


# Travel Mode Detection Using GPS Data and Socioeconomic Attributes Based on a Random Forest Classifier

Bao Wang, Linjie Gao, and Zhicai Juan 

**Abstract**—The past few years have witnessed the rapid growth in the collection of large-scale GPS data via smartphone-based travel surveys around the world, following which transportation modes detection received significant attention. A mass of methods varying from Criteria-based rules to Machine Learning technology were employed to recognize the travel modes. However, the limited sample size, deficient feature selection and the less emphasis on addressing confusion modes, which leave room for improvement. This paper therefore sought to develop and evaluate a Random Forest classifier combined with a rule-based method to detect six travel modes (subway, walking, bicycle, e-bike, bus and car). Seven GPS-related variables are selected as feature set from the initial list of 22 variables. Consequently, more than 98% subway trips were correctly identified and the overall accuracy of the rest five modes classification is obtained as high as 93.11%. More than 85% trips were successfully identified for each mode except for the bus. More importantly, results show that socioeconomic attributes data could significantly improve the prediction of e-bike and address the confusion between bus and car modes. The employment of ROC curve provides a statistical proof to the excellent classification capacity of Random Forest in this study. Besides, the comparison with two representative classifiers demonstrates the applicability of Random Forest classifier for travel modes detection incorporating multi-source attributes.

**Index Terms**—GPS data, socioeconomic attributes, rule-based method, random forest classifier, feature selection, ROC curve.

## I. INTRODUCTION

TRAVEL surveys are fundamentally essential for transportation researchers, engineers and governors to study human behaviors, as well as plan, design and manage the transportation system. Travel surveys collect trip data such as origin, destination, travel mode, purpose, distance, duration, socioeconomic and demographic data for analysis [1]. These collected data were used to enhance our understanding of travel in relation to choice, location and scheduling of daily

activities, which enables us to enhance our travel forecasting methods and improve our ability to predict changes in daily travel patterns [2]. Nevertheless, these travel surveys also play an important role in activity-based modeling with day-to-day variability of travel behavior.

Conventional methods for travel survey collection have been experienced multiple stages like paper-and-pencil interviews (PAPI), computer-assisted telephone interviews (CATI) and/or computer-assisted self-interviews(CASI). However, these methods rely on respondents completing the travel log at the end of the day or the entire survey period. Some drawbacks, including the travel time overestimating [3], trip underreporting [4], surrogate reporting and/or confusion of trip purpose [5], may decrease the quality of collected data and dispirit the respondents. Since the late 1990s, technologies such as Global Positioning System (GPS) devices have been utilized as a supplement to traditional travel surveys. GPS data collection has become an important mean to investigating travel behavior since such data ideally provides far more detailed information on travel choice and travel patterns over a longer time period than those could be obtained from traditional travel survey methods [6]. The increased sensing capabilities of smartphones combined with their extensive market penetration, easy programmability, and effective distribution channels for third party applications, have contributed to smartphones maturing into an effective tool for unobtrusive monitoring of travel behavior [7].

Despite the accuracy and usefulness of smartphones embedded with GPS and other sensors, there are still challenges to overcome. Considering the complexity and size of collected data by these applications, more efficient and advanced algorithms are required to infer travel information. Previous researches apply particular rules to infer trip ends and employ GIS sources to detect trip purposes [5], [8], [9]. In contrast, travel mode is inferred with either particular rules or special classifier [10]–[12]. Some studies detect travel modes with the help of GIS sources [13]. Although these studies obtain a high overall accuracy, they could not provide reasonable solution to distinguish some confusing travel modes. Besides, the sample size they used is fairly small, thus the lack in characteristic diversity of sample makes it difficult to detect distinct travel modes.

In the following section, we will summarize existing studies related to travel mode detection and concentrate on limitations and main challenges. Data collection and preparation are

Manuscript received December 16, 2016; revised May 4, 2017; accepted June 27, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 51478266 and in part by the Fundamental Research Funds for the Central Universities under Grant 16JCCS24. The Associate Editor for this paper was J. Miller. (Corresponding author: Linjie Gao.)

B. Wang and L. Gao are with the School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ljgao@sjtu.edu.cn).

Z. Juan is with the Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200240, China.

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/TITS.2017.2723523

described in section 3. Then, the methodology, including the feature selection and algorithm development is discussed. The results and the main findings are presented in section 5. Finally, some conclusions and the implications of this research are provided.

## II. RELATED WORK

This section lists the main methods and data sources used in existing studies. Then we will draw some limitations in some aspects and challenges for further improving mode detection accuracy. Lastly, our proposed two-stage method will be briefly described.

### A. Data Source and Methods Used for Travel Modes Recognition

Transportation mode detection can be considered as a significant step in activity recognition. The two main tasks of travel mode detection are: (i) determining whether the participant is moving; and (ii) what kind of transportation tools he/she uses [14].

Various data sources have been input for transportation mode detection in existing researches. An early attempt by Patterson *et al.* [15] used a Bayesian model with Expectation Maximization method, GIS information and GPS data being employed as input variables, which achieved an overall accuracy of 84%. Another study by Tsui and Shalaby [16] used Fuzzy Logic model algorithm with and without GIS data, while the overall rate for accurate mode detection remained the same as 91%. Results show that using GIS could improve bus detection rate from 76% to 80% by taking advantage of the bus route information. In the USA, Gonzalez *et al.* [1] differentiated walk, car and bus through the GPS data collected by deploying the TRAC-IT mobile application to Java ME-enabled cell phones. Two sets of data base, namely all GPS points and critical GPS points, were used in the research and the input variable varies depending on the data set. For critical points case, the detection accurate rate achieved 91.2%, which is 2.6% higher than that of the all points case. Bohte and Maat [9] employed Criteria-based and Fuzzy Logic approach in a mode detection effort in the Netherlands. The overall detection rate is 70%, with 72% for bicycle, 75% for car mode, while 35% for rail and bus mode was not included.

Recently, a Bayesian Belief Network was applied by Feng and Timmermans [11] to compare mode detection accuracy using three data combinations: GPS only, accelerometer only, combined GPS and accelerometer. Results indicate that the method of combining accelerometer and GPS data outperforms the other two ones. Assemi *et al.* [17] developed and evaluated three statistic models to classify four prevalent travel modes in New Zealand. After the model validation, the nested logit model was selected for the final model to detect travel modes. Of the initial 31 variables, eight independent variables were statistically significant in mode detection by using the advanced categorically Lasso method with bootstrapping.

