IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 27, 2020, accepted October 18, 2020, date of publication November 17, 2020,
date of current version December 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3038831

Passenger Travel Behavior in Public Transport Corridor After the Operation of Urban Rail Transit: A Random Forest Algorithm Approach

XIAOFEI LI[1], YUEER GAO[1], HUIZHEN ZHANG[2], AND YANQING LIAO[1]
[1]College of Architecture, Huaqiao University, Xiamen 361021, China
[2]College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China

Corresponding authors: Yueer Gao (gaoyueer123@gmail.com) and Huizhen Zhang (zhanghz@hqu.edu.cn)

This work was supported in part by the Promotion Program for Young and Middle-Aged Teachers in Science and Technology Research of Huaqiao University under Grant 600005-Z18X0022, in part by the National Natural Science Foundation of China under Grant 52078224, and in part by the Subsidized Project for Postgraduates' Innovative Fund in Scientific Research of Huaqiao University under Grant 18013085016.

ABSTRACT A competitive relationship has been generated between urban rail transit and bus transit since the operation of the former. Despite different roles in providing services in public transport corridor, they affect each other in actual situations. In terms of urban transportation planning and policy formulation, it is necessary to explore and master the rules of passengers' travel mode choice from different means. In order to study their choices between the rail transit and the bus transit after the operation of the former, taking Xiamen, China as an example, this article analyzed the overall travel features of passenger flow before and after the operation of rail transit by using the public transit IC card data from two consecutive weeks in November 2017 and November 2018. Some features of travel distance, travel time, travel cost, whether to travel in peak hours, the number of collinear stations between bus transit and rail transit or that of rail transit stations are sorted out. With random forest algorithm, a model is set up for the travel mode choice of passengers after urban rail transit is put into use to find out the impact of different travel features. The result shows that travel cost is the most crucial factor that affects passengers' decisions, followed by the number of collinear stations between bus transit and rail transit or that of rail transit stations, travel time and travel distance. Whether to travel in peak hours have less impact on their choices. This study is constructive for cities in the stage of facing competition between newly opened rail transit and bus transit and support transportation decision-making.

INDEX TERMS Travel mode choice, feature importance, after the operation of newly urban rail transit, random forest algorithm.

I. INTRODUCTION

Urban rail transit is built up in an increasing number of cities with the aim of alleviating the deteriorating traffic congestion with a public transport corridor operating mainly with the rail transit and supplemented with the bus transit [1], [2]. However, the newly opened rail transit in some cities not only fails to become the main mode of transportation in the public transportation corridor, but also forms a competitive relationship with the bus transit. Most passengers' trips are not transferred to rail transit which results in a small passenger flow of rail transit. Therefore, when facing the competition between the two travel modes, will passengers those who could only travel by bus transit change their choices and what factors affecting their decisions? These questions rise as necessary ones to be studied. Exploring the travel rules of passengers with their travel behaviors can lead to targeted solutions for small passenger flow in some rail transit routes caused by the competition of different travel mode in public transport corridor, which can improve its transit efficiency and promote urban development. Besides, it can also provide important reference for policy formulation and optimization of the transportation system [3].

Many scholars have studied the impact of different factors on the passengers' choices on travel modes for which some of them have even made prediction. Zhou *et al.* [4]

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

found that when choosing from shared bikes and taxis, travel distance and the number of parks and recreational facilities at the destination were important factors affecting individual passenger's decision. Cheng *et al.* [5] found that the choice of travel mode is affected by a variety of factors. When applying the random forest algorithm to simulate the traveler's choice, it was found that travel time plays the most important role in predicting travel mode choice. At the same time, the construction surroundings of the studied area and passengers' age are the two major factors affecting their choices. Omrani [6] used a variety of machine learning methods to predict the travel mode choice of individual passenger with multiple factors such as travel cost, personal factors of travelers (e.g. income, age, and gender), the number of bus stops in the residence, region of residence and type of work area. Zhao *et al.* [7] selected the individual's travel attributes (such as travel time, travel cost and waiting time, transfer frequency), socioeconomic and demographic attributes (e.g. car access, economic status, gender, and identity) as characteristics to predict the choice of travel mode and analyze travel behavior. Julian *et al.* [3] chose individual's trip attributes (e.g. travel distance), individual attributes (e.g. gender, age, and ethnicity) as characteristics to study the importance of each variable for different classifiers and travel modes. Liu *et al.* [8] chose the trip characteristics, including access and egress mode, trip purpose, travel time, combined the built environment and individual characteristics to analyze the travel characteristics and access mode choice of elderly urban rail riders in Denver.

The above-mentioned researches are mainly on passengers facing different travel modes or even includes private cars. The data is mainly from passenger questionnaires or resident travel surveys. It is subjective and with high acquisition cost, long period, and is usually limited by the small sample size and short observation period. The factors that affect passengers' travel mode choice mainly include personal characteristics such as the passenger's age, gender, occupation, etc., the characteristics of the travel process such as the passenger's travel time, travel cost, travel purpose and travel distance, and external environment characteristics such as weather conditions and built environment. In addition, passengers' travel mode choices are also subject to the socio-economic conditions of passengers' residence and related transportation policies [15]. Appropriate features can improve the detection accuracy of the model and reduce the computational complexity of the algorithm.

In recent years, the development of big data has made it easier to obtain public transportation system data. Compared with traditional survey data, public transit smart cards cannot record passengers' subjective feelings or the environment during passengers' trips. The traveler's information such as some personal attributes cannot be directly obtained, but some accurate travel characteristics of passenger such as boarding time, travel time, travel distance, travel cost, etc. can be obtained via transit smart cards data, the amount of which is much larger than the survey data. Zhao *et al.* [11] used transit smart card records in London to focus on the passenger's three travel attributes-start time, origin, and destination and proposed a new model based on the Bayesian n-gram model used in language modeling to predict passenger travel attributes-start time, origin, and destination. Gabriel *et al.* [12] use London's smart card data to obtain passenger's characteristics such as the entry and exit stations records and travel time and combine with individual travel activities to measure the regularity of passengers' travel behavior. In addition, some scholars have begun to choose other data sources such as mobile phone signaling data to study passenger travel behavior. Lu *et al.* [13] combined the mobile phone signaling data with the resident travel survey data and use random forest algorithm to set up the travel model of individual travelers, mainly using personal attributes such as individual age, gender and travel OD point and travel time characteristics. The results display that travel distance attribute is more important than the time attribute. Normally, when facing the two modes of public transport namely the rail transit and the bus transit, travel cost is an important factor that passengers take into account, which was also proved by some scholars' studies [9], [10], and travel distance and travel time are the following factors considered [4], [5], [9], [10], [12], [13]. In addition, according to the research of Zhao *et al.* [11] and Yang *et al.* [16], passengers' boarding time has also a great influence on their choices. For example, during peak hours, passengers prefer public transport and choose bus or rail transit. Combined with the data used in this study, the above-mentioned characteristics are taken as studied factors influencing passengers' travel mode choice.

