Shared Subway Shuttle Bus Route Planning Based on Transport Data Analytics

Xiangjie Kong, *Senior Member, IEEE*, Menglin Li, Tao Tang, Kaiqi Tian, Luis Moreira-Matias, *Member, IEEE*, and Feng Xia, *Senior Member, IEEE*

***Abstract*— The development requirements of shared buses are extremely urgent to alleviate urban traffic congestions by improving road resource utilization and to provide a neotype transportation mode with good user experiences. The key to shared bus implementation lies in accurately predicting travel requirements and planning dynamic routes. However, the sparseness and the high volatility of shared bus data bring a great resistance to accurate prediction of travel requirements. Based on the consideration of user experiences, optimization objectives of shared bus route planning are significantly different from traditional public transportation and shared bus route planning is far more challenging than online car-hailing services due to the relatively high number of passengers. In this paper, we put forward a two-stage approach (SubBus), which is composed of travel requirement prediction and dynamic routes planning, based on various crowdsourced shared bus data to generate dynamic routes for shared buses in the "last mile" scene. First, we analyze the resident travel behaviors to obtain five predictive features, such as flow, time, week, location, and bus, and utilize them to predict travel requirements accurately based on a machine learning model. Second, we design a dynamic programming algorithm to generate dynamic, optimal routes with fixed destinations for multiple operating buses utilizing prediction results based on operating characteristics of shared buses. Extensive experiments are performed on real crowdsourced shared subway shuttle bus data and demonstrate that SubBus outperforms other methods on dynamic route planning for the "last mile" scene.**

***Note to Practitioners*—This paper is inspired by the problem of shared subway shuttle bus dynamic route planning for the "last mile" scene, and it is also applicable to other scenes, including commuting scenes, urban transportation hub scenes, and destination scenes of the tourist market. Shared bus operation routes at such scenes are usually aimed at trips with fixed destinations. Existing approaches to planning routes are generally designed for traditional transportation, such as traditional buses and taxis. In this paper, we propose a novel two-stage dynamic route planning approach (SubBus) based on the operation characteristics of shared subway shuttle buses.**

We perform a resident travel behavior analysis to improve the accuracy of travel requirement prediction. After that, we combine the prediction results and station properties to gain shared bus optimal routes. We then display how to apply SubBus to optimize shared bus operation status based on crowdsourced shared subway shuttle bus data generated by Panda Bus Company. We keep a continuous collaboration with the company to optimize the approach details and experimental effects, which demonstrate that our approach can generate effective routes for shared subway shuttle buses to optimize operation status on the "last mile" issue.

***Index Terms*— Crowdsourced data, passenger flow prediction, route planning, shared buses.**

I. Introduction

DRIVEN by the rapid development of information technologies, sharing economy has turned into a burgeoning economic paradigm. Sharing economy realizes cooperative consumption by simplifying sharing of services and the right to use things, to increase resource utilization from the perspective of resource redistribution [1], which promotes green consumption and sustainable development in smart cities. The solution to resource shortage thus ushers in an important turning point. How to expand practical applications of sharing economy in smart cities to ease resource shortage and to bring about huge benefits for citizen living and economic development has become an issue of close concern to scholars in multiple fields [2], [3].

Ridesharing is to integrate the same trips of passengers within a certain period, and such a period is usually short, such as 0.5 h. Ridesharing can be regarded as the application of sharing economy in the field of transportation and its concrete implementation forms include shared bicycles, shared cars, and shared buses [4]. Online car-hailing services and car rent services are considered to belong to the application field of shared cars. With the rapid development of ridesharing transportation mode represented by Didi Travel and Mobike in recent years, the travel mode of urban residents has undergone significant changes. In the field of personal travel, the one-stop service platform represented by Uber and Didi Travel has increased the operational efficiency of online car-hailing services and, meanwhile, has crossed the stage of travel

Manuscript received October 24, 2017; revised June 11, 2018; accepted August 7, 2018. Date of publication September 10, 2018; date of current version October 4, 2018. This paper was recommended for publication by Associate Editor W. Tan and Editor M. P. Fanti upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 61572106, in part by the Natural Science Foundation of Liaoning Province, China, under Grant 201602154, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18JC09. *(Corresponding author: Feng Xia.)*

X. Kong, M. Li, K. Tian, and F. Xia are with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China (e-mail: xjkong@ieee.org; cookies.s@outlook.com; cagetian@outlook.com; f.xia@ieee.org).

T. Tang is with the Chengdu College, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: tang.tau@outlook.com).

L. Moreira-Matias is with NEC Laboratories Europe, 69115 Heidelberg, Germany (e-mail: luis.moreira.matias@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2018.2865494

resource integration moving toward the direction of intelligent travel [5].

As the most important part of urban transportation, public transportation takes on the majority of urban residents' travel [6]. Therefore, the development status of public transportation directly affects the traffic conditions of the entire city [7]. Moreover, public transportation has the advantages of high road utilization, environmentally friendly, and so on, which private transportation does not own. However, from the perspective of ridesharing, the development of public transportation lags far behind personal travel. Most cities still stay in the stage of information sharing, that is, real-time queries of locations of buses and vehicles. The public transportation system does not support online booking, ticket purchase, and other services, so it cannot meet passengers' demand for direct travel to destinations just relying on the existing lines. The public transportation system cannot deploy transportation resources according to the real-time situation, so it is difficult to further increase resource utilization and the cost remains high. The poor travel experience of public transportation makes passengers incline to personal transportation. As mentioned in *Harvard Business Review*, commercial competitions in the Internet era make the market more sensitive to service quality than prices. What is more, Chinese President Jinping Xi also emphasized in the report of the 19th National Compress that the major social contradictions in China have been transformed into the contradiction between the people's ever-growing needs for a better life and unbalanced uneven development. People are increasingly demanding high quality of life. A convenient and quick public transportation mode with a good user experience is urgent.

Under such pressing social needs, shared bus is at the historic moment. Shared bus is committed to developing a transportation mode to make up the gap among current online car-hailing services, taxis, and traditional public transportation, providing a convenient and inexpensive door-to-door travel experience. Shared bus integrates the same fragmented trips in multiple scenarios and dynamically allocates the transportation resources to provide bus services. Its operational scenarios embrace commuting scene, airport or railway station and other urban transport hub subscenes, destination scene of the tourist market, and the "last mile" issue. By integrating fragmented trips, the shared bus can improve resource utilization effectively and reduce the operating costs of public transportation to promote the long-term development of public transportation in smart cities. Compared with traditional transportation modes, shared bus has the following characteristics.

1) *Inexpensiveness:* The fare for shared bus is much lower than that for taxis and online car-hailing services and slightly higher than buses and subways.
2) *Convenience:* Shared bus is a kind of short-distance dynamic shuttle based on human travel needs, enabling users to take the bus such as taking a taxi.
3) *High Resource Utilization:* In addition to the efficient use of vehicle resources, shared bus can also effectively save road resources, which is quite important under current unprecedented tension in per capita road resources in large cities.

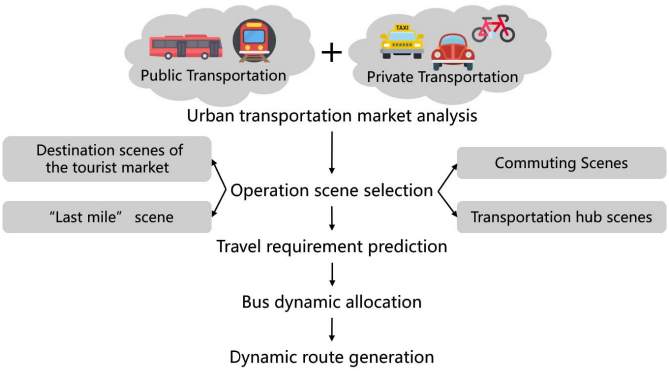


Fig. 1. Implementation process of shared bus.

4) *Environmentally Friendly:* The wide popularity of shared bus can reduce the emissions of greenhouse gas in cities.

As shown in Fig. 1, the implementation process of shared bus is analyzing the urban transportation operation market first to find and select valuable and potential operation scenes, then analyzing human travel patterns under such scenes and predicting travel requirements, and allocating buses dynamically to generate flexible routes in the end. In this way, fragmented trips are integrated into dynamic, nonfixed bus operational lines based on passenger travel requirements. Keys to shared bus implementation lie in accurately predicting travel requirements and planning dynamic routes, which provide the motivation for this paper.

We focus on how to predict travel requirements and plan dynamic routes in this paper. As we know, route planning is a multiobjective optimization problem. Oriented by the characteristics of the shared bus, we set an operating distance as the main optimization goal and passengers' number as a constraint condition to do route planning. The optimization goal of traditional bus route planning is usually the number of passengers. Thus, bus lines are not the shortest routes, which lead to excessive travel time and a decrease in user experience. The cost of a taxi can be shared by passengers. However, for a bus, if there are too few passengers, the operation cost cannot be covered. On the contrary, too many passengers bring poor travel experiences and security risks. Therefore, the ideal number of passengers should close to and not exceed the number of bus seats. Both being dynamic travel, the technical requirements of a bus are much higher than a taxi, because a trip has at least 10 passengers, which makes all aspects of the application scene difficult [8].

