscopic and microscopic views, respectively.

A. Macroscopic Model

In the macroscopic model, we do the analysis and contrast test with the real traffic condition described in the Beijing Traffic Development Annual Report. In order to show the performance of our model, we also take the existing gravity model and PWO model as the core of the dataset generation and conduct comparative experiments.

1) Traffic Flow: According to the report, from the overall situation of road traffic in Beijing, the expressway and the main road are the main channels for carrying out traffic operations. Therefore, we analyze the actual traffic flow data and the generated traffic flow data of the main roads in Beijing. Fig. 6 is a comparison of the traffic flow of major roads in Beijing. The real data is the main road traffic flow data obtained by the Beijing Traffic Development Research Center in 2012. From the results, both the actual data and the generated data show that the all-day flows of the West Fourth Ring Road and the East Fourth Ring Road are at the forefront. The flow rates of the West Fifth Ring Road and the South Second Ring Road are lower, indicating that the traffic burdens are lighter. From the overall comparison results, except for the South Fifth Ring Road and the East Fifth Ring Road, the data generated by us is more consistent with the real data. By contrast, the gravity model method is very inaccurate in describing the condition of South Second Ring Road, East Second Ring Road, and North Second Ring Road. In contrast, the PWO model's method is weak in the South Fifth Ring Road, East Fifth Ring Road, East Fourth Ring Road, and South Second Ring Road.

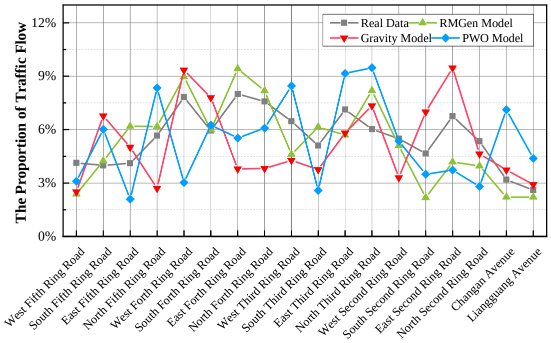
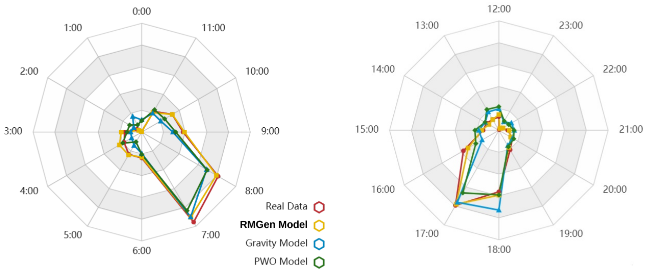


Fig. 6. Traffic Flow of Major Roads in Beijing.

2) Travel Range: For human mobility analysis, travel time distribution and distance distribution are two crucial parameters. By studying the amount of travel in different periods, researchers can propose better travel optimization programs to alleviate road conditions and improve travel efficiency. Besides, studying the distribution of human travel distance also plays an important role in road planning and travel prediction. Therefore, we use the generated trajectory data to analyze the distributions of travel time and distance. The accuracy of the simulation data is evaluated by comparison with real data.



(a) 0:00-12:00 (b) 12:00-24:00

Fig. 7. Travel time distribution.

Fig. 7 is a distribution of travel volume of residents' travel time. The data that participated in the comparison includes the official statistics of the data generated by several methods. The result proves that the trajectory data we generated has the same travel characteristics as the real data. In addition, as can be seen from Fig. 7, 7:00-9:00 and 17:00-19:00 are two peak travel periods. The amount of travel in these two time periods accounts for about 50% of the total daily travel volume. The data generated by our RMGen model is more consistent with the real situation.