In the study of passenger travel prediction methods, some scholars conducted a comparative analysis of machine learning methods and traditional methods. Zhao [7] used a variety of machine learning methods and logit models to predict individual travel. The results of the study show that the random forest algorithm has the highest prediction accuracy, and machine-learning and logit models largely agree on variable importance and the direction of influence that each in-dependent variable has on the choice outcome. Wang *et al.* [14] compared the multinomial logit model (MNL) model with Xgboost and found that the Xgboost model has higher prediction accuracy than the MNL model. Julian *et al.* [3] found that the random forest algorithm has the best performance and better effect than the commonly used MNL model when studying the choice of residents' travel modes.

Above all, few scholars pay attention to the change in passengers' travel modes when they get more choices in the public transport corridor and the study on the factors causing it is rare. The first metro line of Xiamen in China was put into use in January 2018. However, its total passenger flow in 2018 failed to reach 10% of the total of public transport (including the bus transit, BRT, and the rail transit). The average load factor is 8.38% (taking that of the morning and evening peaks on November 5, 2018 as the example), which was far beyond operational expectation. Against this back-

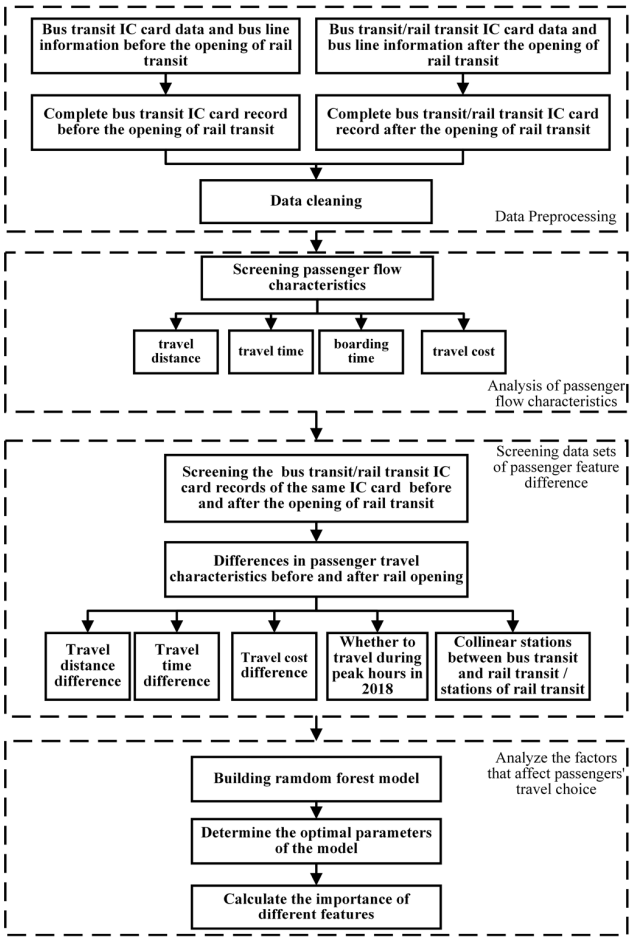

FIGURE 1. The technical route of this article.

ground, this article studies passengers' travel mode choice in the public transport corridor between the bus transit and the rail transit after the operation of the latter in Xiamen in 2018. To begin with, two-consecutive-week bus and rail transit IC card (Integrated Circuit Card) data in November 2017 and November 2018 was collected to sort out different travel characteristics of the overall passenger flow and analyze the travel behavior. And thereafter, the passenger flow of rail transit and that of bus transit before and after the operation of the rail transit are analyzed as well to screen the difference in features of passengers' travel behavior. In this article, in addition to the travel time, travel distance, travel cost, and whether to travel in peak hours mentioned earlier in this article, the number of collinear stations between bus transit and rail transit or that of rail transit stations is also used to establish a model with random forest algorithm for passengers' travel modes choice after the operation of the rail transit to analyze the importance of various factors that affect passengers' travel mode choice. The technical route of this article is shown in Fig.1.

II. DATA COLLECTION AND RESEARCH METHODS

A. DATA SOURCES

By November 2018, Xiamen has a built-up area of approximately 388.58 square kilometers and a resident population of 4.11 million, with one metro line and 359 bus lines. The data used hereby is the IC card data of the bus transit collected from November 6th to 19th, 2017 and that of both the bus transit and the rail transit from November 5th to November 18th, 2018. The reason for this is that there was no holiday in these four normal working weeks, the data is representative and suitable for comparison, and conducive for the extraction of features. IC card data of the bus transit includes IC card identification, transaction date, boarding and alighting time and stops, and latitude and longitude information; rail transit IC card data includes IC card identification, transaction date, entrance and exit time and stations, latitude and longitude information. (As shown in table 1). Based on these data, the transfer time threshold of 60 minutes is used to screen out the transfer travel records of passengers including the records of transferring from the bus to the rail and those from the rail to the bus.

B. DATA PROCESSING

1) TRAVEL FEATURES SELECTION

In November 2018, there was only Xiamen Metro Line 1 and 359 bus transit lines. In this article, when a bus stop is located within the service scope of a rail station (no farther than 800m from the station for this article), we hold that the bus stop and the rail station have a co-stop relationship, and the bus line where the bus stop is located has a collinear relationship with the rail line. Generally, the overlap degree between the bus transit and the rail transit increases as the number of collinear stations between them rises, and the competition between the bus transit line and the rail line will become stronger as well. We count the number of different bus transit lines that have a collinear relationship with the rail lines. There are 227 lines that have a collinear relationship with the rail transit. And the bus line with 2 collinear stations account for the majority. The number of bus lines with collinear stations decreases as the number of collinear stations increases, but there are still 95 bus lines with 5 or more collinear stations which means there is a strong competitive relationship between bus transit and rail transit (As shown in Fig.2. A collinear station may belong to different bus lines). Before the opening of rail transit, passengers can only travel by the bus transit (option 1). After the opening of rail transit, when passengers' origin and destination (OD) are not covered by the rail transit (option 3), they can only travel by bus or choose to transfer between bus transit and rail transit (option 2). When there are both bus stops and rail stations at the passenger's OD, the passenger can choose from them or choose to transfer between them (as shown in Fig.3). The focus of this article is the influence of different factors on passengers' choices of travel mode when they are offered bus or rail modes.