### B. Limitations and Main Challenges

Over the past decade, a plethora of studies has made efforts to detect transportation modes using various data collected

TABLE I  
A SUMMARY OF TRANSPORTATION MODE DETECTION STUDIES

Study	Sample size	No. of modes	Accuracy
Liao et al. [23]	4 participants	3	90%
Gonzalez et al. [1]	114 trips	3	All points:88.6% Critical points:91.23%
Zheng et al. [24]	65 participants	4	76%
Gong et al. [13]	49 participants	5	82.6%
Zhou et al. [12]	12 participants	4	93.8%

by sensors embedded in smartphones. These data sources mainly comprise of GPS/GIS data and accelerometer data. Correspondingly, the detection/classification methodologies can be subdivided into three main categories: (i) Machine Learning technology (a branch of Artificial Intelligence), such as Bayesian Network [18], Support Vector Machine [10], [19], Multi-Layer Perceptron Neural Network [20], and Conditional Random Field [18]. (ii) Probability method, such as Fuzzy logic rules [16] and probability matrix [21]. (iii) Criteria-based method [13], [22].

Although many studies have achieved relatively high detection rates (see Table I), there are some limitations with either sample size or feature selection for mode classification, which are summarized as below:

Sample size used for mode detection is too small to represent the mode diversity. It is easy to train several participants' mode characteristic and build a classification model, thus the detection rate of test set is bound to be high.

High priority should be given to feature selection, which is an important procedure before applying advanced classification to divide transportation mode. Proper features could significantly enhance the detection accuracy and decrease the complexity of the algorithm computing. Most existing studies detect travel modes straightly based on pre-defined or empirical features. However, some redundant features may be unnecessary and thus reduce the prediction accuracy, so it is possible obtaining a high detection accuracy by using more refined variables.

Another issue to be addressed is the limited number of the predicted transportation modes. Considering the various traffic configuration over the world, some researchers detect prevalent local travel modes without some unpopular means. Besides, there is an urgent need for further classification strategies used to detect confusing means (e.g. bus and car).

Given these challenges, we launched a large-scale smartphone-based travel survey and sought to develop and evaluate a Random Forest classifier combined with a rule-based method to detect six travel modes. An attribute selection procedure were conducted to select the optimal feature set before mode classifying. Besides, socioeconomic attributes were used to further distinguish confusing modes.

### III. DATA DESCRIPTION

Abundant data should be collected to develop the transportation mode detection algorithm. Considering the sample size and model validation, the data should contain various travel modes and different traffic conditions, as the input variables used in mode classification are highly affected by traffic flows [25].

To meet these requirements, we developed a smartphone-based travel survey launched in Shanghai from mid-October 2013 to late-April 2015. An application was developed for collecting location-based data by our research group. Android and IOS were selected as platforms on account of high marketing penetration in China. The application could record time, longitude, latitude, altitude, heading and the number of satellite in-view every second. To avoid battery drainage, we present each respondent with an external battery package. Also, the application will automatically closed when the smartphone keeps stationary for more than five minutes. It will restart when the smartphone moves again, which could effectively reduce the battery consumption with no adverse effects on normal data recording. After the survey, each respondent would be provided with a mobile recharge card valued at 50 RMB, which in turn attracts more respondents to participate.

Some respondents were recruited by Internet, while others were invited by social networks of our group members. When the positioning application was installed and launched on the respondents' smartphones, unique user ID was assigned. Additionally, respondents were required to complete their socio-demographic attributes online. During the survey, respondents were required to start the application before leaving home and upload GPS records after the last arrival home every day. After uploading the GPS data streams to our server, travel information, including trip ends, travel modes and trip purposes, were derived and displayed on the map. Then the respondents would be called by our group members to validate and correct the travel information if necessary. This intervention aims to help the respondents recalling more details of their trips, which can improve the accuracy of the actual travel information to a maximum extent. A total of 312 respondents were required to complete at least five days survey. According to the completeness and validity, 841 person-day GPS records from 125 participants were collected at last. More details about data collection and trip detection could be referred in our previous study [26].

Subway, walking, bicycle, e-bike, car and bus are considered in our study. After data cleaning and preprocessing, all GPS records are split into single-mode segments according to reported travel information. Every reported travel mode is assigned to the corresponding GPS segment. In total, 2740 single-mode segments are extracted from GPS recordings. As shown in figure 1, the walking share is significantly higher than other modes, probably because walking trip usually plays a transfer role. For instance, one may walk to a bus station after taking off a bicycle when bus and bicycle are combined for a trip. Car mode takes the second largest proportion among all travel modes, i.e., 20% of the total trips. This result is due to the high car ownership in

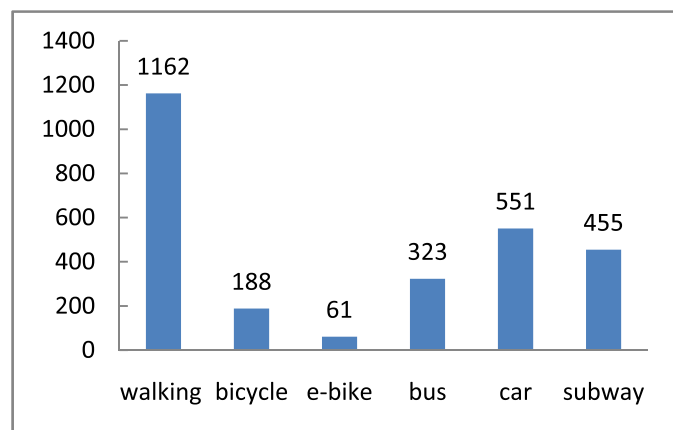


Fig. 1. The number of segments with six reported travel modes.

Shanghai metropolis. The fact that Chinese people prefer driving to working place or friends' home for convenience despite of the serious traffic jam, also contributes to the widely use of cars. On the other hand, bus and subway, which belong to public transportation, undertake 28.4% of the total travel modes and the subway share is higher than that of bus. Actually, subway is competitive for its high travel time reliability and well-connected network, especially in Shanghai. It is worth mentioning that there are 61 e-bike trips due to its convenience and a low cost. Travel modes other than subway will be detected by a Random Forest classifier with a necessary feature selection procedure and the subway trips will be identified with particular rules in the next section.

### IV. METHODOLOGY

In this study, we employ particular rules to detect subway trips based on its significant difference from the other travel modes. A Random Forest classifier was applied to divide the rest five means. Before employing classifier, we utilized several feature selection methods to obtain an optimal feature set from 22 variables extracted from GPS records.

#### A. Detecting Subway Trips With Particular Rules

Unlike the other travel modes, most subway lines in Shanghai operate underground, which results in a serious signal loss. Even some lines travel above ground, GPS signals may be incomplete due to the closed metal body. Generally, there does not exist any GPS points during a subway trip or only exists a small quantity of points. These special characteristics make subway trips distinct from the trips taking by other transportation modes. In view of the interaction of all feature variables on detecting travel modes, it is more appropriate to apply rule-based method to identify subway modes than advanced classifier, which also could significantly decrease the computation cost and enhance the detection accuracy of the other modes. According to the quality of GPS data streams, two scenarios, without GPS signal and with incomplete GPS signal, were considered for subway trips detection.