Commuting time has been a headache for most workers, especially for "last mile" scene. Therefore, we focus on solving the shared bus route planning problem under "last mile" scene. To be more especially, the scene is from residential regions to nearby subway stations. However, the solution for "last mile" issue is also applicable to other application scenes where passenger travel requirements change dynamically and the destination is fixed. In this paper, we propose a two-phased approach (SubBus) to plan the dynamic routes for shared subway shuttle bus. To the best of our knowledge, this is the first route planning work specifically for shared buses. First, based on a passenger behavior analysis, we identify the multidimensional properties to predict passenger travel

requirements, which are the distribution and volume of passengers at different time intervals. Then, we design a dynamic programming algorithm to get the optimal routes based on the prediction results. Our major contributions can be generalized as follows.

1) We put forward five predictive features, such as flow, time, week, location, and bus, to predict travel requirements of shared buses accurately to overcome the difficulty caused by the sparseness and high volatility of crowdsourced shared bus data.
2) We design a dynamic programming algorithm to generate dynamic, optimal routes with fixed destinations for multiple operating buses in the "last mile" scene based on the operating characteristics of shared buses.
3) We integrate the designed dynamic programming algorithm with the five predictive features into a shared bus dynamic route planning approach (SubBus) to plan flexible routes based on the predicted dynamically changing travel requirements.
4) We evaluate SubBus with three state-of-the-art prediction models utilizing various metrics and a dynamic route algorithm based on real shared bus data to demonstrate the effectiveness and stability of SubBus.

The rest of this paper is structured as follows. In Section II, we review the related work on traffic flow prediction and traffic route planning. In Section III, we formulate the route planning problem of shared buses. In Section IV, we present the details of SubBus. Data description and experiment result analysis are displayed in Section V. Finally, we conclude this paper and chart the future directions in Section VI.

II. Related Work

This section discusses the prior studies closely related to this paper.

A. Traffic Flow Prediction

Traffic flow prediction has high application value in many fields, such as smart cities, intelligent transportation services, vehicle social networks, and route planning. Traffic flow prediction models can be roughly divided into the following four categories: linear prediction model, nonlinear prediction model [9], artificial neural network prediction model [10], and hybrid models combining of the above-mentioned models [11]. A linear prediction model utilizes the historical data to predict traffic flow based on the periodic change rule of urban traffic travel [12], [13]. The linear prediction models mainly include a linear regression model, Kalman filter model, and time series statistical model [14], such as autoregressive integrated moving average model [15]. Zhang *et al.* [16] construct a two-step real-time prediction linear model based on the historical and current patterns to predict passenger flow in the future. As urban traffic has great volatility and randomness, the further analysis of this characteristic needs a nonlinear theory, which contains a nonparametric regression method, analysis based on wavelet theory, and other methods [17]. With the continuous complication of urban transportation networks, a traditional linear prediction theory cannot satisfy people's requirements for traffic flow predict accuracy. Artificial neural networks can simulate complex nonlinear mapping relationships between multiple variates quite closely, so artificial neural networks are increasingly widely used in traffic flow prediction. Zhang *et al.* [18] put forward a prediction model based on a deep neural network. Yang *et al.* [19] use a deep learning approach, that is, neural network approach, to optimize the structure of traffic flow forecasting model. In addition to artificial neural network traffic flow forecasting models, some scholars focus on a hybrid predicting model based on a neural network [20], [21]. A deep belief network and a multitask regression method are combined to form a traffic prediction approach to predict the traffic flow of multitask output and single-task output [22]. For traditional traffic data, such as subway transaction card data, and taxi GPS data, current predictive models have a good performance on accuracy and stability. However, for shared bus data, which is involved in this paper, the fluctuation of data is marked and the data are sparse. Traditional methods cannot provide the satisfactory prediction effects. Therefore, we propose a multifeature-based shared bus passenger flow prediction method according to the analysis of human travel behaviors.

B. Traffic Route Planning

Shared bus route planning falls into the general topic of urban traffic route planning. For urban travel, how to choose an optimal route that has less travel time, shorter distance, more passengers, and other advantages directly affects the quality of resident lives [23]. Buses and taxis are the most primary and most important transportation modes in cities [24], and their route planning issues have always attracted the continuous interests of numerous scholars [25], [26]. A great number of city's bus lines constitute a complex bus network, which leads scholars to optimize bus lines from graph-based algorithms [27]. Chen *et al.* [28] propose a two-phase approach for bidirectional night-bus route planning and develop a bidirectional probability based on a spreading algorithm. Bastani *et al.* [29] propose an optimal single flexible route discovery algorithm on graph searching. Wang *et al.* [30] present timetable labeling, an efficient indexing technique for bus route planning on timetable graphs. Liu *et al.* [31] focus on the identification and optimization of flawed region pairs with problematic bus routing according to people's real demands for public transportation. Compared to fixed bus routes, taxi route planning places a greater emphasis on dynamics and flexibility [32]. Yuan *et al.* [33] propose a time-dependent landmark graph to model the intelligence of taxi drivers and the properties of dynamic road networks and design a two-stage routing algorithm to compute the practically fastest route. Li *et al.* [34] present a new experiential approach that computes optimal paths by mining floating car trajectories to do fast path finding through a flexible hierarchical road network. The research on route planning of public transportation and taxis is relatively mature. Nevertheless, as far as we know, route planning methods, specifically for sharing buses, have not yet emerged. The dynamics of operating routes, the particularity of the optimization goal, and the constraint of passengers' number for

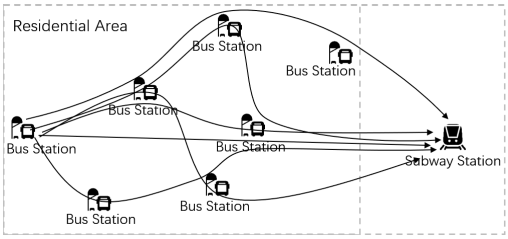


Fig. 2. Problem formulation.

shared bus route planning make a specialized route planning method quite necessary, which is exactly the focal point of this paper.

III. Problem Formulation

The operating scenes of shared bus mainly include commuting scenes, airports, railway stations, and other urban transportation hub scenes, and destination scenes of the tourist market. Commuting time has always been a headache for many office workers, especially the last few kilometers. Therefore, in this paper, we focus on an important subscene in the commuting scene: "last mile" issue. More specially, this paper is from residential areas to nearby subway stations, as shown in Fig. 2. Thus, the issue that we attempt to deal with is how residents can quickly and conveniently reach nearby subway stations from residential areas.

On an abstract level, this paper is a route planning problem from multiple identified stations to a single destination and routes change with human travel requirements. Oriented by the operating characteristics of shared buses, we set an operating distance as the main optimization goal and passengers' number as a constraint condition to do route planning. In other words, our method is devoted to finding dynamic and flexible routes with short operating distance, less operating hours, whose passengers' number is close to but do not exceed the number of bus seats. Based on the characteristics of shared bus route planning, we intend to tackle the formulated route planning problem from perspectives of travel requirement prediction and dynamic route planning, and our proposed approach SubBus is described in detail in Section IV. The approach proposed for "last mile" issue is also applicable for other scenes with similar features.

IV. SubBus Approach

In this section, we first introduce the framework of our proposed approach SubBus. Then, the detailed description of the shared bus dynamic route planning approach is presented from two aspects: travel requirement prediction and dynamic route planning.

A. Route Planning Framework

Fig. 3 presents the framework of SubBus. It consists of three major components: 1) data preprocessing; 2) travel requirement prediction; and 3) dynamic route planning.

1) Data Preprocessing: Typical traffic data preprocessing operations include data cleaning, data organization, data mapping, and data aggregation [35]. For crowdsourced shared bus order data and GPS data, we first perform the data filtering

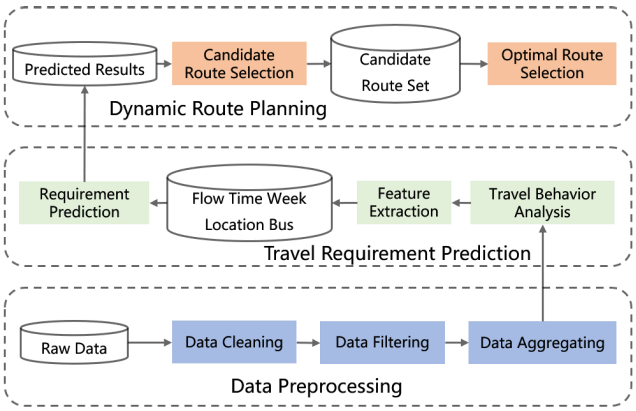


Fig. 3. Framework of SubBus.

operation to extract useful information from raw data, then clear out errors and outliers, and zeroize missing values during the data cleaning phase. Finally, the data are aggregated in the time dimension. According to the operational law of shared buses, we divide the study time period into small intervals of 10 min.