After the opening of the rail transit, passengers choose their travel mode according to the actual OD of a certain trip. They will move to the rail transit or keep choosing bus transit, or choose to transfer between bus transit and rail transit, so this article assumes that passengers make decision according to different travel features of different travel modes. We mainly study on the passengers' choices in a certain trip. When a bus

TABLE 1. Example of bus transit and rail transit IC card data.

Data field	IC card identification	Transaction date	Bus boarding time or rail transit station entrance time	Bus alighting time or rail transit station exit entrance time	Bus boarding stop or rail transit station	Latitude and longitude of the stop or station	Bus alighting stop or rail transit exit station	Latitude and longitude of the stop or station
Examples	80234121	20171106	63200	65200	xxxx 站	118.1231xxx 24.23123xxx	xxxx 站	118.1231xxx 24.23123xxx

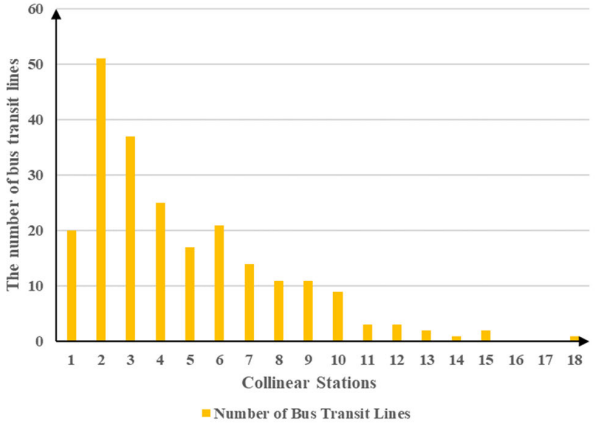

FIGURE 2. Number of bus transit lines that have collinear stations with rail transit in Xiamen City.

TABLE 2. Symbols and examples of selected features.

Selected features	Symbol	Examples
Travel time	$P_{(time)}$	45
Travel distance	$P_{(distance)}$	4
Travel cost	$P_{(cost)}$	0.8
Whether to travel during peak hours in 2018	$P_{(rush\_time)}$	1
The numbers of collinear stations between Bus transit and rail transit or that of rail Transit stations	$P_{(col\_sta)}$	3

transit route is included in a passenger's itinerary, in order to check the overlap degree of the bus transit route with the rail transit, the number of collinear stations along the chosen bus route is used to looking into the competitive relationship in space between them. When the passenger travels by rail, this feature is used to indicate the number of rail transit stations. For those who choose to transfer this feature represents the sum of the number of collinear stations and the number of rail stations. The data symbols and examples are shown in table 2.

2) ESTABLISHMENT OF DATA SETS

13,482,628 records of the bus transit IC card data in two consecutive weeks are collected from November 6th to 19th 2017. And 14,662,199 bus IC card records and 1,718,778 rail IC card records in two consecutive weeks are collected from November 5th to November 18th, 2018.

With the latitude and longitude information when passengers board and alight (or enter and exit the rail station) at a certain time and the information of bus lines and rail lines, the travel time, travel distance and whether to travel in peak hours can be sorted out. And with the latitude and longitude information of the bus and the rail transit, whether the bus the passenger chose in 2018 has collinear stations with rail transit or not can be also found out. Thereafter, the travel features of bus passenger flow before the operation of rail transit and those of both bus and rail passenger flow after the operation of rail transit can be analyzed. When constructing the data set for random forest model, the article uses the passengers' bus IC card data in 2017 as the reference to Screening the 2018's IC card records. Only the cards which are still used in 2018 for buses or the rail transit can be taken for the analysis. In 2018, there was only one rail line in Xiamen, and the rail transport network yielded to be constructed. Therefore, the accessibility of traveling by rail transit is far lower than that of by bus which has a complete transportation network. In order to study the passengers' choice against the competition between the rail transit and the bus transit, the distance between the boarding and alighting stops (entrance or exit stations) in 2017 and 2018 is controlled for purpose to ensure that the passenger's travel trajectory remains unchanged. The distance between both the boarding stops and the alighting stops of the selected IC card in 2017 and 2018 should be within 1000 meters. as shown in Fig. 4.

And in order to ensure the condition that the travel mode in November 2018 was chosen by passengers between the bus transit and the rail transit, in other words, passengers can either take the bus or the rail or transfer. Therefore, the bus stops chosen for this study should be within the 800-meter sphere of a railway station (as shown in Fig.3) according to Calvo's [10] research on the impact of rail transit stations and the actual situation in Xiamen.

Finally, this article takes the travel features sorted out from IC card records in 2018 after the operation of the rail transit to subtract the corresponding features in 2017 when there was only bus transit. Then the differences of travel time, travel distance, and travel cost between 2017 and 2018 can be obtained. Equation (1) shows how the values of the three features are calculated.

$$Feature_{value} = \sum_{\substack{0\le i\le m \\ 0<j<n}} P^{i}_{2018} - P^{j}_{2017} \tag{1}$$

In this formula, $P \in (P_{time}, P_{distance}, P_{cost})$, letter m and n represent the total number of travel records of passengers in November 2018 and November 2017, P^{j}_{2017} meaning the travel feature P of a passenger in his or her no. j trip in November 2017, likewise, P^{i}_{2018} meaning the travel feature P of a passenger in his or her no. i trip in November 2018.

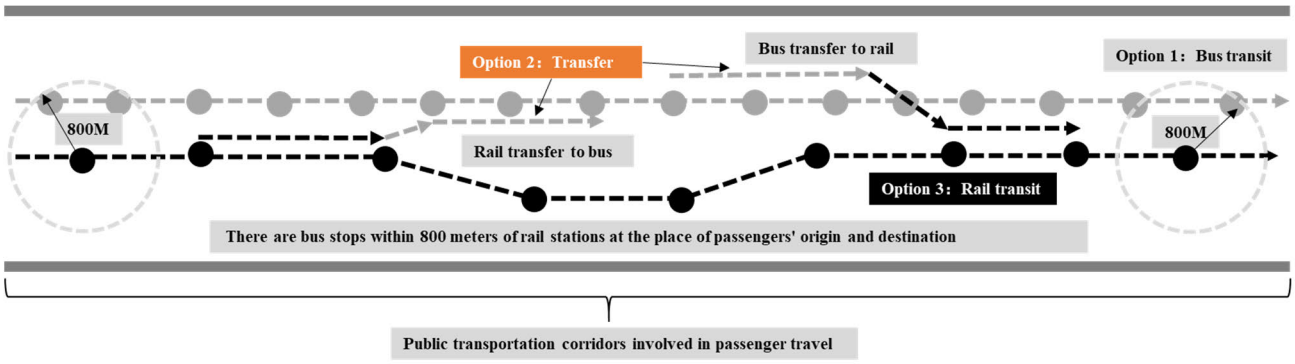

FIGURE 3. Competition for passengers between bus transit and rail transit in public transportation corridors.