Subway lines and stations are underground in most cases, thus it is probable not to receive any GPS signals for smartphones in an underground metro train. In this situation,



we recognize it as a subway trip by matching the origin and destination of the trip segment to some certain subway stations. If a trip segment satisfies the requirements that: (i) the distance between the start point of the trip segment with a certain station entrance/exit is less than the critical distance  $d_1$ ; (ii) the distance between the end point of the trip segment with a certain station entrance/exit is less than the critical distance  $d_2$ , then a subway trip was flagged. We collected coordinates of all station entrances/exits instead of the stations as a whole so as to decrease the critical distance and enhance the detection accuracy. For this algorithm, the critical distance  $d_1$  and  $d_2$  were tested with different thresholds. In reality, GPS can be received until one enters the station from a certain entrance, while it may not recover immediately after one walks out of an exit due to cold/warm start [27]. Thus, the critical distance for the end point is set to be larger than that of the start point.

Another situation with incomplete GPS records takes place when the metro train operates above ground. GPS points recorded during the subway trip are along with the metro line and the distance between these points and the subway line differs because of positioning error. A segment with incomplete GPS records were recognized as subway trip when it satisfies the following conditions: (i) firstly, it meets all the requirements for detecting subway trips without GPS signals; (ii) the points during a trip must be located in a distance less than the critical distance  $d_3$  with the subway line.

To sum up, the trip segments without GPS signal or with incomplete GPS records once meeting the above requirements will be identified as subway trips. This rule-based matching procedure was conducted in Matlab R2012a.

### B. Detecting the Other Modes

In this study, we employed a Random Forest classifier to divide the rest five modes: walking, bicycle, e-bike, bus and car. To obtain a best feature set, three different search methods combined with two evaluation strategies were applied to select the optimal input variables. This classification stage was completed in a data mining software Weka 3.6(Waikato Environment for Knowledge Analysis).

1) *Identifying Variables for Mode Detection*: A wide range of variables were selected and evaluated in this study to determine an optimal feature set that could classify five modes with the highest accuracy based on previous studies [11], [14], [17], [25] and the characteristics of our collected data. These selected variables can be divided into four main categories including: (1)speed, (2)acceleration, (3)orientation, (4)distance/duration. Table II summarizes the initial list of 22 variables and their abbreviations.

Speed-related variables are the most frequently used to detect different travel modes. Mean speed of a trip segment reflects the nature travel pattern under normal traffic circumstance. The maximum speed is usually applied for mode detection, as the speed limits are different across various modes. For example, the maximum speed of a car is obviously greater than that of bicycle. We adopted 95 percentile speed instead of the maximum speed to eliminate the disturbance

TABLE II  
INITIAL LIST OF PROPOSED INPUT VARIABLES

Category	Variables
Speed	Mean(SMEAN), variance(SVAR), 25 percentile(S25), 50 percentile(S50), 75 percentile(S75), 95 percentile(S95), interquartile range (SIQR), skewness(SSKEW) and kurtosis(SKURT) of speed distribution, and the ratio of GPS points with each of the following range of total GPS points: below 0.5m/s(SBELOW0.5), below 1m/s(SBELOW1), below 1.5m/s(SBELOW1.5), below 2m/s(SBELOW2).
Acceleration	Mean(AMEAN), 95 percentile(A95), variance(AVAR), skewness(ASKEW) and kurtosis(AKURT) of acceleration distribution.
Orientation	Maximum change of orientation(OMAX), and average change in orientation(OAVG).
Distance/Duration	Trip distance(DIST) and duration(DURA).

of extreme value. To further explore and utilize the speed distribution, interquartile range, which equals to 75 percentile speed minus 25 percentile speed, was proposed to distinguish diverse travel modes. Also, skewness and kurtosis of speed distribution served as input variables, as they indicated the speed pattern of a trip segment.

Considering different speed profiles for each mode, the ratio of GPS points with speed below certain threshold were extracted. For instance, the ratio of speed less than 1 m/s, is expected to capture the characteristics of periodical stops of buses. Therefore, four cases of the ratio of GPS points with speed below certain threshold were used in this study: below 0.5m/s; below 1m/s; below 1.5m/s; and below 2m/s.

Acceleration-based variables were another important feature set commonly used to detect travel modes. Given the existing literature findings, mean, 95 percentile, variance, skewness and kurtosis of acceleration distribution were incorporated in the developed model to identify various modes. Absolute value of all acceleration/deceleration were averaged to determine the mean, which took all acceleration and/or braking into consideration.

Orientation changes can also differentiate distinct travel modes. For example, a motorized vehicle can only drive on roads and may not usually turn or change to a new lane unless necessary. However, orientation may change frequently in some travel modes with greater randomness (e.g. walking, bicycle). Therefore, two orientation-related variables were applied in the classifier: average change in orientation and the maximum change of orientation.

Finally, considering the transfer capacity of different travel modes, additional two variables were included in the initial list: distance traveled and the trip duration. Intuitively, the distances and duration traveled by the motorized modes are much greater than those of non-motorized modes.

2) *Feature Selection*: Previous studies classify different travel modes by inputting proposed variables directly without a feature selection procedure. In fact, all input variables

TABLE III  
SUBWAY TRIPS DETECTION RESULTS

Trial	Parameters			Results			
	Critical $d_1$	Critical $d_2$	Critical $d_3$	No. of correctly detected	Accuracy(%)	No. of correctly detected	Error(%)
1	75	150	20	401	88.13	9	1.98
2	75	150	30	412	90.55	13	2.86
3	75	200	30	429	94.29	14	3.08
4	75	250	30	434	95.38	18	3.96
5	100	150	20	428	94.07	15	3.29
6	100	150	30	436	95.82	18	3.96
7	100	200	30	446	98.09	19	4.18
8	100	250	30	447	98.24	25	5.49
9	150	200	20	430	94.51	20	4.39
10	150	250	20	432	94.95	24	5.27
11	150	200	30	447	98.24	27	5.93
12	150	250	30	449	98.68	32	7.25

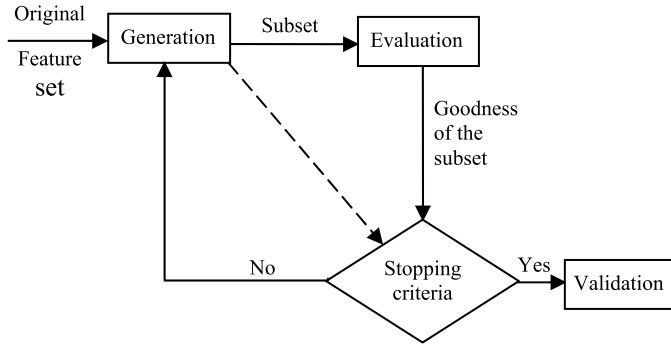


Fig. 2. Feature selection process with validation.

may not always obtain the highest detection accuracy. Thus, it is necessary to select a feature set incorporating a high distinctiveness with an attribute selection method.