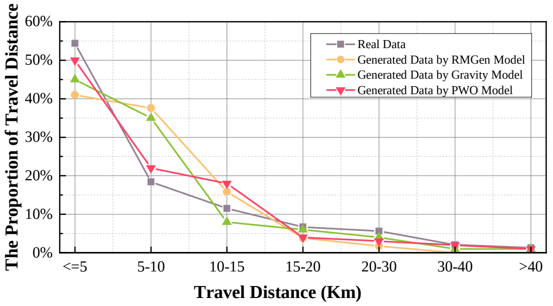
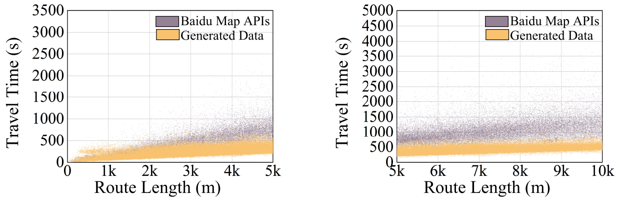


Fig. 8. Travel distance distribution.

In terms of the distribution of residents' travel distance, Fig. 8 is the result of our analysis. We compare the generated data with the official travel distance distribution data. All these generated data are consistent with the real situation. From the distribution of travel distance, the number of trips is inversely proportional to the overall distance. As the distance increase, the number of trips decreases. For driving, humans prefer to have short distance trips of 0-5 kilometers, which accounts for more than 40%. Considering the specific circumstances, when the travel distance is too long, people will choose the train, subway and other modes of travel instead of driving, due to the consideration of fuel consumption, time spent, and so on.

3) Traffic Condition: From the macroscopic traffic flow situation, the overall travel distances and travel times of the vehicle are regular. Referring to [29], we use the navigation service of Baidu Map APIs to validate our generated datasets. Today, with the growing use of mobile devices, most of us have experience using map service applications. When using these software, we input the origin and destination positions and select the appropriate travel mode, such as walking, bus, and private car, to get the estimated route lengths and travel times.



(a) 0 - 5 km (b) 5 - 10 km

Fig. 9. Traffic condition comparison of travel time and distance.(a) Route lengths are less than 5km. (b) Route lengths are between 5km and 10km.

In the following study, we use the navigation services of Baidu Map APIs to compare with our generated datasets. From Fig. 8, we can see that more than 70% of the vehicles travel within distances of less than 10 km. Therefore, we put the focus of the study on the travel of vehicles within 10 km. Fig. 9 shows a scatter plot of the travel times and route lengths estimated from the Baidu Maps APIs and generated by us. Fig. 9(a) represents the travel situation of vehicles within 5 km, and Fig. 9(b) represents the travel of vehicles ranging from 5 to 10 km. We note that in both ranges of route lengths, the generated travel times are overlapped with the values provided by the navigation service. This also reflects that the generated vehicle speed distribution is more in line with the speed distribution of real vehicles. In addition, compared with the distance ranging from 5 to 10 km, the generated data of travel time is more in line with the real situation during the short distances travel within 5 km, which proves that our proposed model is more suitable for generating travel trajectories with shorter distances.

Fig.10 shows the visual comparison of the vehicle trajectory geographic information between the generated data and the real situation. Fig.10(a) is the real morning rush hour (7:00-8:00) vehicle trajectory geographic information map, from the government's traffic report, the red part represents the road with a high traffic density. For comparison, we use the generated trajectory data to extract vehicle position information from 7:00 to 8:00, depicting the distribution of vehicle trajectories during the period, as shown in Fig.10(b). The color from green to red represents the traffic density from small to large. Comparing Fig.10(b) with Fig.10(a), we can see that at the rush hour, our generated data is similar to the geographical distribution of vehicle trajectories of real data. In addition, we add vehicle trajectory analysis during off-rush hours (14:00-15:00), as shown in Fig.10(c). During the off-rush hours, the number of vehicle travel trajectories is significantly reduced, and the city's overall road traffic density is reduced considerably.

B. Microscopic Model

In this section, we adopt the method of analyzing and evaluating the accuracy of the generated data from the perspective of acceleration and relative distance, which is presented by Punzo *et al.* [39].

TABLE V
JERK ERROR STATISTICS OF REAL AND GENERATED TRAJECTORIES.