2) Travel Requirement Prediction: The travel requirement prediction component contains three operations: travel behavior analysis, feature extraction, and requirement prediction. Before predicting travel requirements, it is necessary to fully understand travel behaviors. Therefore, based on the preprocessed data, we first analyze the passenger travel behaviors of shared buses and obtain the highly time-dependent, location-dependent, sparse, and other characteristics of travel behaviors. Then, based on the above-mentioned characteristics, we define multiple features that have a significant impact on travel requirement prediction, including flow, time, week, location, and bus, and extract them from the data. Based on the extracted features, an effective machine learning model, XGBoost, which is an optimized distributed gradient boosting library, is utilized to predict the travel requirements of shared buses.

3) Dynamic Route Planning: As we know, the route planning issue is actually a multiobjective optimization problem. Therefore, as shown in Fig. 3, the dynamic route planning component includes possible route generation and optimal route selection. First of all, we determine the candidate origins based on station information and passenger travel requirements. Then, starting from the candidate origin set, combined with road network information of the research environment, a candidate route set is generated. Finally, we design a dynamic programming method, using predicted results, taking the operation distance as the optimization goal and the number of passengers as a constraint condition, to implement the dynamic route planning of multiple buses operating at the same time.

B. Travel Requirement Prediction

Travel behaviors of residents actually reflect a kind of human mobility, which contains a wealth of information [36], [37], such as passenger preferences, that is, when passengers prefer to go out and which stations they would like

TABLE I
QUANTITATIVE DESCRIPTION OF THE FIVE PREDICTIVE FEATURES

Feature	Measure	Annotation
Flow	1DayBefore	Passengers' number at corresponding time interval of a day before the target day
	2DaysBefore	Passengers' number at corresponding time interval of two days before the target day
	3DaysBefore	Passengers' number at corresponding time interval of three days before the target day
Time	IntervalLabel	The number of time intervals
Week	Week	In Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, the one the target day belonged to
Location	TimeGap	Travel time from the target station to the destination
Bus	DaysDifference	Difference in days between the target day and the abrupt day

to take buses at. However, influencing factors of travel behaviors are complex and varied, such as mood, economic levels, weather, and so on [38]. Passenger travel behaviors have a high degree of time dependence and location dependence. Time dependence is not only reflected in the multilevel periodicity of human mobility but also in the complex sequential transition regularities [39]. In cities, human travel patterns in residential areas and workspace are significantly different, which is exactly the typical manifestation of location dependence of travel behaviors. Travel requirements are hidden in the patterns and changes in travel behaviors. Thence, for travel requirement prediction, significance and challenges coexist. Combined with the characteristics of passenger travel behaviors, we define the following five features that are critical for travel requirement prediction of shared buses.

We define the five prediction features: flow, time, week, location, and bus as follows. Table I presents the quantitative description of the five predictive features.

1) Flow: As we mentioned earlier, travel behaviors have complex sequential transitional regularities. Therefore, in passenger flow prediction, the most relevant influencing factor is historical flow. Traffic flow prediction studies based on linear models employ historical flow as the only input [15]. In recent studies, prediction models are increasingly complex and multiple factors are considered, but the historical flow has always been the most important element that cannot be ignored [19]. Similarly, in our prediction model, we take the historical flow as the most crucial feature and quantify it as passengers' number at the corresponding time interval on one day before the target day, two days before the target day, and three days before the target day. After the data are aggregated in the time dimension, the value of this feature can be obtained.

2) Time: Travel behaviors are highly time-dependent and have multilevel periodicity. Daily travel behaviors have a strong regularity, such as similar peaks and troughs. Under the increasingly serious traffic conditions, when to go out can escape traffic jams is the question people must consider before going out. That is, time is a key factor that affects traffic flow. We use the number of time intervals as quantification values of this feature. The specific calculation process is as follows, where T_k is the number of time intervals and θ is the length of a time interval, and the values of $\text{Time}_{\text{lower}}$ and $\text{Time}_{\text{upper}}$ can be set according to the special experiments

$$\begin{cases} T_k = k_0, & \text{Time} < \text{Time}_{\text{lower}} \\ T_k = [k\theta, (k+1)\theta], & k = k_1, k_2 \ldots k_{n-1} \\ T_k = k_n, & \text{Time} > \text{Time}_{\text{upper}}. \end{cases} \tag{1}$$

3) Week: In addition to the peak trough law in a day, the weekly nature of traffic flow is also obvious. People go to work or school at a fixed time on weekdays, which results in fixed morning peaks, noon peaks, and evening peaks. At weekends, their travel time is relatively scattered and the peak trough law is relatively weak. Moreover, even equally being weekdays, the travel rules of Monday and Friday are not the same, especially on Monday morning and Friday evening. Therefore, in order to reflect the weekly nature of traffic flow, we value Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday as 0, 1, 2, 3, 4, 5, and 6 to form the weekly property.

4) Location: Everyone who has taken a bus has such an experience; some stations are overcrowded, while some stations are almost no one, which is location dependence of travel behaviors. Station locations can be a good explanation for this phenomenon. Consequently, for travel requirement prediction, station's location information is a property that we cannot ignore. The number of neighbor stations around the station, the distance, and travel time between the station and the nearest station, and the distance and travel time between the end and the station all can affect the number of passengers at the station. After thinking over, we choose the time from the station to the destination as the measure of such location feature and its value can be obtained from the GPS data. The travel time from the station to the destination is varied over time. This property can also reflect the time law from a new perspective.

5) Bus: For traffic flow prediction based on taxi data or bus data, the data performance is stable and the fluctuation is small because of large data volume. In this paper, we use crowdsourced shared bus data, which is sparse, and the changes of buses' number produce significant fluctuations of passenger flow. We take the above-mentioned characteristics into consideration in travel requirement prediction and quantify it as the difference in the days between the target day and the abrupt day when the number of buses changes. In Section V, we find that this feature is extremely effective for improving the prediction accuracy.

Utilizing the five features we defined, we employ the XGBoost model to travel requirement prediction of shared buses to lay the foundation for dynamic, flexible route planning.

C. Dynamic Route Planning

The goal of the second phase of our proposed approach (SubBus) is to plan dynamic routes for shared

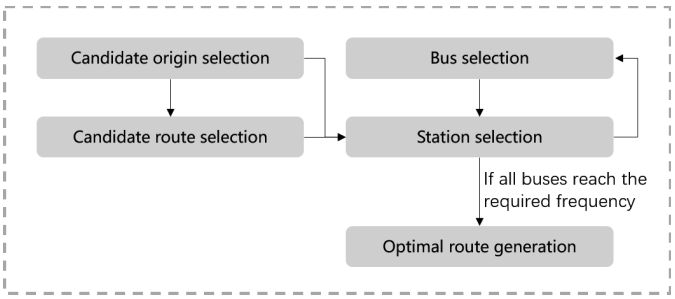


Fig. 4. Two-step dynamic route planning framework.

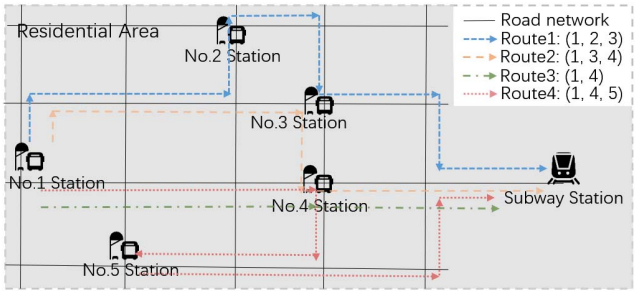


Fig. 5. Example of candidate route selection.

buses based on predicted travel requirements. In this section, we describe in detail the dynamic route planning process. As shown in Fig. 4, the whole dynamic route planning process consists of the following two steps: 1) select candidate origins and then select candidate routes and 2) select buses and select stations based on the candidate origin set and the candidate route set and finally generate optimal routes if all buses reach the required frequencies. The idea of generating a candidate set first and then selecting from it is widely applied in route planning researches and is verified to be effective [28].

1) Candidate Route Selection: We have already emphasized that the dynamic route planning problem in this paper is with a single destination. From the residential area to the nearby subway station, there are countable multiple stations within the spatial scope. Therefore, based on the station distribution topology and given origins, the idea to exhaust all routes that can reach the destination is practically feasible, but it is also obvious that not all accessible routes are suitable and an excessive number of routes increase the difficulty of selecting optimal routes. From the perspective of the optimization goal, such as travel distance, we limit the routes to be not nonbackward and loopfree. Apparently, backward and ringed routes will plus actual operating distance and damage user experiences. Therefore, we first extract actual routes from GPS data and select origins from all stations based on the empirical values to build a candidate origin set. Then, starting from candidate origin set, we exhaust all nonbackward, loopfree accessible routes combining with road network information of the study area. In this way, the candidate origin set and the candidate route set are all generated.