The issue of feature selection is essentially a combination optimal problem, which determines a best set of features from the initial set of features so as to reduce the dimensions of original set. Searching from all candidate subsets is usually applied in mathematics to select the optimal combination. As shown in figure 2, Dash and Liu [28] put forward a basic framework of feature selection. In which, four function parts constitute a loop searching process.

**Generation** determines the search strategies, including the starting point for the search, direction and the style of generating the next subset. **Evaluation** assesses the new subset generated by previous step based on some certain criteria. This process known as the evaluation calculation will update the current optimal subset if the new one is better than previous one. After **Evaluation**, the searching process will cease by pre-defined **Stopping criteria**. Finally, in the **Validation** step, the selected feature set will be employed for training and predicting on the real data to compare the effectiveness. The comparison usually includes the time spent on training and predicting, the complexity of the model or the prediction

accuracy of classifier. In summary, **feature selection is actually a searching process**, in which the searching strategies and attribute evaluator determine the result of the selected subsets. In this study, three different searching strategies (Bestfirst, Greedy stepwise, and Ranker) combined with two main attribute evaluators (Filter and Wrapper) were employed to select the optimal feature set.

3) *Classifier Selection*: Each single travel mode segment incorporates 22 attributes. Different travel modes have various characteristic features and there are similarity/confusion and difference between them. For instance, bus and car may have the similar mean speed while car generally travels a farther distance than bus. Thus the detection of travel modes could be treated as a classification problem on large sample under high dimensions.

**Despite of feature selection, some attributes of sample are still severely biased**. Considering the large size of sample, the absolute number of these biased attributes is not small. In addition, several percentiles and ranges are simultaneously used to describe the speed distribution pattern in the same feature set, which are somewhat correlated.

To sum up, the sample extracted from each travel mode has several characteristics, e.g., large amount, high dimension and attributes correlated. In this case, the random forest [29] is chosen as the classifier. This method incorporates the following advantages. First, the application of randomly selecting variables and data to generate plenty of classification trees equips this classification a good ability to resist noise. Second, it is adaptable for diverse data sets and can deal with high-dimensional data. Besides, it can process not only discrete data, but also continuous data without normalization. Third, the fast training speed makes it feasible for implementing Parallel Arithmetic, which avoids a high computation cost.

As shown in figure 3, Random Forest classifier works as follows:

- 1) Selecting  $K$  subsets to form the training set  $X$  from original training data  $S$  by using the bootstrap sampling

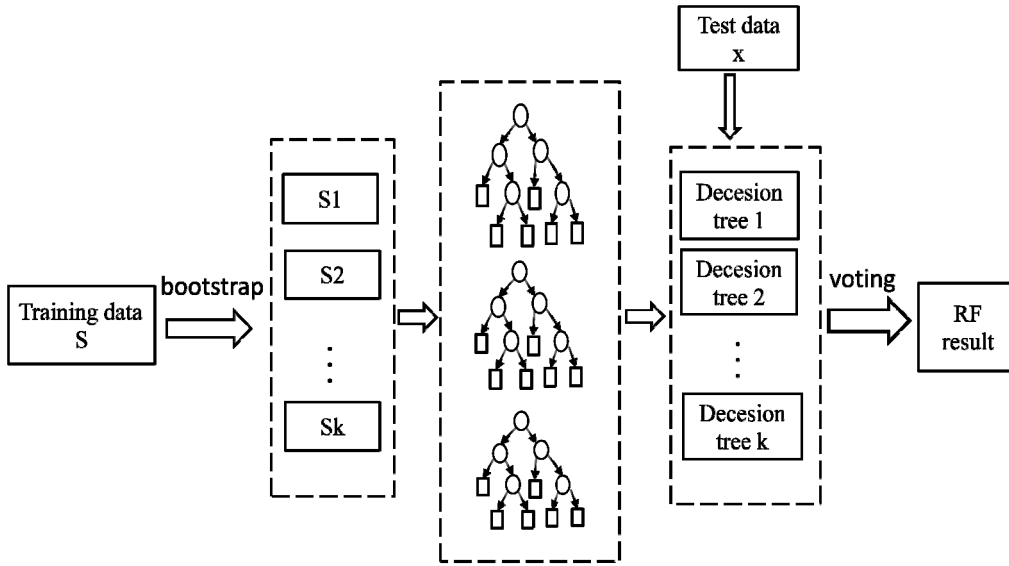


Fig. 3. Random forest theory.

method:

$$X = \{S_1 S_2 \dots S_K\} \quad (1)$$

- 2) The algorithm automatically builds up  $K$  decision tree models for each subset with a random vector  $\theta_k$ :

$$D = D_1 D_2 \dots D_K \quad (2)$$

All the random vectors  $\{\theta_k, k = 1, 2 \dots K\}$  are independent and identically distributed.

- 3) Every decision tree grows freely without pruning so as to get a 'forest', a combination of all trees:

$$\{h(X, \theta_k), k = 1, 2 \dots K\} \quad (3)$$

- 4) Each developed decision tree model has the final voting rights to determine the classification result of a new input variable  $x$ .

The classification result is:

$$H(x) = \max_Y \sum_{i=1}^k I(h_i(x) = Y) \quad (4)$$

Where  $H(x)$  represents the classification result;  $h_i(x)$  denotes classification result of  $i$ th decision tree;  $Y$  is the target category and  $I(\bullet)$  stands for characteristic function.

## V. RESULTS AND DISCUSSIONS

### A. Subway Trips Detection

As illustrated in 4.1, subway trips were identified with particular rules and a total of 12 parameter combinations was experimented in Matlab R2012a.

An optimal combination should be selected based on the highest accuracy with a lowest error. However, none of the combination list satisfied this requirement (see Table III). Combination 12 achieved the highest accuracy of 98.68%, while there was still an error of 7.25%. Consequently, combination 7 was preferred due to its comprehensive performance in detecting subway trips. This combination achieved an

accuracy as high as 98.09% with a relatively low error of 4.18%. For this parameter combination, the critical distance of  $d_1$ ,  $d_2$ ,  $d_3$  were 100 m, 200 m and 30 m, respectively.

### B. Results of the Remaining Five Modes Detection

1) *Feature Selection Results:* A necessary feature selection procedure was suggested before using the Random Forest classifier. **Seven feature selection combinations composing of three searching strategies** as well as two main attribute evaluators were conducted in the Weka 3.6. Full data was trained to evaluate the selected variables by the overall accuracy.