Indicator	Original (7:00-7:15)	Original (7:15-7:30)	Original (14:00-14:15)	Generation (7:00-7:15)	Generation (7:15-7:30)	Generation (14:00-14:15)
%Of jerk values > abs (3(m/s^3))	3.276272	2.994038	5.472061	5.270598	3.719935	4.974231
Maximum jerk(m/s^3)	7.275132	25.000003	8.173139	11.591182	11.598726	10.991288
Minimum jerk(m/s^3)	-10.648148	-27.500009	-20.500009	-11.590758	-11.590758	-13.127592

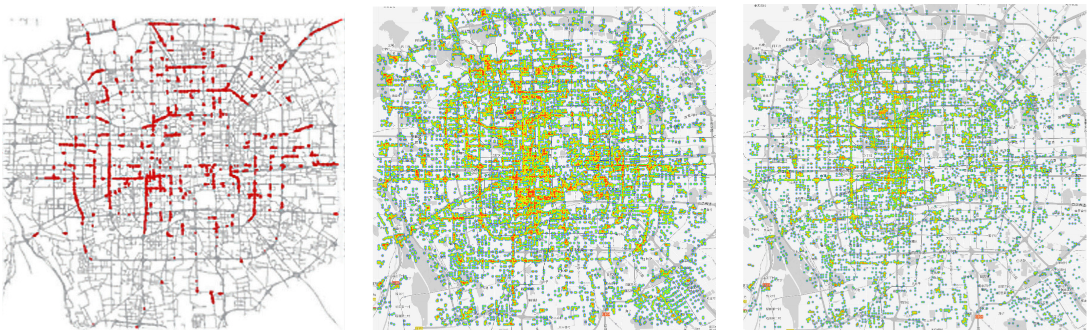

(a) Morning rush hour in reality (b) Morning rush hour of generated data (c) Off-rush hour in generated data

Fig. 10. The visual comparison of the vehicle trajectory geographic information.

In this part, we choose the morning rush hour (especially from 7:00 to 7:30) and off-rush hours (14:00-14:15) of the workday as an analysis period. In addition, we use six sets of observations as a contrast, with the time interval of 15 minutes for each data collection. This data is the traffic trajectories of 7:00-7:15, 7:15-7:30, and 14:00-14:15 on weekdays extracted from the original downloaded dataset and the generated dataset.

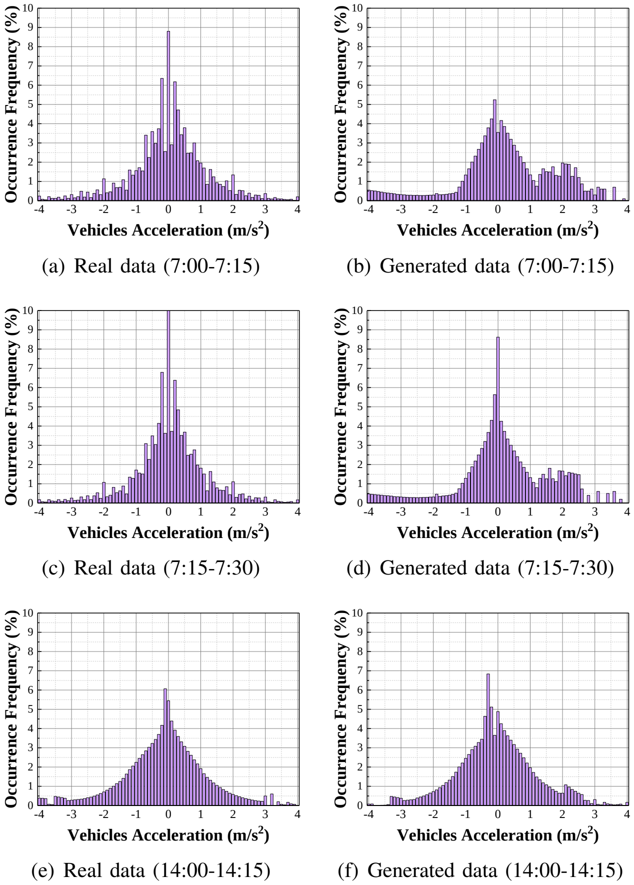
1) Jerk Analysis: Vehicle acceleration is an important part of vehicle dynamics and traffic flow evaluation. So it is necessary to validate the accuracy of the acceleration of our generated trajectory data. There is an obvious way to validate acceleration data, which is to check its distribution across the entire dataset. Fig. 11 shows the acceleration frequency of the originally downloaded datasets and generated datasets. From Fig. 11, we can see that whether real data or generated data, the frequency distributions of vehicle acceleration are usually normally distributed.