Fig. 5 shows an example of candidate route selection. In the residential area, there are five bus stations and we choose No. 1 station as the origin. From No. 1 station, multiple routes can reach the nearest subway station. We list some of all accessible routes in Fig. 5. Among them, the route indicated by the dashed lines produces a backtracking, which increases the travel distance, so the route does not belong to candidate routes. Therefore, in this example, the candidate route set from No. 1 station consists of routes 1–3.

2) Optimal Route Selection: After acquiring the candidate origin set and the candidate route set, we focus on how to select optimal routes. Dynamic route planning of shared buses is a route planning problem with multiple buses operating at the same time. Compared with the planning issue of a single optimal route, such a problem is pretty difficult. We need to consider the operation status of all buses and travel requirements of all stations. We cut out a time slot of shared buses under operating status: among multiple buses and multiple stations (travel requirements are varied over stations), which bus should we choose to pick up at which site? Therefore, two core issues are involved in the dynamic route planning: bus selection and station selection. Continuous station selection from the origin to the destination results in a route. An optimal route means that each station it contains is optimal, which is exactly in line with the idea of splitting the optimal solution into multiple optimal substructures in dynamic programming. Thus, we split the optimal route selection into a continuous selection of optimal stations, and the interstation and interbus selections affect each other and will generate the linkage effect.

a) Bus selection: When multiple buses are all in operation, how to select a bus? We set a weight for each bus and prioritize buses to perform the next step based on their weights. We quantify this weight as the current time of the bus, that is, the time after picking up passengers at stations. Now, the bus is in the state of finding the next station. For buses not in routes, their time is the departure time. The bus with the least weight has the priority for the next station selection. For example, compared with the bus at 7:20, the one at 7 o'clock has the priority. For the whole route planning process, each bus will be assigned to stations to form routes, so bus selection does not affect scheduling results. However, the bus selection operation is essential when multiple buses operate at the same time. The pseudocode of bus selection is shown in Algorithm 1.

b) Station selection: After selecting buses, we can get the previous station of a bus. In conjunction with the station and candidate route set, we can generate an optional set of the next station for the bus, $\text{Sta}_{\text{optimal}}$. Which of these stations should be selected as the next station? Similarly, we set a weight for each station. The calculation of this weight is shown in 3. We first calculate the arrival time of a bus, T_{arr}, reaching each station in set $\text{Sta}_{\text{optimal}}$. Then, the number p of the time interval that T_{arr} lies in is also available. Based on the prediction results, the number of passengers Pass on each station at the pth time interval is known. T_{arr} is likely to fall within a time interval. How many passengers should the bus pick up? Here, we assume that the number of passengers on stations is evenly distributed over time, i.e., passengers randomly reach stations within time intervals. The number of passengers that the bus can pick up is calculated according to the proportion of time length, as shown in 2. In addition to passengers at

Algorithm 1 Bus Selection

Input: Rou_i: list of routes, which contain the information of stations, time, and seats;
$Rou_i.Sta[]$: list of stations;
$Rou_i.T[]$: list of current time;
$Rou_i.Sea[]$: list of seats;
i: bus ID, (0, 1...n-1)
Output: i: next bus id for selecting stations
1.**procedure:** select a bus to prioritize station selection
2.if $Rou_0.Sta = \Phi$ do
3. $i \leftarrow 0$;
4. return i;
5.else do
6. for each route $\in Rou_i$ do
7. $T_{previous}[] \leftarrow Rou_i.T[-1]$ \\ $T_{previous}[]$ is the list of previous station arriving time;
8. end for
9. $Rou.sort(T_{previous})$\\ Sort Rou_i according to $T_{previous}$ by increasing order;
10. $i \leftarrow 0$;
11. return i;
12.end else
13.end if
14.end procedure

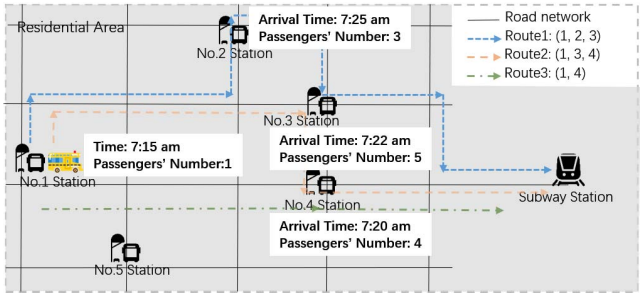


Fig. 6. Example of station selection.

the pth time interval, there may be passengers who have not been picked up before the pth time interval. Therefore, it is necessary to add the number of historical passengers in 3. Note that during the station selection process, we update the data of pass dynamically, so $\sum_{k=0}^{p-1} \text{Pass}_k$ represents the number of historical passengers. In this way, we can get the weight of each station and choose the station with the highest weight as the next station. After the bus picks up the passengers, we update the number of passengers in the bus

$$\text{rate} = \frac{T_{\text{arr}} - T_{\text{start}}[p]}{\theta} \tag{2}$$

$$\text{Pass}_{\text{pickup}} = \sum_{k=0}^{p-1} \text{Pass}_k + \text{Pass}_p * \text{rate}. \tag{3}$$

Algorithm 2 presents the pseudocode of station selection. Fig. 6 shows an example of station selection. The bus just picked up passengers at No. 1 station at 7:15 A.M. Based on the candidate route set, there are No. 2 station, No. 3 station, and No. 4 station, and the three stations can be selected. For the bus, it takes 10, 7, and 5 min to reach at No. 2 station, No. 3 station, and No. 4 station, respectively. Assuming that the length of time intervals is 10 min, only the arrival time of No. 4 station does not fall in the time interval, and the other two stations need to consider historical passengers. Assume that the historical numbers of No. 2 station and No. 3 station are 1 and 4, respectively, and the number of people in the current time interval is 4 and 5. The No. 4 station only considers the current number of passengers, which is 4. The station with the largest number of passengers is 3, that is, the bus selects No. 3 station as the next station.

At the same time, we need to limit the number of passengers in buses due to the seats' number. When the number of passengers reaches the number of seats, the bus will directly head to the destination. In real life, most drivers will choose this approach to avoid excess passengers. Our method can sense such path changes and make the corresponding improvements in the subsequent bus scheduling. In the actual operation, buses generally have restrictions of frequencies, so in our method, we also consider this factor. When a bus reaches the required frequency, it will no longer participate in the schedule. The value of frequencies is based on empirical values. In general, the setting of frequencies can satisfy travel requirements and will not affect scheduling results.

V. Experiments

We conduct a case study of SubBus to demonstrate its effectiveness and efficiency from the perspective of crowdsourced shared bus data set. In this section, we first describe the shared bus data set, including data preprocessing and data analysis, and then present the experimental results and effectiveness evaluation of station passenger flow prediction and route planning.

A. Data Description

The data set used in our experiments is shared subway shuttle bus data, a kind of representative mobile crowdsourced data. It is generated by shared buses in Shanghai by Panda Bus Company, which covers from April 1, 2017 to September 6, 2017. The data set includes the order data of shared buses' passengers and GPS data of shared buses. Order data contains various fields, such as order ID, city code, region code, passenger ID, order type, date and time of order creation and passengers boarding, station ID of passengers boarding and alighting, the number of passengers, order status, cash passengers flag, order cancellation flag, and so on, as shown in Table II. The GPS data contains latitude, longitude, and time. The data contain 44 817 passenger records of 8 vehicles in 10 stations in Shanghai Yongkang City.

1) Data Preprocessing: Data preprocessing contains data filtering, data cleaning, and data aggregating. We focus on the route planning from the residential area to the subway station nearby, so the records in the afternoon from subway stations to residential areas are filtered. For the operation mode of shared buses, if passengers use online payment, time and location of passengers boarding that influence the number of passengers can be obtained as soon as they get on the bus. However, if passengers pay in cash, their information will generate

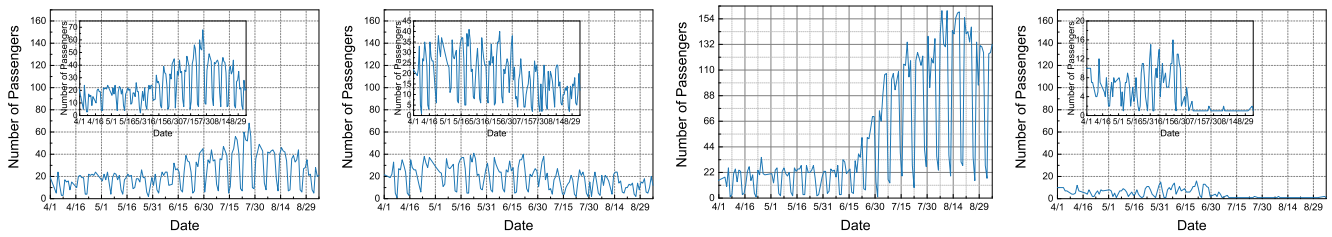


Fig. 7. Passenger flow of four stations from April 1, 2017 to September 6, 2017. (a) Danguiyuan station. (b) Xiangzhangyuan station. (c) Baomingyuan station. (d) Yongsongyuan station.