As shown in Table IV, the highest overall accuracy was achieved by serial number 2 and 3. These two methods selected the same seven variables including trip distance(DIST), mean speed(SMEAN), 50 percentile(S50), 75 percentile(S75), 95 percentile(S95), average change in orientation(OAVG), and **the skewness(SSKEW) of speed distribution**. Figure 4 exhibits the pattern distribution of the selected feature set. Overall, four speed-related variables are similar in the shape, which indicates the stability and effectiveness of them. Walking segments could be separated from the trips undertaken by other modes successfully based on the speed-related variables. 95 percentile speed seems to classify the five modes as three classes: walking as a single class, bicycle and e-bike as a class, the other two as a class. In contrast, trip distance may also fall the travel modes into three categories: walking and bicycle as a class, e-bike and bus as a class and car as a single class. As one might expect, walking and bicycle modes have the greatest value of average orientation change, which distinguishes them from the rest three means. Regarding the skewness of speed distribution, bicycle mode might be correctly picked out due to its special distribution.

2) *RF Classification Result:* Based on the selected variables, the rest five modes were detected by the Random Forest classifier. 2285 trip segments were split into 60% and 40% for model training and testing, respectively. Classification results

TABLE IV  
THE RESULT LIST OF FEATURE SELECTION COMBINATIONS

Serial number	Searching strategies	Attribute evaluator	No. of selected variables	Overall accuracy
1	Bestfirst	CfsSubsetEval	6	83.84%
2	Bestfirst	WrapperSubsetEval	7	88.88%
3	Greedystepwise	WrapperSubsetEval	7	88.88%
4	Greedystepwise	CfsSubsetEval	6	82.36%
5	Ranker	ReliefFAttributeEval	6	87.53%
6	Ranker	GainRatioAttributeEval	8	86.08%
7	Ranker	InfoGainAttributeEval	8	84.86%

TABLE V  
CONFUSION MATRIX OF GPS DATA FOR TESTING SET (914 INSTANCES)

Actual	Predicted					Precision (%)	Recall (%)	TPR	FPR	F-measure
	Walking	Bicycle	e-bike	Bus	Car					
Walking	463	1	0	0	0	99.4	99.8	0.998	0.007	0.996
Bicycle	3	75	2	3	0	89.3	90.4	0.904	0.011	0.898
e-bike	0	4	15	1	1	75.0	71.4	0.714	0.006	0.732
Bus	0	3	2	76	45	68.5	60.3	0.603	0.044	0.641
Car	0	1	1	31	187	80.3	85.0	0.850	0.066	0.826

are seen in Table V with respect to: (i) TPR: the rate of true positives (instances correctly classified as a given class), (ii) FPR: the rate of false positives (instances falsely classified as a given class), (iii) Precision: the proportion of instances that are actually of a class divided by the total instances classified as that class, (iv) Recall: the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate), and (v) F-Measure: A combined measure for precision and recall was calculated as:  $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ .

A total of 914 instances were tested for the developed model and 816 travel modes were correctly classified with an overall accuracy of 89.28% (see Fig. 5). Among them, nearly all the walking trips are correctly classified, which is due to the fact that walking has distinct characteristics from the other modes in terms of trip distance and average speed. Bicycle trips are detected with second highest precision, since the skewness of speed distribution distinguishes bicycle from the others effectively. In contrast, the precision of bus is unsatisfactory and there are still 45 bus trips misclassified as car trips. Correspondingly, the FPR of car is highest, which is almost 11 times of the e-bike. In other words, an incorrectly detected mode is most probably classified as the car, which indicates the car mode has less significant features distinguishing it from the other travel modes. Besides, the precision of e-bike is less-than-desirable for that its wild range of speed distribution makes it inclined to be misclassified as the other travel modes. Therefore, we need to extract additional attributes to differentiate e-bike from the other means. More

importantly, it is urgent to employ more data to address the issue of great confusion between bus and car trips.

Apart from the GPS data, socioeconomic attributes were also gathered in the survey period. Respondents completed their socioeconomic information online and the address is now available at <http://www.sjtuits.net/scheduling/gps.php>. Four socioeconomic attributes were extracted: whether or not having a bus card; the number of household bicycles; the number of household e-bikes and the number of household cars. These four attributes incorporating seven selected GPS variables were used to classify travel modes based on the Random Forest classifier. Also, the data were split into 60% and 40% for model training and testing, respectively.

Table VI shows the classification results of employing GPS data and socioeconomic attributes for testing data. The overall accuracy is 93.11%, which indicates a large improvement by adding socioeconomic data to detect travel modes. Though an improvement of 3.83% compared with that of using GPS data seems not to be great, it is rather difficult to enhance the results on the condition that the total accuracy is nearly 90%. As can be seen from Table VII, it is not surprising that the performance of walking detection does not improve anymore considering a F-measure of 0.996, which stands for the absolute advantage in correct classification. The precision of e-bike increases by 10% after incorporating the GPS data and the socioeconomic attributes. The most striking improvement reflects in the detection of bus and car. The number of increased correctly classified modes is 23 and 10 for bus and car, respectively, and the confusion number between