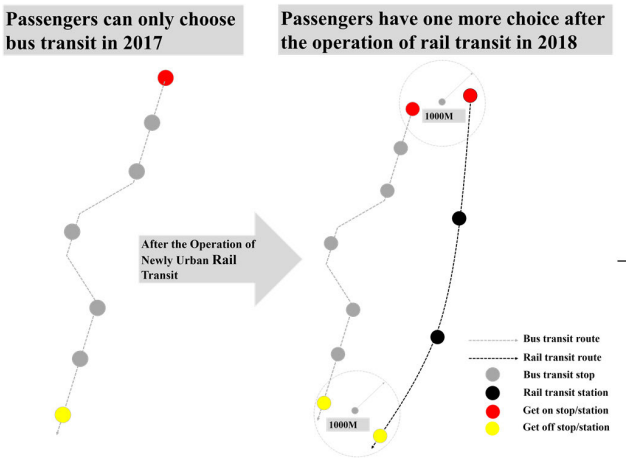

FIGURE 4. Rail transit has been added to the options available to passengers.

TABLE 3. Symbols and examples of features in the data set.

Features	Symbol	Examples	Type of the variable
Travel time difference	$time\_difference$	40	numerical
Travel distance difference	$distance\_difference$	2	numerical
Travel cost difference	$cost\_difference$	0.8	numerical
Whether to travel during peak hours in 2018	$rush\_hour$	1	categorical
The numbers of collinear stations between bus transit and rail transit or that of rail transit stations	col_sta_num	3	numerical
Choice of travel mode	tag	1	categorical

In addition to the values of the above three features, each set of records also contains the mark of whether to travel in peak hours, and the number of collinear stations between bus transit and rail transit or that of rail transit stations. At the end, each comparison group will be marked with labels, number 0 for the bus transit and number 1 for the rail transit and 2 for the transfer between them according to the travel mode in 2018. After that, the obtained data will be discretized to facilitate analysis (as shown in table 3):

In the end, a data set contains 733,734 records is left for the study, with 597,947 labeled as the bus transit and 121,904 labeled as the rail transit and 13,883 as the transfer. We can see that little data labeled as transfers. The data set is a typical unbalanced data set which means that even if all data are predicted to be label 0, the accuracy of the model is 81%, but the recall rate of labels with little data is 0%. In order to solve this problem, we adopt SMOTE [20] (Synthetic Minority Oversampling Technique) algorithm for data sampling. SMOTE is an improved scheme which based on random sampling algorithm. Random oversampling takes the strategy of simply copying samples to increase samples in minority, so it is easy to make the model overfitting. The basic idea of SMOTE is to analyze minor samples and add new samples artificially composed with minor samples into the data set. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. [20]. In this way, the quantity of the labels of minor samples is increased to equal to that of the largest sample.

C. EXPERIMENT METHODS

The study analyzes the general features of all travel records to obtain the overall behavior features of passengers. Then in order to study the impact of different travel features on passengers' choice of travel modes, a random forest algorithm is used to classify passengers' decisions according to the selected features, and then the influence degree of each feature is obtained. The random forest algorithm is an integrated learning algorithm based on the idea of Bagging [19]. It uses bootstrap to randomly fetch multiple samples from the original sample to build up each decision tree in the forest, and then vote on the prediction of each tree to get the final classification. Random forest can be used for classification with many features, especially suitable for situation with discrete features. In the process of the random forest model, the problem of high variance of a single tree classifier can be reduced by introducing randomness in the selection of samples and the node splitting in the decision tree, which can reduce and accommodate noise and outliers. At the same time, as a supervised integrated learning

classification method, the comprehensive performance indicators of random forests (such as classification accuracy, algorithm efficiency, etc.) are more friendly for high-dimensional data than other single classifiers and integrated classifiers [15].

In order to better compare the impact of various features on travel mode choice, the importance of different features in the data samples is evaluated during the establishment of the random forest algorithm model to analyze their impact on the sample classification or in other words, on passengers' choices. And this importance will be accessed via mean decrease impurity, which is a feature importance measurement method based on information gain. Generally, greater information gain of a feature leads to higher importance of it [20]. The principle of mean decrease impurity is that the more important a feature in the data set, the more constructive this feature is on the purity increase of the classification of decision trees. The increase of purity can be measured by the Gini index. Assuming that the sample has m features, namely $X_1, X_2, X_3, \ldots, X_m$, and the sample has K categories, then the formula for calculating the Gini index of the feature X_j in the decision tree is as follows:

$$GI_{X_j} = \sum_{k=1}^{K} P\left(X_j = L_k\right) * \left(1 - P\left(X_j = L_k\right)\right) \tag{2}$$

In this formula, X_j represents feature no. j, GI_{X_j} the Gini index of feature X_j, and $P(X_j = L_k)$ the estimated probability when feature X_j is include into category L_k.The importance of the feature X_j in a node c in the decision tree. i is represented by the change in the GINI index of node c before and after branching.

$$Importance^{ic}_{X_j} = GI_{X_j} - GI^{l}_{X_j} - GI^{r}_{X_j} \tag{3}$$

In this formula, $GI^{l}_{X_j}$ and $GI^{r}_{X_j}$ stand for the GINI index of the two new nodes branched from node c. Then when feature X_j appears X times in the decision tree, its importance can be obtained by:

$$Importance^{i}_{X_j} = \sum_{x=1}^{X} Importance^{icx}_{X_j} \tag{4}$$

And its importance in random forest is:

$$Importance_{X_j} = \frac{1}{n} \sum_{i=1}^{n} Importance^{i}_{X_j} \tag{5}$$

The n marks the number of decision trees in random forest.

And the research plan is as following:

(1) Obtain and process passenger IC card data.

(2) Analyze and compare the data to summarize the overall travel features of passenger flow by bus transit and by rail transit between November 2017 and November 2018.

(3) Select the data set. With the card number as a unique identifier, compare and find out the passenger's travel features in November 2017 and November 2018 according to their travel records.