TABLE II
DESCRIPTION OF SHARED SUBWAY SHUTTLE BUS ORDER DATA

Field	Annotation
OrderID	Shared bus order ID
CityCode	City code
RegionID	Region ID
PassengerID	Shared bus passenger ID
Type	Orders type
CreateDate	Time and data of orders creating
CheckTicketDate	Time and date of passengers boarding
UpStopID	Station ID of passengers boarding
DownStopID	Station ID of passengers alighting
RideCount	the number of passengers
SmallDriverID	Shared bus deriver ID
SmallVehicleID	Shared bus ID
OverStatus	Order status
OrderFlag	Order cancellation flag
CashFlag	Cash passenger flag

after drivers confirm in the cellphone application Panda Bus. Therefore, the delay operation of drivers may generate wrong information. In order to avoid cash passengers' impact on passenger flow prediction, we screen out cash passengers records, whose order type is 3 in data. Then, we clean up errors and outliers, such as canceled orders and order record from the subway in the morning, and zeroize missing values. Finally, we aggregate the data in the time dimension to divide the study time period into small time intervals of 10 min long according to the operation law of shared buses.

2) Data Analysis: In order to better understand data characteristics to achieve better experimental results, we carry on a data analysis on the relevance of passenger flow and time. There are six operating shared buses in the study area before June 1 and the number of buses increases to 8 after June 1. As shown in Fig. 7, we list the passenger flow of several stations. It shows that the number of passengers at the Danguiyuan Station and Baomingyuan Station has a substantial increase after June 1, while the number of passengers on the Xiangzhangyuan Station and Yongsongyuan Station even decreases at the same time. Therefore, in passenger flow prediction, the factor that the number of shared buses changed on June 1 cannot be ignored, and we quantify the factor based on its impact of stations' passenger flow to generate a feature to be the model's input to optimize prediction effects, that is, the feature Bus.

We can get from Fig. 7 that traffic law of each station is different from each other. Considering of this point,

TABLE III
STATISTICAL ANALYSIS OF PASSENGER FLOW DATA

Station	Maximum	Minimum	Variance	Standard Deviation
No.1	68	2	216	15
No.2	41	1	117	11
No.3	161	1	2580	51
No.4	32	1	73	9
No.5	35	1	95	10
No.6	16	1	16	4
No.7	36	1	70	8
No.8	43	1	113	11

we build independent models for each station to do passenger flow prediction. The maximum number of passengers at the Yongsongyuan Station per day is only 16, and the maximum number of other stations is only up to 161 or so. Compared with the other traffic data, such as taxi or subway data set, the amount of this data set is relatively small [40], [41], which brings challenges for passenger flow prediction. Another bad influence of relatively small data amount is that passenger flow has great fluctuations. We can get from Table III that the variance of all stations is quite large. For instance, the number of passengers changes a lot when it rains. Station passenger flow has poor resistance to emergency, so the changing law of passenger flow gets more difficult to find. Therefore, in station passenger flow prediction, we try to optimize the prediction effects from all aspects and strive to achieve higher accuracy. The difficulty brought by data demonstrates the effectiveness of our proposed model further, which will be introduced in detail in Section V-B.

B. Passenger Flow Prediction Evaluation

In the first step of experiments, we first predict passenger flow at each station in the targeted area at different intervals. The order data cover the period from April 1, 2017 to September 6, 2017, with 158 valid days. Following the process of our approach (SubBus), we extract four important features from data based on the analysis of travel behaviors, which are flow, time, week, and location. According to the data analysis, we extract another special property: Bus, to simulate the impact of the increase in the number of buses on passenger flow at each station. In order to optimize experimental results, we do the following feature processing.

1) Feature Processing: Based on the specific data type, how to optimize features and how to maximize the use of features

Algorithm 2 Station Selection

Input: T_i: list of departure time;
Rou_{can}: candidate route set;
Ori: candidate origin set;
Tim: travel time between two stations;
$Pass$: the number of passengers in each time interval at each station;
i: bus ID, (0, 1...n-1);
s: station ID, (0, 1...m-1);
τ: the number of seats in a bus;
F: frequency of buses;
D: the destination;

Output: Rou_i: list of routes, which contain the information of stations, time, and seats;
$Rou_i.Sta[]$: list of stations;
$Rou_i.T[]$: list of current time;
$Rou_i.Sea[]$: list of seats;

1.**procedure:** select the next station of a bus
2.$Sta_{optional} \leftarrow$ \\ The set of optional next stations and the travel time from the previous station to these stations;
3.$i \leftarrow 0$\\ Initialization, No.0 bus is the first to select stations;
4.**while** $i < n$ do
5. **if** $Rou_i.Sta.count(D) < F$ do
6. **if** $Rou_i.Sta = \Phi$ do
7. **for** each $s \in Ori$ do
8. $Sta_{optional}[s].time \leftarrow 0$;
9. **end for**
11. **else if** $Rou_i.Sta[-1] = D$ do
12. **for** each $s \in Ori$ do
13. $Sta_{optional}[s].time \leftarrow T[D][s]$;
14. **end for**
15. **else do**
16. $Sta_{previous} \leftarrow Rou_i.Sta[-1]$\\ get previous stations;
17. Sta\\get all possible next stations according to Rou_{can};
18. **for each** $s \in Sta$ do
19. $Sta_{optional}[s].time \leftarrow T[Sta_{previous}][s]$;
20. **end for**
21. **end else**
22. **end if**
23. **for** each $s \in Sta_{optional}$ do
24. $T_{arr} \leftarrow Rou_i.T_s[-1] + Sta_{optional}[s].time$\\ the time arriving at station s;
25. **if** $s = D$ do
26. select D as the next station;
28. **end if**
29. $rate \leftarrow \frac{T_{arr} - T_{start}[p]}{\theta}$\\ p is the number of time interval that T_{arr} lies in, and T_{start} is the beginning of pth interval;

Algorithm 2 *(Continued.)* Station Selection

30. $Pass_{pickup} \leftarrow \sum_{k=0}^{p-1} Pass_k + Pass_p * rate$\\ $Pass_{pickup}$ is the number of passengers picked up;
31. select $Station$ with the max $Pass_{pickup}$;
32. **if** $Sea_i + Pass_{pickup}[Station] \geq \tau$ do \\ the number of passengers after picking up passengers
33. add $Station$ to $Rou_i.Sta$;
34. add $T_{arr}[Station]$ to $Rou_i.T$;
35. add $Pass_{pickup}[Station]$ to $Rou_i.Sta$;
36. update Pass;
37. $Station \leftarrow D$;
38. select $Station$ as the next station;
39. **end if**
40. **end for**
41. add $Station$ to $Rou_i.Sta$;
42. add $T_{arr}[Station]$ to $Rou_i.T$;
43. **if** $Station! = D$ do \\ D is not the next station
44. add $Pass_{pickup}[Station]$ to $Rou_i.Sea$;
45. update Pass;
46. **end if**
47. $i \leftarrow$ **Bus Slection Algorithm**;
48. **else do**
49. $i \leftarrow i + 1$;
50. **end else**
51. **end if**
52. **end while**
53.return Rou_i;
54.end procedure

determine effects of algorithms. Therefore, according to the characteristics of data and problem, we perform the following processing steps to maximize the extraction of features from raw data for the use of algorithms and models to achieve the best prediction effects.

a) Normalization: Feature normalization is the proposed solution for the great disparity in features' value range that affects the outcome of our model adversely. Normalization needs to calculate the mean and standard deviation of the feature value, as shown in the following equation:

$$x' = \frac{x - \bar{x}}{S} \tag{4}$$

where x is the value of raw features, $\bar{x}$ stands for the mean, S is the standard deviation, and x' is the value of processed features. In this way, the value of all the features is processed to the same value range.

b) Discrete features processing: Features are not always continuous values, and may be classified values, that is discrete values, just as the features week and time in this paper. Even converted into digital representation, the data are also not suitable for using directly in our model, which defaults that data are continuous and ordered. In order to solve this problem, one of the possible solutions is to use one-hot encoding, also known as one-bit valid encoding. Its method is to use N-bit status register to encode N states, and each state has a separate register bit and only one of them is valid at any time. It can be understood that for each feature, if it has m possible values, it becomes m binary features after one-hot encoding. What is more, these features are mutually exclusive and only one is active for each time. As a result, the data become sparse.