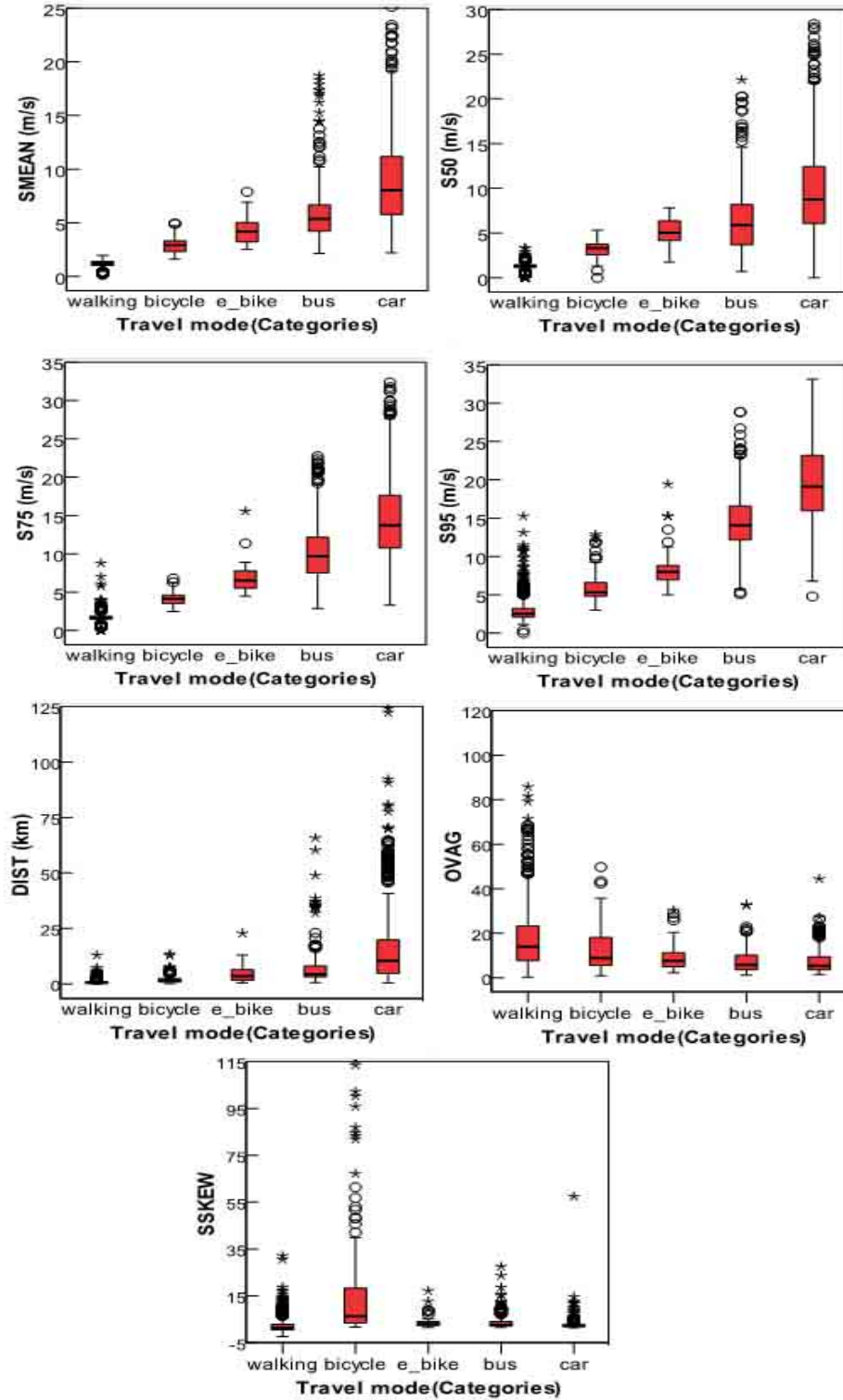


Fig. 4. Pattern distribution of selected variables.

these two modes decreases by 30. This encouraging enhancement indicates the excellent distinctiveness of socioeconomic data in addressing the confusion between bus and car modes. In addition, the FPR of car mode detection decreases by 0.3, nearly the half of that with only GPS data.

Statistically, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance

of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR as y axes) against the false positive rate (FPR as x axes) at various threshold settings, which depicts the trade-offs between true positive (benefits) and false positive (costs). Each prediction result or instance of a confusion matrix represents one point in the ROC space. The best possible prediction result



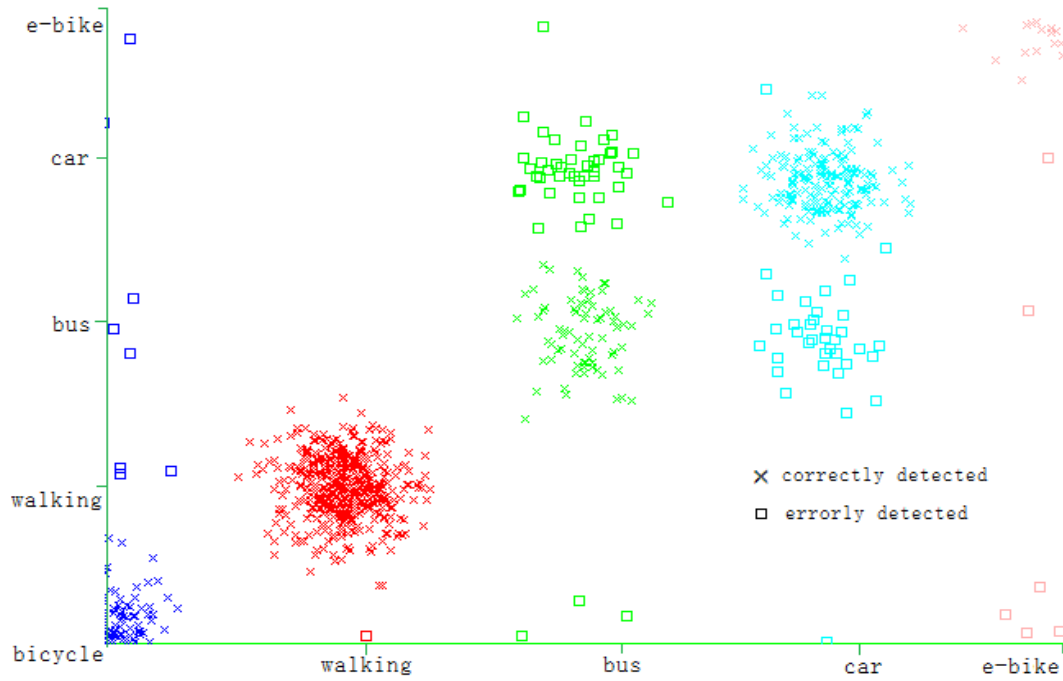


Fig. 5. Classification plot of GPS data for testing set.

TABLE VI  
CONFUSION MATRIX OF GPS DATA AND SOCIOECONOMIC ATTRIBUTES FOR TESTING SET(914 INSTANCES)

Actual	Predicted					Precision (%)	Recall (%)	TPR	FPR	F-measure
	Walking	Bicycle	e-bike	Bus	Car					
Walking	463	1	0	0	0	99.4	99.8	0.998	0.007	0.996
Bicycle	3	75	2	2	1	92.6	90.4	0.904	0.007	0.915
e-bike	0	2	17	2	0	85.0	81.0	0.810	0.003	0.829
Bus	0	2	1	99	24	79.2	78.6	0.786	0.033	0.789
Car	0	1	0	22	197	88.7	89.5	0.895	0.036	0.891

TABLE VII  
DIFFERENCE BETWEEN CONFUSION MATRIX OF THE GPS DATA VS. GPS DATA AND SOCIOECONOMIC ATTRIBUTES

Actual	Predicted					Precision (%)	Recall (%)	TPR	FPR	F-measure
	Walking	Bicycle	e-bike	Bus	Car					
Walking	0	0	0	0	0	0	0	0	0	0
Bicycle	0	0	0	-1	+1	+3.3	0	0	-0.004	+0.017
e-bike	0	-2	+2	+1	-1	+10	+9.6	+0.096	-0.003	+0.097
Bus	0	-1	-1	+23	-21	+10.7	+18.3	+0.183	-0.011	+0.148
Car	0	0	-1	-9	+10	+8.4	+4.5	+0.045	-0.030	+0.065

would yield a point in the upper left corner or coordinate (0,1), which is also called a perfect classification. A completely random guess would give a point along the diagonal line from the left bottom to the top right corners. The area under the curve (often referred to as simple the AUC) is equal to

the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [30]. Generally, AUC ranges from 0.5 to 1 and the greater AUC is, the better the classification performance is. When the AUC is greater than 0.9, it indicates that the model has an