(4) Confirm the optimal parameters and use the optimal parameters to build a random forest algorithm model to obtain the optimal model of it. Construct a random forest model, and determine the optimal number and depth of decision trees by searching for stable values on the out-of-bag data (OOB) error curve. Calculate the importance of each feature through the optimal model with mean decrease impurity, and analyze the rationality of the results. Finally, the accuracy score, precision score and recall score are used to evaluate the performance of classified prediction of the model for different labels. The accuracy refers to the proportion of the number of correctly classified samples divided by the total number of samples in the test set. The precision means the proportion of samples that are actually of a class divided by the total samples classified as that class in the prediction results. The recall means the proportion of samples classified as a given class divided by the actual total in that class in the test set.

III. DATA ANALYSIS AND RESEARCH RESULTS

A. TRAVEL FEATURES OF THE BUS AND THE RAIL TRANSIT BEFORE AND AFTER THE OPERATION OF THE RAIL TRANSIT

This article compares the general features of the bus passenger flow in November 2017 and November 2018 first, and pick out the passengers who had traveled on both the bus transit and the rail transit under the same OD in November 2018 to contrast their travel features. It is difficult to make a good comparison with the data of direct travel by bus or by rail since the transferred data is in small amount, so we listed them separately, as shown in Fig. 5i-l.

It is notable that the number of passengers traveling by bus increased significantly in 2018, which shows that after the opening of the rail transit, it didn't decline but rather rise. According to the data analysis of four aspects of travel distance, travel time, boarding time and travel cost, the results are shown in Fig.5. We find that there is no significant difference of the bus transit passenger flow in terms of travel distance in 2017 and 2018, with overall travel distance within 20 kilometers (km)(Using one-way ANOVA on these two groups of data, p-value was greater than 0.1). Passengers who travelled within 10 km account for over 80% among which most passengers travelled from 1 to 2 km, indicating that most passengers were doing short- and medium-travel. With the same OD, when travel distance is within 10 km, the passenger flow of the rail transit is equally distributed. When it exceeds 10 km, the distribution is less even. And the total passenger flow tends to decrease as the travel distance increases. Meanwhile, the bus passenger flow is far larger than that of the rail transit. The travel distance of most bus transit passengers is within 8 km, and the longer the travel distance, the smaller the passenger flow is. In general, most bus transit passengers travel in short and medium distance with the same OD. While for rail transit passengers, some of them travelled in short- and medium-distance and some

FIGURE 5. **The overall passenger flow characteristics.**

in long distance. And when it is over 10km, the number of passengers decreased. And it shows that most passengers did not choose the rail transit for long-distance trip even after the operation of it. In contrast, for transfer passengers, although the amount of data is small, it presents different travel characteristics. Passenger's traveling for 8 to 9 km account for the majority among those who choose to transfer, which is obviously different from those who travel by bus or by rail. And most of the cost is among 3 to 4 yuan, which is consistent with rail travel. Besides, there is also obvious morning and evening peak.

According to the comparison of travel time, it can be found that there is no visible change in bus passenger flow of 2017 and 2018, with records traveling for 7 or 8 minutes as the majority. With the same OD, the passenger flow of bus transit with travel time of 2-3 minutes is the most, and that of rail passenger flow is 19-20 minutes. On the other hand, the number of passengers who choose to transfer travel in a longer time, mostly 39 to 40 minutes. It is because firstly they spend much time in transferring and secondly, they travel in a long journey after transferring. In contrast, most of the bus passenger flow is a short-term travel, with travel time less

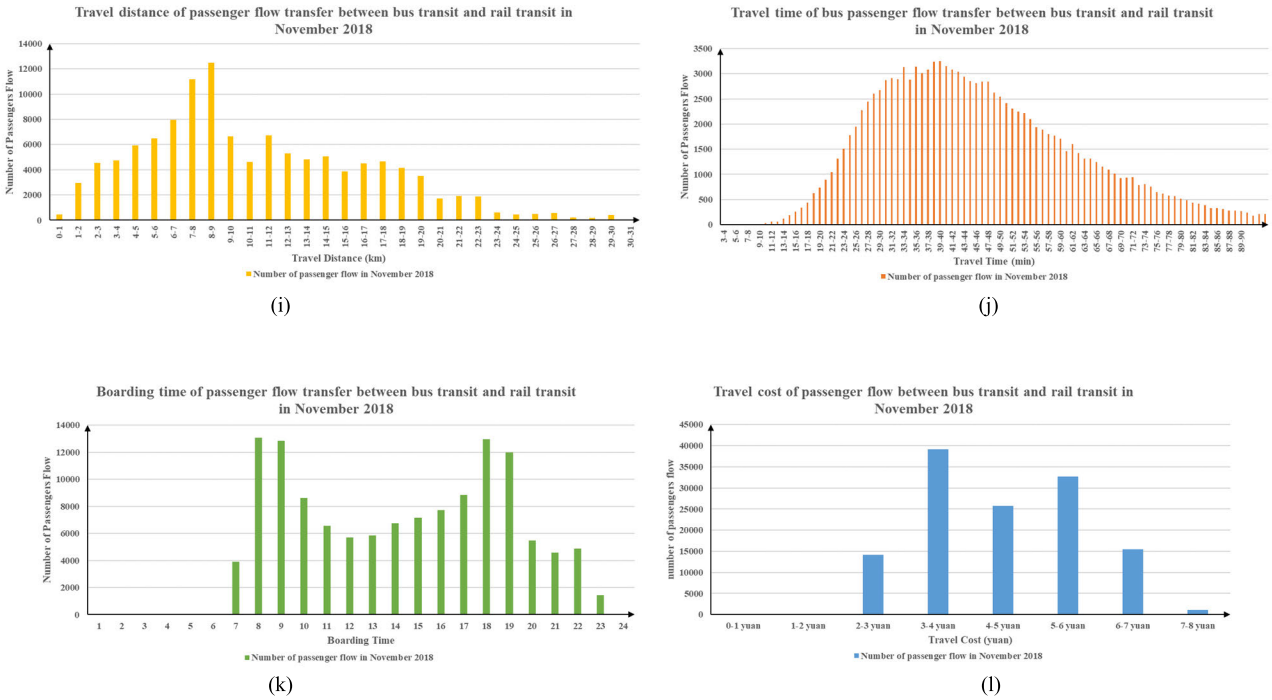

FIGURE 5. ***(Continued.)*** **The overall passenger flow characteristics.**

than 20 minutes. In conclusion, most passengers choose the bus transit for short trip.

It can be seen from the distribution of passenger flow at different boarding time that there is an obvious morning and evening peaks in both 2017's bus transit and 2018's bus and rail transit or transfer with the same OD. The morning peak is from 8:00 am to 9:00 am, and the evening peak is from 6:00 pm to 7:00 pm. And for the travel cost, most of passengers decide to spend 0.8 yuan to travel by the bus transit. Among rail transit passengers, the people spending 3 to 4 yuan account for the majority while those spending 7 yuan the minority which is almost the same as the cost of passengers choose to transfer.