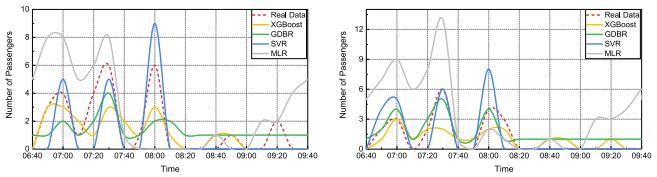
There are two main benefits of one-hot encoding: first, it solves the problem that our model is not suitable for processing discrete data, and second, it plays a role in the expansion of features to a certain degree.

c) Polynomial feature construction: As we mentioned earlier, one-hot encoding can play a role in expanding features. The purpose of polynomial feature construction is exactly to expand features. We use a basic function to construct linear fit in the higher dimension space of features. Thus, the model has the flexibility to adapt to a wider range of data. Common data transformations are polynomial-based, exponential function-based, and logarithmic function-based. This operation not only increases the number of features but also constructs the features that we may ignore in the feature extraction process. Polynomial features consider the nonlinear features of the input data to increase the model's complexity and capture the high-order and interacting terms of features. Here, we use the polynomial-based data transformation, which is named polynomial feature construction. We use a polynomial conversion formula of degree 3.

2) Parameter Tuning: As mentioned earlier in the part of data analysis, instability and too small values of fields bring great difficulties to the prediction, which does not exist in traditional traffic data prediction, such as subway data and taxi data. Therefore, we try to improve prediction effects from various aspects, including data analysis, feature selection, feature processing, and so on. Hyperparameter tuning is a way to improve the effects of models. We mainly adjusted the following parameters. The maximum depth of the tree, represented by max_depth, is used to control overfitting and its usual range is 3–10. min_child_weight is used to control overlearning, and a higher value prevents the learning relationship of the model. γ determines the minimum required loss reduction and its default is 0. subsample controls the proportion of randomly sampled for each tree. reg_alpha = 1 applies regularization to reduce the fit.

The similarity of passenger flow patterns at each station is low, so we build a prediction model for each station. Then, the parameter tuning operation is performed separately. Taking the Danguiyuan Station for example, based on the results of default parameters, we first adjust the values of max_depth and min_child_weight and finally update the max_depth value to 10. Followed by γ and subsmaple, their default values are maintained to be the best. Finally, we tune the regularization parameters. The parameters tuning process of other stations is basically the same, and there are differences in concrete values of parameters.

3) Prediction Results: The prediction results demonstrate the effectiveness of our method. Meanwhile, we use other three models to carry out contrast experiments, which are support vector machine regression (SVR) model, gradient boosting regression model (GDBR), and multiple linear regression (MLR) model. Fig. 8 shows the passenger flow change over time intervals based on real data and predicting data of four models on September 4 and 5. We can get from Fig. 8 that there are two flow peaks at 7:30 and 8:00 around in real data. For predicting data, in addition to our approach, SVR and GDBR have also predicted this trend. However, the difference



(a) (b)

Fig. 8. Prediction results of four models on two typical days. (a) September 4, 2017. (b) September 5, 2017.

between the predict data and the real data is large. There is a small gap between the real data and the predicted values of our approach. The prediction effect of MLR is quite bad in our experiment, although it shows good results in the passenger flow prediction based on subway data. On the whole, our approach can provide good predictive results.

4) Prediction Evaluation: However, from passenger flow-chart of September 5 [Fig. 10(d)], we found that the predictive effects of GDBR and SVR seem pretty good. The prediction data of one day are contingent clearly, and then, how is the prediction effect of our method on the earth? In order to solve this problem, we use several evaluation indicators: correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE) to further measure the overall predictive effect of multiple methods.

1) *CC:* According to different research objects, CC has a variety of definitions, and the one used most commonly is Pearson CC, which is used in this paper. CC can reflect the degree of correlation between variables. We define CC in this paper as the following equation:

$$\mathrm{CC} = \frac{\mathrm{Cov}(p_i, y_i)}{\sqrt{\mathrm{Var}[p_i]\mathrm{Var}[y_i]}} \tag{5}$$

where p_i and y_i denote the predicted values and real ones, respectively. $\mathrm{Cov}(p_i, y_i)$ is the covariance of p_i and y_i, $\mathrm{Var}[p_i]$ is the variance of x, and $\mathrm{Var}[y_i]$ is the variance of Y. In traffic flow prediction, CC is often used to analyze the linear correlation between the predicted data and the real data, as an important reference for accuracy.

2) *RMSE:* It is used to verify the deviation between predicted values and real values, which shows the model accuracy from a perspective of predicting deviation. In this paper, the RMSE is given by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} |p_i - y_i|^2}{n}}. \tag{6}$$

3) *MAE:* It is the average of the absolute values of the deviation between all the individual predicted values and real values. The MAE can reflect the actual situation of predict error well

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |p_i - y_i|. \tag{7}$$

We selected index data of all models at five stations to display in Fig. 9. For CC, our approach is superior to other models at most stations. At station 5th, the predictive effect of MLR

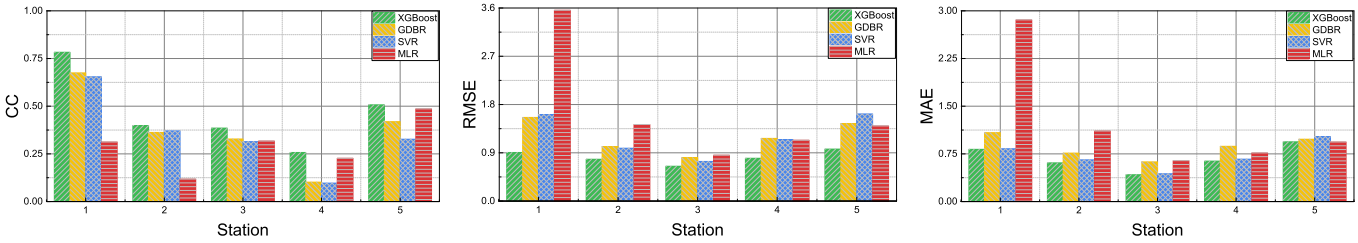


Fig. 9. Predictive metrics results of four models at five stations. (a) CC. (b) RMSE. (c) MAE.

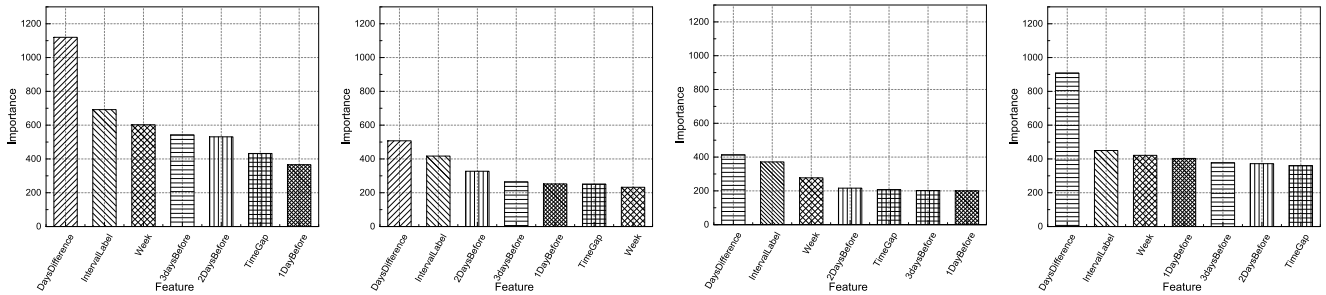


Fig. 10. Importance of each models' features. (a) Danguiyuan station. (b) Xiangzhangyuan station. (c) Pujianglu station. (d) Shenghuayuan station.

is better than our approach, while it shows very poor effect at other stations. The predict accuracy of our approach at some stations can even reach 0.8. The smaller the value of RMSE and MAE, the better the model in performance. From the point of view of RMSE and MAE, the shortcomings of MLR instability are more pronounced. Our approach is basically kept at a minimum in multiple stations. Under good predict accuracy, our approach can maintain the stability of prediction.

5) Importance of Features: After obtaining the prediction results of travel requirements, we analyze the importance of features. We choose the importance of features of several station prediction models shown in Fig. 10, where the values of the ordinate only represent the importance of features. The higher the value, the more important the feature is, and the value has no practical significance. The discrepancy in importance between features varies over models, but for all models, DifferentDays is the most important, indicating that this feature owns a strong ability to improve the feature accuracy. The number of vehicles is significant for shared bus travel requirement prediction. IntervalLabel is also an important feature, which shows the strong time dependence of travel requirements.

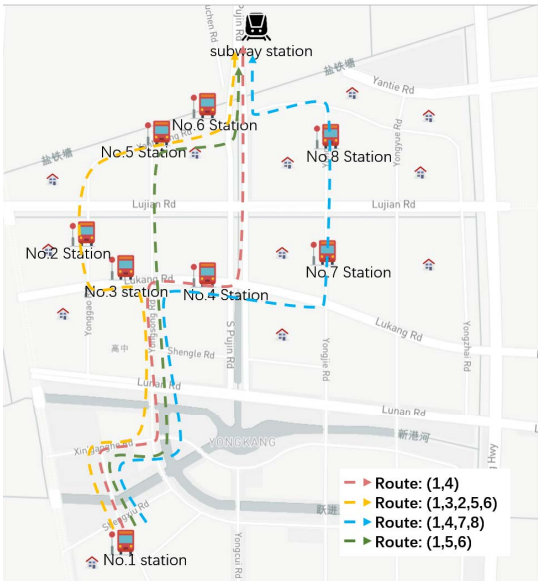


Fig. 11. Candidate route set of No. 1 station.