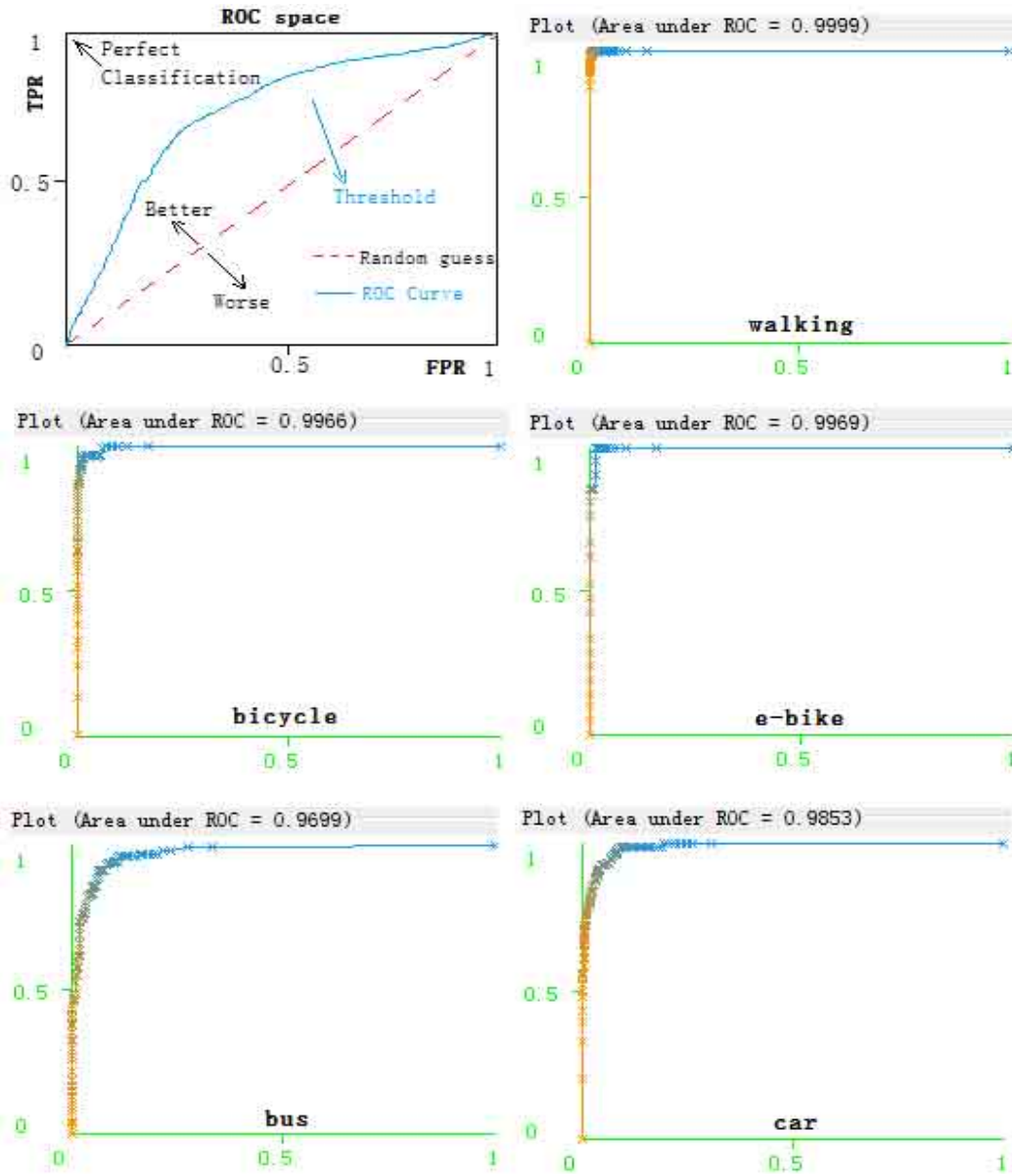


Fig. 6. ROC plot for five modes detection results.

excellent prediction ability in pattern recognition. The AUCs for five modes detection in our study are 0.9999, 0.9966, 0.9969, 0.9699, and 0.9853 for walking, bicycle, e-bike, bus and car, respectively (see Fig. 6). These encouraging AUCs give a statistical proof to the excellent classification capacity of Random Forest in this study.

3) *Comparison Between Random Forest and Other Classifiers*: To compare the Random Forest classifiers with other typical model in travel mode detection from GPS track data, two representative classifiers, i.e., the artificial Neural Networks (ANNs) and Support Vector Machines(SVM) are employed to the same selected feature set. The ANNs is conducted to classify travel modes based on the study by Xiao *et al.* [31]. The SVM is employed on account of the model proposed by Bolbol *et al.* [10], which applies a fixed-length moving window to obtain the best detection accuracy. The sample is also split into 40% and 60% for model training and testing, respectively.

TABLE VIII  
COMPARISON OF TRAVEL MODE DETECTION

Classifier	Testing Set(914 instances)	
	# of correctly classified	Accuracy (%)
ANNs	779	85.23
SVM	728	79.65
RF	851	93.11

As shown in Table VIII, the Random Forest classifier achieves the highest accuracy among all the classifiers for the testing set. Therefore, there is no doubt that the Random Forest classifier has an advantage over the other two representative classifiers in this sample. This result demonstrates the

applicability of Random Forest classifier for detecting travel modes from GPS track data. Compared with other classifiers, Random Forest has a good ability to resist noise due to the application of randomly selecting variables and data to generate plenty of classification trees. It can process not only discrete data, but also continuous data. The result proves that it is adaptable for diverse and high-dimensional data sets, which enables the RF classifier in mode detection incorporating multi-source attributes.

## VI. CONCLUSIONS AND IMPLICATIONS

In this study, particular rules were employed to detect subway trip segments and the Random Forest classifier were developed to identify the rest five prevalent travel modes in Shanghai. **GPS records and socioeconomic attributes from 312 respondents** were collected based on a smartphone travel survey. From the initial list of **22 variables** extracted from GPS data, seven distinctive variables were selected as the feature set to detect travel modes. These **seven variables** include trip distance, mean speed, 50 percentile speed, 75 percentile speed, 95 percentile speed, average change in orientation and the skewness of speed distribution. **Finally, subway trips were detected with an accuracy of 98.09% and the critical distance for  $d_1$ ,  $d_2$ ,  $d_3$  were 100m, 200m and 30m, respectively, experimented with 12 parameter combinations.** The inclusion of socioeconomic attributes has improved the overall detection accuracy of the rest five travel modes by 3.83%, which is a great enhancement in view that the overall accuracy of such a large sample has been close to 90%. Nearly all walking trips were correctly classified, and more than 85% trips were successfully identified for each mode except for the bus. Besides, it is worth mentioning that the socioeconomic attributes data could significantly improve the prediction of e-bike as well as address the confusion between bus and car modes.