Through the above analysis, it can be concluded that with or without the rail transit, the overall features of the bus passenger flow remain the same. In other words, the operation of the rail transit has no obvious effect on the change of passenger travel features. There are obvious morning and evening peaks in the rail transit or transfer between bus and rail, but their passengers are less compared with bus transit.

B. ANALYSIS OF INFLUENCING FEATURES FOR PASSENGERS' CHOICES AFTER THE OPERATION OF RAIL TRANSIT

To begin with, the data set is divided into two data sets. 70% of it is made as the training set for the training of random forest algorithm prediction model and 30% of it is the test set for verifying the prediction accuracy of the model. The two most important parameters in a random forest are the number and the depth of the decision trees. More decision trees and deeper depth lead to higher prediction accuracy, but they can also result in over-fitting of the model and increase of program runtime. Therefore, it is necessary to control reasonable number and depth of decision trees. At the same time, in order to prevent the model from overfitting, the prediction accuracy of the test set and OOB are taken as reference to determine the optimal parameters for the model.

With the increase of the number of decision trees at different depths, the prediction accuracy of the test set and OOB have a gradual increase (see Fig.6). We obtained, with Grid Search, the parameters of the random forest when the prediction accuracy peaks, with 30 decision trees and the depth of the tree being 13. (We use random forest classifier package in sklearn to build the model in Python. In grid search, the search range of the estimators is 10-100, and the depth range is 3-14.)

Under this situation, the prediction accuracy of OOB and the test set reach 95.9%. The OOB error, overall prediction accuracy of the final model and the accuracy and recall rates of each category are shown in the table 4 below.

When the model is established, the importance of each feature in the model can be obtained. As shown in Fig.7, the sum of the importance of the five features is 100%, with the difference in passenger travel cost accounting for the highest proportion, 65.71%. Compared with it, other factors have little effect on passengers' choices. The weight of the travel time is 12.77%, and that of the number of collinear stations between bus transit and rail transit or that of rail transit stations is 11.36%.

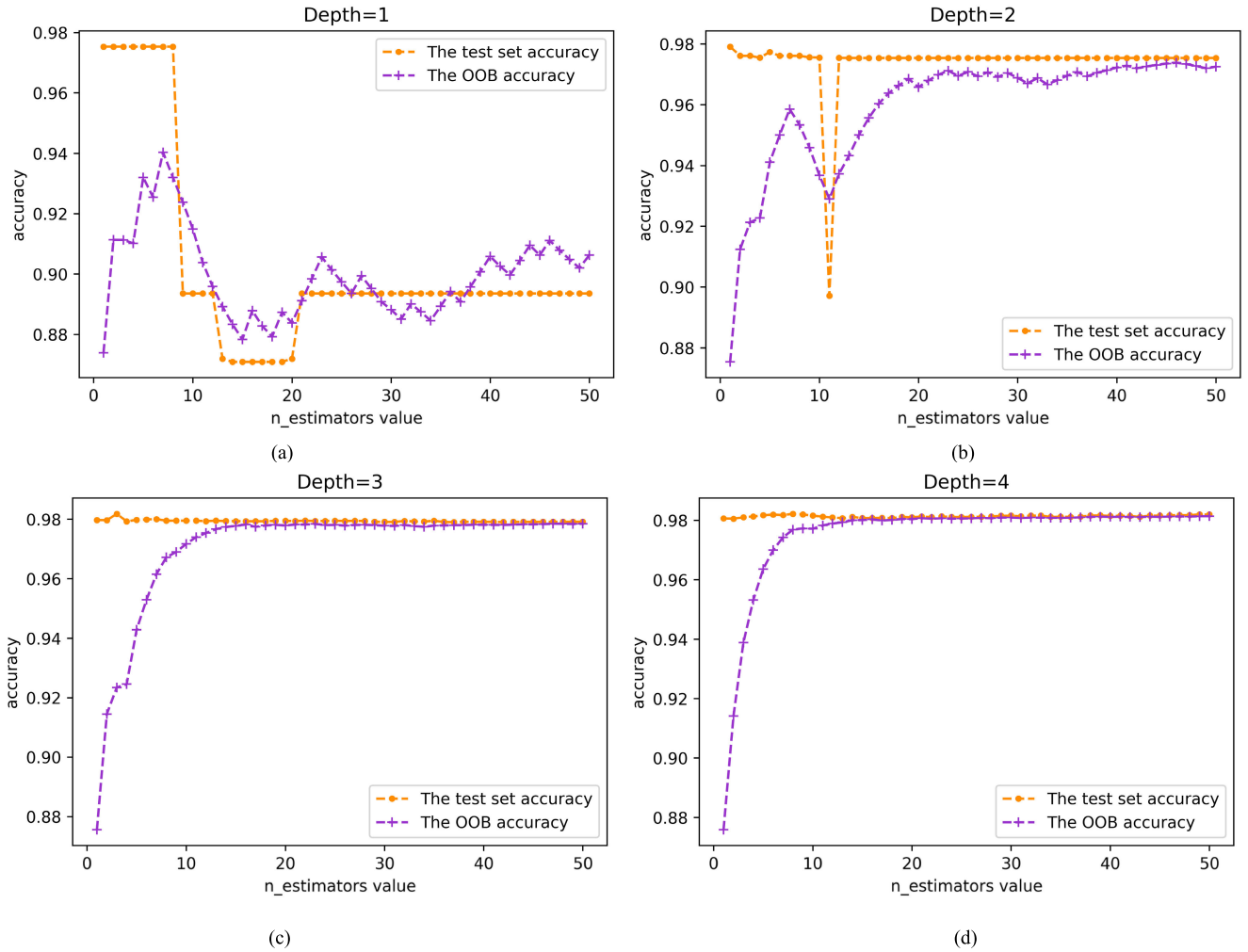

FIGURE 6. Influence of the number of decision trees in random forest on model accuracy at different depths.

TABLE 4. The OOB score, prediction accuracy and the accuracy and recall.