C. Dynamic Route Planning Evaluation

In the dynamic route planning phase, we first do candidate route selection and the candidate routes of No. 1 station are shown in Fig. 11. Then, utilizing accurate prediction results of travel requirements, we plan dynamic routes for subway shuttle shared buses. The buses scheduling results from 06:55 A.M. to 07:05 A.M. are presented in Fig. 12, and we can see that the routes change over time because the travel requirements are different at different time intervals. Because passenger flow is relatively concentrated at No. 1 station from 6:50 A.M. to 7:20 A.M., there are more vehicles departing from No. 1 station. Table IV shows the examples of route planning results. We can get the operating routes of some vehicles and time at which vehicles arriving at each station of the route. Moreover, for the same shared bus, its operating routes at different times are likely to be different, which means the dynamic route. Our algorithm chooses the optimal routes for the shared bus based on the change of passenger flow at each station. Compared with traditional

TABLE IV
ROUTE PLANNING RESULTS OF COMPARATIVE EXPERIMENTS

BusId	Frequency	Route	Time
4	1	[8, 7, 4, 3, 2, 5, 6, 9]	['07:27:57', '07:29:38', '07:32:11', '07:33:20', '07:34:04', '07:35:59', '07:36:13', '07:49:35']
1	1	[8, 7, 4, 5, 6, 9]	['07:03:00', '07:05:30', '07:08:04', '07:10:28', '07:10:37', '07:19:46']
1	2	[6, 5, 2, 3, 4, 9]	['07:28:39', '07:28:49', '07:30:32', '07:31:12', '07:31:35', '07:46:45']
1	3	[8, 7, 4, 3, 2, 5, 6, 9]	['07:57:48', '07:59:28', '08:01:52', '08:02:41', '08:03:10', '08:04:52', '08:05:09', '08:27:36']
6	4	[1, 4, 9]	['08:18:22', '08:19:36', '08:23:49', '08:24:28', '08:25:15', '08:26:55', '08:27:08', '08:42:59']

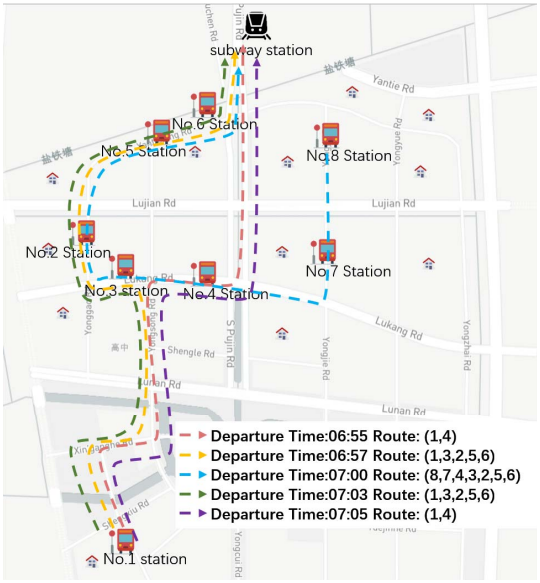


Fig. 12. Dynamic routes from 06:55 A.M. to 07:05 A.M..

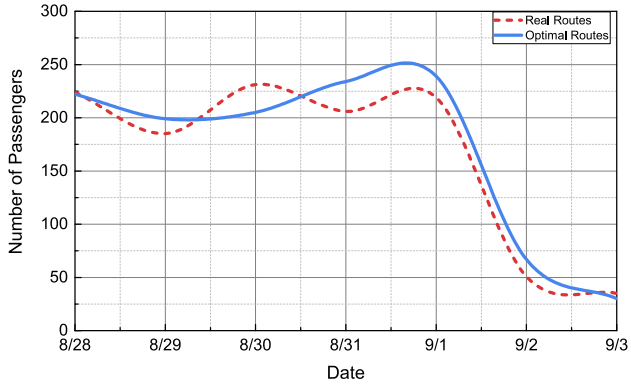


Fig. 13. Passengers' number of operating routes.

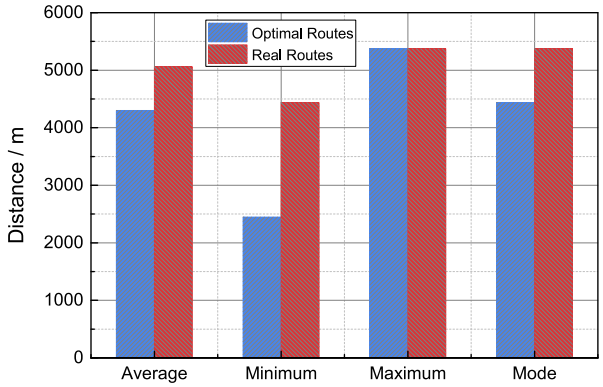


Fig. 14. Distance statistics of operating routes.

public transport route planning, the rough use of the idea is that the largest loop connects the largest number of passengers, and there exist some very short routes in our planning results, because such routes contain stations with a relatively large number of passengers and there is no need to spend extra time from the farthest station. We also analyze the results of our route planning from the perspective of operation distance and passengers' number.

From the aspect of passengers' number, the superiority of our approach is presented. In Fig. 13, we select operation passengers' number data of our optimal routes and real operating routes from August 28 to September 3, which exactly covers a week. It can be seen that for most days in the week, the passengers' number of our optimal routes is larger than the real routes. We calculate the operation distance of all planning routes and real operation routes in a day and display its average value, minimum, maximum, and mode in Fig. 14. We can see from the figure that the average distance, minimum distance, and mode distance of our planning routes are all obviously shorter than the real routes. Therefore, based on our planning results, shared buses can lower costs and improve operational efficiency by reducing operating distances.

In addition to real operation routes, our method is also superior to other dynamic route planning methods. Here, we choose a prediction-based unobstructed route planning method [42], which has the similar framework: planning dynamic routes based on prediction. In their work, crowded stations are first detected and unobstructed routes can be found. However, such an effective method is not suitable for dealing with route planning problem of shared buses. We display the part route planning results of the comparative method in Table V. From each origin, there is only a single route, and among all eight stations, all routes only cover four stations. What is worse, routes are fixed.

In our approach, we will consider the characteristics of the shared bus route planning and provide a high capacity, short operating distance, and dynamic route planning methods.

In a word, our proposed approach (SubBus) can provide effective suggestions for shared bus dynamic route planning, especially from the aspects of operating distance and passengers' number.

TABLE V
ROUTE PLANNING RESULTS OF COMPARATIVE EXPERIMENTS

BusId	Frequency	Route	Time
1	1	[1, 5, 6, 9]	['06:55:26', '06:59:48', '07:00:06', '07:02:00']
4	1	[1, 5, 6, 9]	['07:03:00', '07:07:22', '07:07:40', '07:09:34']
7	1	[1, 5, 6, 9]	['07:12:07', '07:16:29', '07:16:39', '07:20:21']
1	2	[8, 5, 6, 9]	['07:12:41', '07:15:02', '07:15:20', '07:19:11']
2	2	[8, 5, 6, 9]	['07:15:11', '07:18:44', '07:18:54', '07:20:48']
6	2	[8, 5, 6, 9]	['07:26:30', '07:30:03', '07:30:19', '07:30:19']
7	2	[6,5, 8, 9]	['07:26:07', '07:26:22', '07:29:55', '07:31:25]
4	3	[6,5, 8, 9]	['07:32:24', '07:32:38', '07:36:15', '07:37:59']
6	2	[6,5, 8, 9]	['07:38:31', '07:38:45', '07:42:22', '07:44:06']

VI. CONCLUSION

In this paper, we put forward a dynamic route planning approach named SubBus for shared subway shuttle buses based on crowdsourced mobile data, which contains station passenger flow prediction and dynamic route planning. Based on the real shared subway shuttle bus data, we carry out extensive experiments to demonstrate that our approach can generate effective operation routes to optimize the operation status of shared buses to promote their development. We perform a resident travel behavior analysis to extract multiple important features and to predict passenger flow utilizing a machine learning method. Though the data are very volatile, the predict accuracy at several stations can reach 80%. Based on the candidate origin set and candidate route set we generated, we obtain the optimal routes for shared buses by our designed dynamic programming algorithm. Experiment results show that our planning routes have shorter operation distance and more passengers than real routes. Our proposed approach (SubBus) can generate routes for shared subway shuttle buses to optimize operation status on the "last mile" issue.