Apart from the distribution of the acceleration values, the jerking factor also gives an important indication of data quality. The jerking factor $j[m/s^3]$, means the variations of acceleration in time, which is the derivative of the acceleration. In our study, we consider the jerking values in reality around $\pm 3m/s^3$ as an acceptable value for applications of traffic flow microscopic simulation [40].

Therefore, we propose these indicators to do the jerk analysis:

- The percentage of trajectory data with j higher than the threshold of $\pm 3m/s^3$;
- Maximum and minimum j in the generated datasets.

The jerk error statistics results of the analysis are reported in TABLE V. As we can see, the percentages of jerk values higher than $\pm 3m/s^3$ are 3.28% (Original (7:00-7:15)), 2.99% (Original (7:15-7:30)) and 5.47% (Original (14:00-14:15)), while the generated data has similar result (5.27%, 3.72%



(a) Real data (7:00-7:15) (b) Generated data (7:00-7:15)

(c) Real data (7:15-7:30) (d) Generated data (7:15-7:30)

(e) Real data (14:00-14:15) (f) Generated data (14:00-14:15)

Fig. 11. Acceleration frequency of real datasets and generated datasets.

TABLE VI
PLATOON CONSISTENCY (SPACING INDICATORS) ERROR STATISTICS OF REAL AND GENERATED TRAJECTORIES.

Indicator	Original (7:00-7:15)	Original (7:15-7:30)	Original (14:00-14:15)	Generation (7:00-7:15)	Generation (7:15-7:30)	Generation (14:00-14:15)
Total no. of vehicle pairs	4248	4416	3283	4931	4021	3671
No.vehicle pairs with spacing $< 5m$	24	18	16	94	50	33
%Vehicle pairs with spacing $< 5m$	0.56	0.41	0.55	1.91	1.24	0.90
Mean observation period of vehicle pairs(s)	19.62	24	20.33	27.38	27.69	28.80
The longest observation period for a vehicle pair(s)	49.87	49.85	49.77	50.11	49.98	50.02

and 4.97%). Horizontal comparison, the general data set error is less than 10%, indicating that the data set is relatively reasonable. What is more, maximum and minimum jerk values reach relative irrationality values in all the chosen data.

2) Consistency Analysis: During the driving process, vehicles must keep reasonable distances with other vehicles; otherwise, traffic accidents are likely to occur. When we consider this point, we can validate the rationality of the generated dataset from a distance between the vehicles.

When we focus on two following vehicles, previous consideration shows that the spacing between vehicle pairs can be used to quantify the error of the estimated trajectory [40]. In fact, the vehicle spacing in a vehicle pair at a certain moment can be measured directly from the position of two vehicles at this moment.

To simplify the calculation processing, we assume that the vehicles always travel along straight lines on the road. Then the distance between the vehicles can be calculated directly from the projected coordinates corresponding to the geographical coordinates of the vehicles:

$$\Delta s_{np}^{obs}(k) = \sqrt{(\overline{x_{n,k}^{obs}} - \overline{x_{p,k}^{obs}})^2 + (\overline{y_{n,k}^{obs}} - \overline{y_{p,k}^{obs}})^2}, \qquad (13)$$

where $\Delta s_{np}^{obs}(k)$ represents the vehicle spacing between vehicles n and p at time k. The points $\overline{P_{n,k}^{obs}} = (\overline{x_{n,k}^{obs}}, \overline{y_{n,k}^{obs}})$ and $\overline{P_{p,k}^{obs}} = (\overline{x_{p,k}^{obs}}, \overline{y_{p,k}^{obs}})$ are the actual vehicle positions of n and p.

In general, when $\Delta s_{np}^{obs}(k)$, at least for one instant, decreases below a threshold of $5m$, there will be a collision between the two cars, which leads to a traffic accident. If there are a number of abnormal values in the datasets, the dataset is most likely problematic.