Features	Bus transit	Rail transit	Transfer between bus and rail
OOB score		0.959	
Accuracy		0.959	
Precision	0.999	0.949	0.930
Recall	0.993	0.927	0.957

The travel distance, and whether to travel in peak hours account for 9.72% and 0.45%, respectively. It shows travel cost is the most important factors considered by passengers after the operation of the rail transit. And this is in line with the situation when passengers choose the travel mode at the beginning of the operation of the rail transit, the bus cost in Xiamen was from 0.8 to 1.6 yuan, while the price of rail transit is from 1.8 yuan (10% off from 2 yuan) to 6.3 yuan (10% off from 7 yuan). With the same OD, the travel cost of the rail transit is much higher than that of the bus transit. The impact of the other four features are small on passengers' choices especially the effect of whether to traveling during peak hours. This shows that after the operation of the rail transit, its qualities of high speed, punctuality and comfort does not generate strong competition advantage against the bus transit which is still the primary travel mode for passengers.

As a matter of fact, in most Chinese cities, bus transit and rail transit are usually managed by different companies, and enjoy different subsidy policies from the government, which lead to big difference between their operating cost. And the price of the rail transit is much higher than that of the bus transit. Therefore, we decide to remove the travel cost of bus transit and the rail transit from the experiment in order to further examine the impact of travel time, the number of collinear stations between bus transit and rail transit or that of rail transit stations, travel distance, and whether to travel in peak hours on passengers' travel mode choice. We put aside the travel cost to set up a new model, and conduct iterative calculation on the optimal parameters. Likewise, with Grid Search, when the depth reaches 11 and the number of decision trees 80, the prediction accuracy of OOB and the test set

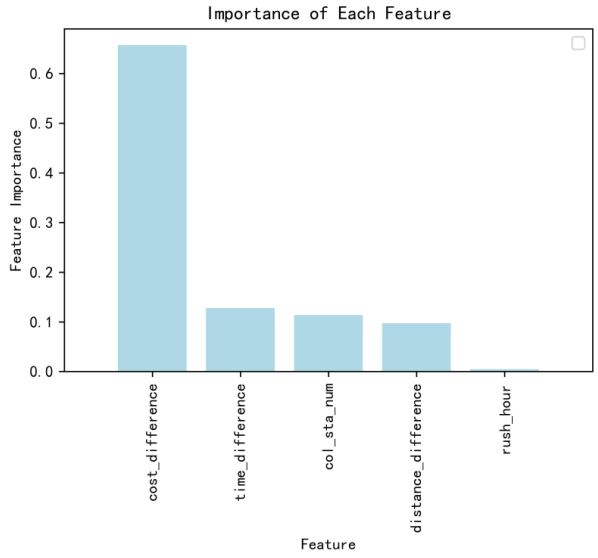

FIGURE 7. Different feature importance.

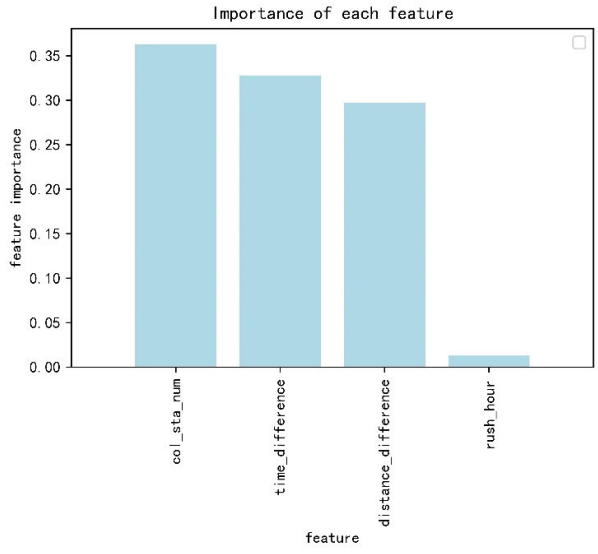

FIGURE 8. Different feature importance without considering the travel cost.

stay stable at 86.6%. The OOB score and overall prediction accuracy of the final model and the precision and recall score of each category are shown in the table 5.

Then it is obtained that the importance of each feature without consideration of travel cost (as shown in Fig.8).We can find that in 2018 the impact of each factor stays the same with or without travel cost, and the most influential factor is the number of collinear stations or that of rail transit stations (accounting for 36.24%) and the travel time accounting for 32.77%. The impact of whether to travel in rush hours stays minimal at 1.28%. This means that when exclude the travel cost in a certain trip, the most important factor that affects a passenger's decision is the number of collinear stations or that of rail transit stations in a certain trip. The travel time and travel distance have kept relatively small influence on passenger travel compared to the number of collinear stations or that of rail transit stations.

TABLE 5. The OOB score, prediction accuracy and the accuracy and recall.

Features	Bus transit	Rail transit	Transfer between bus and rail
OOB score		0.755	
Accuracy		0.755	
Precision	0.746	0.648	0.876
Recall	0.771	0.655	0.838

C. DISCUSSIONS

Combined the importance of each feature with travel cost with the analysis of the overall features of passenger flow, we observed that under the same OD, most passengers choose the bus transit when they are given both the bus and the rail transit mode, which is different from Li's [9] research that as the travel cost increasing, the optimal travel mode is rail transit. The main reason for the difference is that the travel cost by bus is much lower than that by rail transit in this research. Even the highest bus fare is lower than the lowest rail transit fare. Moreover, according to the analysis above, most passengers in Xiamen travel in short and medium trip, it can be found that the competition between rail transit and bus transit is fiercest in the short and medium distance. The travel habits of passengers and the higher ticket price of the rail transit obstruct the rail to attract passenger flow along the public transport corridor. This is a question worthy of attention, because most cities in China face a situation where rail prices are often much higher than that of the bus. In order to make the rail transit function as the major transport mode in the public transport system, the negative impact from the relatively higher ticket prices have to be offset by the convenience and accessibility of the newly-opened rail transit.

As for travel time and travel distance, due to the different bus fare policies in different cities, in this study, when passengers mainly consider the cost in traveling, the impact of travel time and travel distance is relatively small. In order to further examine the impact of other features on passenger travel behavior, we removed the feature of travel cost in the data set and carried out a new experiment. When the travel cost is not taken into consideration, the conclusion from the research of Lu *et al.* [13] shows that the distance attribute is more important than the time attribute. On the contrary, Cheng *et al.* [5] found that travel time played the most important role in predicting travel mode choice but travel distance was not taken into account in his research. However, in this article, when regardless of the travel cost, we found that the biggest factor affecting passenger travel mode choice is the number of collinear stations between bus transit and rail transit or that of rail transit stations. When the bus route is collinear with the rail transit, the more collinear stations there

are, the closer the spatial relationship between the bus route and the rail transit is. And under this situation, the rail transit and the bus transit show outstanding advantages against each other, with higher speed and punctuality of the former and more stops of the latter. This provides guidance for the optimization of the bus lines in the public transport corridor after the operation of the rail transit. In other words, the space direction of the bus lines that have longer overlap with the rail lines should be adjusted in order to mitigate the competition between the rail and the bus to improve the utilization rate of the public transportation resources.