REFERENCES

[1] J. Hamari, M. Sjöklint, and A. Ukkonen, "The sharing economy: Why people participate in collaborative consumption," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 9, pp. 2047–2059, 2016.

[2] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira, "Smart cities and the future Internet: Towards cooperation frameworks for open innovation," in *The Future Internet*. Berlin, Germany: Springer, 2011, pp. 431–446.

[3] X. Kong *et al.*, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3874–3886, May 2018.

[4] B. Cohen and J. Kietzmann, "Ride on! mobility business models for the sharing economy," *Org. Environ.*, vol. 27, no. 3, pp. 279–296, 2014.

[5] J. Wirtz and C. Tang, "Uber: Competing as market leader in the U.S. versus being a distant second in China," in *Services Marketing: People, Technology, Strategy*, 8th ed. Hackensack, NJ, USA: World Scientific, 2016, pp. 626–632.

[6] L. Hong, Y. Yan, M. Ouyang, H. Tian, and X. He, "Vulnerability effects of passengers' intermodal transfer distance preference and subway expansion on complementary urban public transportation systems," *Rel. Eng. Syst. Saf.*, vol. 158, pp. 58–72, Feb. 2017.

[7] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data," *World Wide Web*, vol. 21, no. 3, pp. 825–847, 2018.

[8] D. Wang, W. Cao, J. Li, and J. Ye, "DeepSD: Supply-demand prediction for online car-hailing services using deep neural networks," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, San Diego, CA, USA, Apr. 2017, pp. 243–254.

[9] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, "Short-term traffic prediction based on dynamic tensor completion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2123–2133, Aug. 2016.

[10] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[11] D. Chen, "Research on traffic flow prediction in the big data environment based on the improved RBF neural network," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2000–2008, Aug. 2017.

[12] J. Zhao and S. Sun, "High-order Gaussian process dynamical models for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2014–2019, Jul. 2016.

[13] J. Chen, K. H. Low, Y. Yao, and P. Jaillet, "Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 3, pp. 901–921, Jul. 2015.

[14] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[15] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[16] J. Zhang *et al.*, "A real-time passenger flow estimation and prediction method for urban bus transit systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3168–3178, Nov. 2017.

[17] A. Cheng, X. Jiang, Y. Li, C. Zhang, and H. Zhu, "Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method," *Phys. A, Statist. Mech. Appl.*, vol. 466, pp. 422–434, Jan. 2017.

[18] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "DNN-based prediction model for spatio-temporal data," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Burlingame, CA, USA, Oct. 2016, Art. no. 92.

[19] H.-F. Yang, T. S. Dillon, and Y.-P. P. Chen, "Optimized structure of the traffic flow forecasting model with a deep learning approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2371–2381, Oct. 2017.

[20] Q. Shang, C. Lin, Z. Yang, Q. Bing, and X. Zhou, "A hybrid short-term traffic flow prediction model based on singular spectrum analysis and kernel extreme learning machine," *PLoS ONE*, vol. 11, no. 8, p. e0161259, 2016.

[21] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.

[22] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.

[23] Z. Ning, F. Xia, N. Ullah, X. J. Kong, and X. P. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 16–55, May 2017.

[24] A.-N. Qazi, Y. Nara, K. Okubo, and H. Kubota, "Demand variations and evacuation route flexibility in short-notice bus-based evacuation planning," *IATSS Res.*, vol. 41, no. 4, pp. 147–152, 2017.

[25] W. Y. Szeto and Y. Wu, "A simultaneous bus route design and frequency setting problem for Tin Shui Wai, Hong Kong," *Eur. J. Oper. Res.*, vol. 209, no. 2, pp. 141–155, 2011.

[26] K. Supangat and Y. E. Soelistio, "Bus stops location and bus route planning using mean shift clustering and ant colony in West Jakarta," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 185, no. 1, p. 012022, 2017.

[27] N. Mathew, S. L. Smith, and S. L. Waslander, "Planning paths for package delivery in heterogeneous multirobot teams," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1298–1308, Oct. 2015.
[28] C. Chen, D. Zhang, N. Li, and Z.-H. Zhou, "B-Planner: Planning bidirectional night bus routes using large-scale taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1451–1465, Aug. 2014.
[29] F. Bastani, Y. Huang, X. Xie, and J. W. Powell, "A greener transportation mode: Flexible routes discovery from GPS trajectory data," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Chicago, IL, USA, Nov. 2011, pp. 405–408.
[30] S. Wang, W. Lin, Y. Yang, X. Xiao, and S. Zhou, "Efficient route planning on public transportation networks: A labelling approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Melbourne, VIC, Australia, May 2015, pp. 967–982.
[31] Y. Liu *et al.*, "Intelligent bus routing with heterogeneous human mobility patterns," *Knowl. Inf. Syst.*, vol. 50, no. 2, pp. 383–415, 2017.
[32] Q. Yang, Z. Gao, X. Kong, A. Rahim, J. Wang, and F. Xia, "Taxi operation optimization based on big traffic data," in *Proc. IEEE 12th Int. Conf. Ubiquitous Intell. Comput.*, Beijing, China, Aug. 2015, pp. 127–134.
[33] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013.
[34] Q. Li, Z. Zeng, T. Zhang, J. Li, and Z. Wu, "Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 1, pp. 110–119, 2011.
[35] W. Chen, F. Guo, and F. Y. Wang, "A survey of traffic data visualization," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 2970–2984, Jun. 2015.
[36] S. Chen *et al.*, "Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 270–279, Jan. 2016.
[37] W. Wu *et al.*, "TelCoVis: Visual exploration of co-occurrence in urban human mobility based on telco data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 935–944, Jan. 2016.
[38] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Sci. Rep.*, vol. 5, Mar. 2015, Art. no. 9136.
[39] J. Feng *et al.*, "DeepMove: Predicting human mobility with attentional recurrent networks," in *Proc. World Wide Web Conf.*, Lyon, France, Apr. 2018, pp. 1459–1468.
[40] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generat. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.
[41] X. Zheng *et al.*, "Big data for social transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2015.
[42] S. Shuo, G. Danhuai, L. Jiajun, and W. Ji-Rong, "Prediction-based unobstructed route planning," *Neurocomputing*, vol. 213, pp. 147–154, Nov. 2016.

Xiangjie Kong (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China.

He is currently an Associate Professor with the School of Software, Dalian University of Technology, Dalian, China. He has published over 70 scientific papers in international journals and conferences (with over 50 indexed by ISI SCIE). His research interests include intelligent transportation systems, mobile computing, and cyber-physical systems.

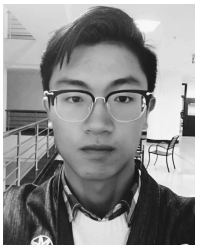
Dr. Kong is a Senior Member of CCF and a member of ACM. He has served as the workshop chair or a PC member for a number of conferences. He has served as a (Guest) Editor for several international journals.

Menglin Li received the Bachelor's degree in software engineering from the Dalian University of Technology, Dalian, China, in 2016, where she is currently pursuing the Master's degree with the Alpha Lab, School of Software.

Her research interests include big traffic data mining and analysis, human mobility behavior analysis, and smart city development.

Tao Tang is currently pursuing the Bachelor's degree in computer science and technology with Chengdu College, University of Electronic Science and Technology of China, Chengdu, China.

His research interests include big data analytics and visualization.



Kaiqi Tian is currently pursuing the Bachelor's degree in software engineering with the Dalian University of Technology, Dalian, China.

His research interests include big traffic data mining and analysis, human mobility behavior analysis, and smart city development.

Luis Moreira-Matias (M'15) received the M.Sc. degree in informatics engineering and the Ph.D. degree in computer science (major in machine learning) from the University of Porto, Porto, Portugal, in 2009 and 2015, respectively.

He is currently a Senior Researcher with NEC Laboratories Europe, Heidelberg, Germany, where he leads R&D of AI-based software for Transport, Retail, and Fintech industries. He has authored over 40 high-impact peer-reviewed publications on related topics. His interests include machine learning, data mining, and predictive analytics in general.

Dr. Moreira-Matias won an International Data Mining competition held during a Research Summer School at the Technical University of Dortmund in 2012. He served in the Program Committee and/or as an invited reviewer for multiple high-impact research venues, such as KDD, AAAI, IEEE TKDE, ESWA, ECML/PKDD, and KAIS, among others. He was invited to give keynotes around the globe, ranging locations from Brisbane (Australia) to Las Palmas (Spain). He encloses a successful track record of real-world deployment of AI-based software products across EMEA and APAC.

Feng Xia (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China.

He was a Research Fellow with the Queensland University of Technology, Brisbane, QLD, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, Dalian, China. He has published two books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, network science, data science, and mobile social networks.

Dr. Xia is a Senior Member of ACM and a Member of AAAS. He serves as the General Chair, the PC Chair, the Workshop Chair, or the Publicity Chair for a number of conferences. He is a (Guest) Editor of several international journals.