Future studies may concentrate on **incorporating GIS sources** to improve overall detection accuracy. For instance, bus detection could be further perfected by matching bus networks. Another novel strategy to enhance the prediction accuracy could be to **examine the rationality and logic of the detected trip chain**. For example, it is unrealistic for a detected trip undertaken by bus and bicycle without a walking segment.

## ACKNOWLEDGEMENTS

The author would like to appreciate anyone who has provided recommendations and comments on this paper.

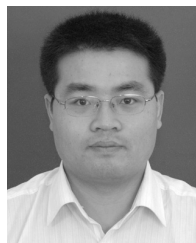
## REFERENCES

- [1] P. A. Gonzalez *et al.*, "Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks," *Intell. Transp. Syst.*, vol. 4, no. 1, pp. 37–49, 2009.
- [2] A. Bolbol, T. Cheng, and I. Tsapakis, "A spatio-temporal approach for identifying the sample size for transport mode detection from GPS-based travel surveys: A case study of London's road network," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 176–187, Jun. 2014.
- [3] P. Stopher, "Use of an activity-based diary to collect household travel data," *Transportation*, vol. 19, no. 2, pp. 159–176, May 1992.
- [4] G. Lei, M. Takayuki, S. Hitomi, and Y. Toshiyuki, "Deriving personal trip data from GPS data: A literature review on the existing methodologies," in *Proc. 9th Int. Conf. Traffic Transp. Stud. (ICTTS)*, 2014, pp. 557–565.
- [5] P. McGowen and M. McNally, "Evaluating the potential to predict activity types from GPS and GIS data," presented at Western Regional Sci. Assoc. 46th Annu. Meet., Newport Beach, CA, USA, Nov. 2007.
- [6] T. K. Rasmussen, J. B. Ingvarsson, K. Halldórsdóttir, and O. A. Nielsen, "Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the greater copenhagen area," *Comput., Environ. Urban Syst.*, vol. 54, pp. 301–313, Nov. 2015.
- [7] T. M. Mitchell, "Mining our reality," *Science*, vol. 326, no. 5960, pp. 1644–1645, 2009.
- [8] J. Du and L. Aultman-Hall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues," *Transp. Res. A, Policy Pract.*, vol. 41, no. 3, pp. 220–232, Mar. 2007.
- [9] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in The Netherlands," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 3, pp. 285–297, Jun. 2009.
- [10] A. Bolbol, T. Cheng, I. Tsapakis, and J. Haworth, "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification," *Comput., Environ. Urban Syst.*, vol. 36, no. 6, pp. 526–537, Nov. 2012.
- [11] T. Feng and H. J. P. Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transport Res. C, Emerg. Technol.*, vol. 37, pp. 118–130, 2013.
- [12] X. Zhou, W. Yu, and W. C. Sullivan, "Making pervasive sensing possible: Effective travel mode sensing based on smartphones," *Comput., Environ. Urban Syst.*, vol. 58, pp. 52–59, Jul. 2016.
- [13] H. Gong, C. Chen, E. Bialostozky, and C. T. Lawson, "A GPS/GIS method for travel mode detection in New York City," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 131–139, Mar. 2012.
- [14] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *Proc. ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2013, p. 13.
- [15] D. J. Patterson, L. Liao, D. Fox, and H. Kautz, "Inferring high-level behavior from low-level sensors," *Ubiquitous Computing*. Berlin, Germany: Springer, 2003, pp. 73–89.
- [16] S. Y. A. Tsui and A. S. Shalaby, "An enhanced system for link and mode identification for GPS-based personal travel surveys," presented at the 85th Annu. Meet. Transp. Res. Board, Washington DC, USA, Jan. 2006.
- [17] B. Assemi, H. Safi, M. Mesbah, and L. Ferreira, "Developing and validating a statistical model for travel mode identification on smartphones," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1920–1931, Jul. 2016.
- [18] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw GPS data for geographic applications on the Web," in *Proc. Int. Conf. World Wide Web*, Beijing China, Apr. 2008, pp. 247–256.
- [19] L. Zhang, S. Dalyot, D. Eggert, and M. Sester, "Multi-stage approach to travel-mode segmentation and classification of gps traces," in *Proc. ISPRS Workshop Geospatial Data Infrastruct., From Data Acquisition Updating Smarter Services*, 2011, pp. 87–93.
- [20] Y. J. Byon, B. Abdulhai, and A. S. Shalaby, "Impact of sampling rate of GPS-enabled cell phones on mode detection and GIS map matching performance," presented at the Transp. Res. Board 86th Annu. Meet., Washington, DC, USA, 2007.
- [21] P. Stopher, C. FitzGerald, and J. Zhang, "Search for a global positioning system device to measure person travel," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 3, pp. 350–369, Jun. 2008.
- [22] C. Chen, H. Gong, C. Lawson, and E. Bialostozky, "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study," *Transp. Res. A, Policy Pract.*, vol. 44, no. 10, pp. 830–840, Dec. 2010.
- [23] L. Liao, D. Fox, and H. Kautz, "Extracting places and activities from GPS traces using hierarchical conditional random fields," *Int. J. Robot. Res.*, vol. 26, no. 1, p. 119, 2007.
- [24] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, "Understanding transportation modes based on GPS data for Web applications," *ACM Trans. Web*, vol. 4, no. 1, pp. 1–36, 2010.
- [25] C. Rudloff and M. Ray, "Detecting travel modes and profiling commuter habits solely based on GPS data," presented at the Transp. Res. Board 89th Annu. Meeting, Washington, DC, USA, 2010.

- [26] W. Bao, G. Linjie, and J. Zhicai, "A trip detection model for individual smartphone-based GPS records with a novel evaluation method," *Adv. Mech. Eng.*, vol. 9, no. 4, pp. 1–10, 2017.
- [27] G. N. Xiao, Z. C. Juan, and J. X. Gao, "Inferring trip ends from GPS data based on smartphones in Shanghai," presented at the Transp. Res. Board 94th Annu. Meet., Washington, DC, USA, 2015.
- [28] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 1, pp. 131–156, 1997.
- [29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [30] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
- [31] G. N. Xiao, Z. C. Juan, and C. Q. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 447–463, Oct. 2016.



**Bao Wang** received the B.E. degree in traffic engineering from the Beijing University of Technology, China, in 2015. He is currently pursuing the M.E. degree in transportation engineering with Shanghai Jiao Tong University, China.



**Linjie Gao** received the B.E. and M.E. degrees and the Ph.D. degree in transportation planning and management from Jilin University, China. He joined the School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, China, as a Lecturer in 2008. His research interests are travel behaviour analysis and modeling, transportation planning, and traffic simulation.



**Zhicai Juan** received the M.E. and Ph.D. degrees in transport planning and management from the Jilin University of Technology, China, in 1984 and 1996, respectively. He is currently a Professor with the Antai College of Economics and Management, Shanghai Jiao Tong University, China.