Also, the result shows that whether to travel in peak hours has little effect on passengers' travel mode choice. This is inconsistent with the research results of Zhao *et al.* [11] and Yang *et al.* [16]. The reasons are that Zhao focused on passengers' a chain of trips and Yang on the choice between public transport and private cars. Because the three passenger flows all show obvious morning and evening peak in the same time, namely 8 am and 6 pm. Considering that its importance only accounts for 1.28%, it can be concluded that whether boarding in peak hours has little impact on passengers' choice of the bus or the rail or transferring between them.

IV. CONCLUSION

In order to study the competitive relationship between the rail transit and the bus transit in public transport corridor after the operation of the former, we use a large volume of IC card data gathered and dealt with via standardized methods to analyze the travel features of overall passenger flow. And then we select the records with the same OD to study passengers' travel mode choice and the affecting features when facing the two travel modes. The random forest algorithm is used to build a passenger travel mode choice model to predict their choice and calculate the importance of different travel features. The prediction accuracy of OOB and the test set reach 95.9% when the travel cost is taken into consideration without which 75.5%. The main conclusions are as follows: (1) the overall travel features have no obvious change after the operation of the rail transit, and the passenger flow of the bus transit has even seen a small increase so the rail transit is less attractive for passengers. With the same OD, travel cost is the most important factor taken into consideration when compared with the other features. When regardless of the travel cost, the overlap degree between bus routes and rail route, travel time and travel distance all affect the travel choice of passengers. And whether to travel in peak hours has the least influence on their choices (2) The random forest algorithm can be used to analyze the importance of passenger travel features after the operation of the rail transit. (3) The research findings can provide a theoretical reference for the formulation of transportation planning and management. For the purpose of increasing of rail transit passenger flow and optimize the allocation of transportation resources, the government can offer preferential treatment for rail transit passengers, and optimize the bus routes which are collinear with the rail.

In addition, there are three limitations in this study. First, due to the limited features that can be extracted from IC card data, this study only considers five features of travel time, travel distance, travel cost, the number of collinear stations between bus transit and rail transit or that of rail transit stations and whether to travel in peak hours. More factors such as socioeconomic characteristics can be introduced into such study in the future. Second, the result of this article may subject to the fact that the rail transit is newly opened so that passengers' travel behavior could change as the rail transit shapes a rail transit network. Third, this article does not study the feasibility of applying the trained random forest model and its parameters to other cities.

REFERENCES

[1] L. Fengjun, *Issues on Urban Express Rail Lines Planning and Construction*. Guangzhou, China: Urban Transport of China, 2020, pp. 9–11.

[2] D. Di and Y. Dongyuan, "Dynamic passenger flow analysis model in urban public transportation corridor," *J. Tongji Univ.*, vol. 42, no. 10, pp. 1523–1529, 2014.

[3] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Syst. Appl.*, vol. 78, pp. 273–282, Dec. 2017.

[4] X. Zhou, M. Wang, and D. Li, "Bike-sharing or taxi? Modeling the choices of travel mode in chicago using machine learning," *J. Transp. Geography*, vol. 79, Jul. 2019, Art. no. 102479.

[5] L. Cheng *et al.*, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behav. Soc.*, vol. 14, pp. 1–10, 2019.

[6] H. Omrani, "Predicting travel mode of individuals by machine learning," *Transp. Res. Procedia*, vol. 10, pp. 840–849, Dec. 2015.

[7] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," *Travel Behaviour Soc.*, vol. 20, pp. 22–35, Jul. 2020.

[8] C. Liu, E. Bardaka, R. Palakurthy, and L.-W. Tung, "Analysis of travel characteristics and access mode choice of elderly urban rail riders in denver, colorado," *Travel Behaviour Soc.*, vol. 19, pp. 194–206, Apr. 2020.

[9] L. Jianqiang, "Selection of optimal travel mode of public transport under competitive conditio," *J. Highway Transp. Res. Develop.*, vol. 36, no. 10, pp. 121–127, 2019.

[10] F. Calvo, O. J. de, and F. Arán, "Impact of the Madrid subway on population settlement and land use," *Land Use Policy*, vol. 31, pp. 627–639, Dec. 2013.

[11] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 19–34, Apr. 2018.

[12] G. G. langlois, N. Haris koutsopoulos, Z. zhao, and J. Zhao, "Measuring regularity of individual travel patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1583–1592, May 2018.

[13] Z. Lu, Z. Long, J. Xia, and C. An, "A random forest model for travel mode identification based on mobile phone signaling data," *Sustainability*, vol. 11, no. 21, p. 5950, Oct. 2019.

[14] F. Wang and C. L. Ross, "Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 47, pp. 35–45, Dec. 2018.

[15] W. W. jing, J. Peng, J. H. fei, and Z. M. hang, "Low carbon travel intention data mining for residents based on K-means clustering and random forest algorithm," *J. South China Univ. Technol.*, vol. 47, no. 7, pp. 105–111, 2019.

[16] B. Wang, L. Gao, and Z. Juan, "Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1547–1558, May 2018.

[17] Y. Liya and L. Juan, "Cross-nested logit model for the joint choice of residential location, travel mode, and departure time," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 53, no. 4, pp. 722–730, 2017.

[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[19] W. Yanhua, T. Shengfeng, H. Houkua, and L. Niandong, "Support vector machine based on weightiness of sample attribute," *J. Beijing Jiaotong Univ.*, vol. 135, no. 5, pp. 87–90, 2007.
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

XIAOFEI LI received the B.E. degree in Internet of Things engineering from the Shenyang Institute of Engineering, China, in 2017. He is currently a Post-Graduate Researcher with the College of Architecture, Huaqiao University, China. His main research interests include traffic big data mining and machine learning.

YUEER GAO received the Ph.D. degree in transportation planning and management from Tongji University, China. She is currently an Associate Professor and a Master's Supervisor with Huaqiao University, China. Her research interests include urban land use and transportation planning and big data analysis.

HUIZHEN ZHANG received the Ph.D. degree in computer systems organization from the University of Science and Technology of China in 2010. He is currently an Associate Professor and a Master's Supervisor with Huaqiao University, China. His research interests include software and hardware system optimization and traffic big data processing.

YANQING LIAO received the B.S. degree in geographical science from Guangzhou University, China, in 2018. She is currently a Post-Graduate Researcher with the College of Architecture, Huaqiao University, China. Her research focuses on transportation and land use and big data analysis.