For the consistency analysis, the following statistics are meaningful:

- The total number of vehicle pairs. This is the total number of vehicle pairs in each chosen datasets;
- The mean and longest observation period of vehicle pairs in the datasets;
- The number and the ratio of vehicle pairs with the spacing less than $5m$.

The consistency statistics results of the analysis are reported in TABLE VI. According to the conclusion, we can learn that the total number of vehicle pairs approximately ranges from 3000 to 5000 vehicles. In all selected datasets, the origin datasets we download have 24 and 18 anomalous vehicle pairs between 7:00-7:15 and 7:15-7:30 in rush hour, accounting for 0.56% and 0.41% of the total vehicle pairs. While in the off-rush hour (14:00-14:15), there are 16 unusual vehicle pairs. As for the datasets we generated, there are 94, 50, and 33 abnormal vehicle pairs in the 7:00-7:15, 7:15-7:30 and 14:00-14:15 period, accounting for 1.91%, 1.24% and 0.90% of the total vehicle pairs. It can be seen that the abnormal data in the generated datasets are maintained in a small range, which proves that the generated data is more reasonable. Besides, in the generated datasets, the average vehicle spacing is about $27m$. What is more, Fig. 12 reflects the occurrence frequency of the spacing length between vehicle pairs. These six sets of observation data have very similar distributions, which means that the generated datasets have similar vehicle mobility patterns with real data.

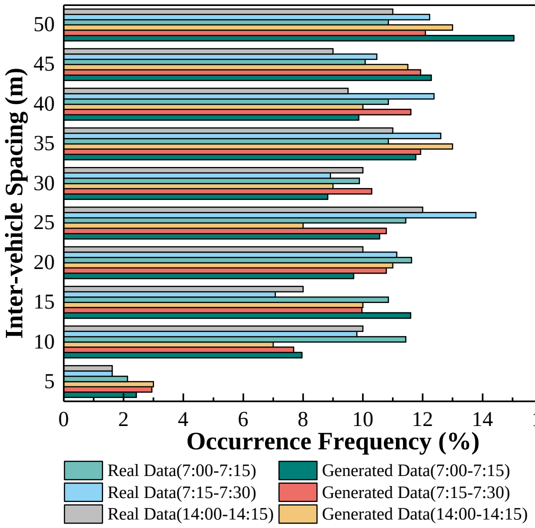


Fig. 12. Inter-vehicle spacing occurrence frequency.

VII. CONCLUSION

The lack of a private car trajectory dataset hinders the research in-vehicle communication and other related fields. To solve this problem, we propose a tri-layer model to generate vehicular trajectory datasets, called RMGen. We discussed the trajectory dataset generation process based on taxi GPS data and urban vehicular social network information and then validate the authenticity and accuracy of the generated data. We randomly generated a trajectory dataset of 13,299,921 private cars on a particular day within the fifth ring road of Beijing. There is no uniform standard for verification of generation trajectory dataset. So we present a novel method, which compares our dataset with the real traffic situations from macroscopic and macroscopic perspectives. The results under our validation model show that our method has high accuracy and universality. Our model provides a broad prospect for the studies which get tough for the lack of relevant data.

Our method performs better in heavy density scenarios than in low density scenarios, which is a limitation of our work. In the future, we will consider multiple factors that may affect travel patterns, such as weather and travel costs, to build a more broadly applicable mobility model. In addition, our proposed ARS method also has specific requirements on the degree of regularity of urban roads in the region division. Results may suffer when urban roads become less regular. In future work, we will further optimize our method to make it more general.

Acknowledgment

The authors would like to thank Di Shang for her contributions in the preparation of this paper.

References

[1] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibaez, "Internet of vehicles: Architecture, protocols, and security," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3701–3709, 2018.

[2] A. M. Vegni and V. Loscr, "A survey on vehicular social networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2397–2419, 2015.

[3] M. Ahmad, F. Abbas, Q. Chen, and M. Ahmad, "A novel hybrid contents oriented communication (coc) technique based on v2x networks," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–5.

[4] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, "Effective urban traffic monitoring by vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 1, pp. 273–286, 2015.

[5] X. Kong, F. Xia, J. Li, M. Hou, M. Li, and Y. Xiang, "A shared bus profiling scheme for smart cities based on heterogeneous mobile crowdsourced data," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1436–1444, 2020.

[6] O. S. Oubbati, A. Lakas, P. Lorenz, M. Atiquzzaman, and A. Jamalipour, "Leveraging communicating uavs for emergency vehicle guidance in urban areas," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 2, pp. 1070–1082, 2021.

[7] O. S. Oubbati, M. Atiquzzaman, P. Lorenz, A. Baz, and H. Alhakami, "Search: An sdn-enabled approach for vehicle path-planning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 523–14 536, 2020.

[8] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1549–15 498.

[9] L. Codeca, R. Frank, and T. Engel, "Luxembourg sumo traffic (lust) scenario: 24 hours of mobility for vehicular networking research," in *2015 IEEE Vehicular Networking Conference (VNC)*, 2015, pp. 1–8.

[10] M. A. Dian Khumara, L. Fauziyyah, and P. Kristalina, "Estimation of urban traffic state using simulation of urban mobility(sumo) to optimize intelligent transport system in smart city," in *2018 International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*, 2018, pp. 163–169.

[11] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2011, pp. 384–388.

[12] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei, "Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method," *Landscape and Urban Planning*, vol. 160, pp. 48–60, 2017.

[13] L. Wang, F. Fang, X. Yuan, Z. Luo, Y. Liu, B. Wan, and Y. Zhao, "Urban function zoning using geotagged photos and openstreetmap," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 815–818.

[14] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 186194.

[15] A. Rahim, X. Kong, F. Xia, Z. Ning, N. Ullah, J. Wang, and S. K. Das, "Vehicular social networks: A survey," *Pervasive and Mobile Computing*, vol. 43, pp. 96–113, 2018.

[16] F. Xia, J. Wang, X. Kong, Z. Wang, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 142–149, 2018.

[17] X. Kong, S. Tong, H. Gao, G. Shen, K. Wang, M. Collotta, I. You, and S. K. Das, "Mobile edge cooperation optimization for wearable internet of things: A network representation-based framework," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5050–5058, 2021.

[18] N. Pourebrahim, S. Sultana, J.-C. Thill, and S. Mohanty, "Enhancing trip distribution prediction with twitter data: Comparison of neural network and gravity models," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. Association for Computing Machinery, 2018, p. 58.

[19] I. Hong and W.-S. Jung, "Application of gravity model on the korean urban bus network," *Physica A: Statistical Mechanics and its Applications*, vol. 462, pp. 48–55, 2016.

[20] A. P. Masucci, J. Serras, A. Johansson, and M. Batty, "Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows," *Physical Review E*, vol. 88, no. 2, p. 022812, 2013.

[21] X. Yan, C. Zhao, Y. Fan, Z. Di, and W. Wang, "Universal predictability of mobility patterns in cities," *Journal of The Royal Society Interface*, vol. 11, no. 100, p. 20140834, 2014.

[22] J. Simon and P. G. Furth, "Generating a bus route o-d matrix from on-off data," *Journal of Transportation Engineering*, vol. 111, no. 6, pp. 583–593, 1985.

[23] M. Friedrich, K. Immisch, P. Jehlicka, T. Otterstter, and J. Schlaich, "Generating origindestination matrices from mobile phone trajectories," *Transportation Research Record*, vol. 2196, no. 1, pp. 93–101, 2010.

[24] V. Caiati, L. Bedogni, L. Bononi, F. Ferrero, M. Fiore, and A. Vesco, "Estimating urban mobility with open data: A case study in bologna," in *2016 IEEE International Smart Cities Conference (ISC2)*, 2016, pp. 1–8.

[25] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks: a survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, pp. 19–41, 2009.

[26] Y. Pigné, G. Danoy, and P. Bouvry, "A vehicular mobility model based on real traffic counting data," in *Proceedings of the Third International Conference on Communication Technologies for Vehicles*, 2011, p. 131142.

[27] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining and Knowledge Discovery*, vol. 32, p. 787829, 2018.

[28] X. Kang, L. Liu, D. Zhao, and H. Ma, "Trag: A trajectory generation technique for simulating urban crowd mobility," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 820–829, 2021.

[29] L. Bedogni, M. Gramaglia, A. Vesco, M. Fiore, J. Härri, and F. Ferrero, "The bologna ringway dataset: improving road network conversion in sumo and validating urban mobility via navigation services," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5464–5476, 2015.

[30] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," *IEEE Transactions on Mobile Computing*, vol. 13, no. 5, pp. 1061–1075, 2014.

[31] R. Ketabi, B. Alipour, and A. Helmy, "En route: Towards vehicular mobility scenario generation at scale," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2017, pp. 839–844.

[32] X. Kong, X. Feng, Z. Ning, A. Rahim, Y. Cai, Z. Gao, and J. Ma, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3874–3886, 2018.

[33] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo-simulation of urban mobility: An overview," in *in SIMUL 2011, The Third International Conference on Advances in System Simulation*, 2011, pp. 63–68.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[35] D. Mimno and A. Mccallum, “Topic models conditioned on arbitrary features with dirichletmultinomial regression,” in *In Uncertainty in Artificial Intelligence*, 2008, pp. 411–418.

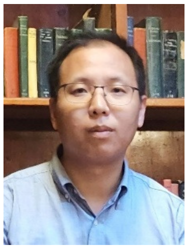
[36] J. Z. Lai, T.-J. Huang, and Y.-C. Liaw, “A fast k-means clustering algorithm using cluster center displacement,” *Pattern Recognition*, vol. 42, no. 11, pp. 2551–2556, 2009.

[37] J. Kim and C. D. Scott, “Robust kernel density estimation,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2529–2565, 2012.

[38] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns,” *Nature*, vol. 484, no. 7392, p. 96, 2012.

[39] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, “On the assessment of vehicle trajectory data accuracy and application to the next generation simulation (ngsim) program data,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.

[40] A. Paz, V. Molano, and C. Gaviria, “Calibration of corsim models considering all model parameters simultaneously,” in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 2012, pp. 1417–1422.

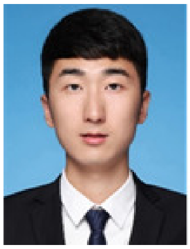


Xiangjie Kong (M13-SM17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Full Professor with College of Computer Science and Technology, Zhejiang University of Technology. Previously, he was an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 130 scientific papers in international journals and conferences (with over 100 indexed by ISI SCIE). His research interests include network science, mobile computing, and computational social science.

Qiao Chen received the B.S. degree in software engineering from Guizhou University, Guiyang, China. She is currently pursuing the M.S. degree in software engineering with the Dalian University of Technology, Dalian, China. Her research interests mainly include social computing, spatio-temporal analysis and graph learning.

Mingliang Hou received the B.Sc. degree from Dezhou University and the M.Sc. degree from Shandong University, Shandong, China. He is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian, China. His research interests include graph learning, city science and social computing.

Azizur Rahim received the M.S. degree in Electrical Engineering from the COMSATS, Islamabad, Pakistan, and the Ph.D. degree in computer application technologies from The Alpha Laboratory, School of Software, Dalian University of Technology, Dalian, China. He is currently working as an Assistant Professor with the Department of Computer Systems Engineering, the University of Engineering and Applied Sciences, Swat, Pakistan. His research interests include mobile and social computing, mobile social networks, and vehicular social networks.



Kai Ma received the master’s degree in software engineering, from Dalian University of Technology, Dalian, China, in 2020. He is currently a operation and maintenance engineer at the data center of China Zheshang bank. His research interests include data analysis, automation of operation and maintenance.

Feng Xia (M07-SM12) received the BSc and PhD degrees from Zhejiang University, Hangzhou, China. He was Full Professor and Associate Dean (Research) in School of Software, Dalian University of Technology, China. He is Associate Professor and former Discipline Leader (IT) in Institute of Innovation, Science and Sustainability, Federation University Australia. Dr. Xia has published 2 books and over 300 scientific papers in international journals and conferences. His research interests include data science, artificial intelligence, graph learning, anomaly detection, and systems engineering. He is a Senior Member of IEEE and ACM.