Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 8582761, 13 pages
https://doi.org/10.1155/2019/8582761

Research Article

Developing a Travel Time Estimation Method of Freeway Based on Floating Car Using Random Forests

Juan Cheng, Gen Li, and Xianhua Chen

School of Transportation, Southeast University, Nanjing 211189, China

Correspondence should be addressed to Xianhua Chen; chenxh@seu.edu.cn

Received 28 May 2018; Revised 2 November 2018; Accepted 18 December 2018; Published 3 January 2019

Guest Editor: Ali Tizghadam

Travel time of traffic flow is the basis of traffic guidance. To improve the estimation accuracy, a travel time estimation model based on Random Forests is proposed. 7 influence variables are viewed as candidates in this paper. Data obtained from VISSIM simulation are used to verify the model. Different from other machine learning algorithm as black boxes, Random Forests can provide interpretable results through variable importance. The result of variable importance shows that mean travel time of floating car $\overline{t}_f$, traffic state parameter X, density of vehicle K_{all}, and median travel time of floating car t_{menf} are important variables affecting travel time of traffic flow; meanwhile other variables also have a certain influence on travel time. Compared with the BP (Back Propagation) neural network model and the quadratic polynomial regression model, the proposed Random Forests model is more accurate, and the variables contained in the model are more abundant.

1. Introduction

Along with economic and populations grow, the number of cars has increased dramatically, causing a series of problems such as traffic congestion, traffic accidents, and environmental pollution [1–3]. To tackle these issues, Intelligent Transportation System (ITS) is applied to the road system. Through the harmonious and close cooperation of people, vehicles, and roads, ITS can improve the efficiency of transportation, ease traffic congestion, improve road network capacity, reduce traffic accidents, lower energy consumption, and decrease environmental pollution. Travel time is the most intuitionistic index to reflect the running condition, which is an important foundation for constructing ITS [4]. Obtaining accurate travel time information, on the one hand, traffic departments improve traffic management decisions; on the other hand, travelers can make better travel choices [5]. Therefore, the accurate travel time of traffic flow is paid more attention by travelers, traffic managers, and scholars. As the basis of ITS, some researchers have conducted special studies on travel time.

Travel time can be achieved directly or indirectly. Direct methods measure travel time using probe vehicle, records at toll stations, tracking of cell phones, and many other technologies [6, 7]. Indirect methods infer travel time using measured traffic volume, speed, and occupancy in point sensors (e.g., loop detector and video camera) along the vehicle trajectory [8].

Recently, GPS on the vehicles and smartphones carried by occupants of motor vehicles can provide data support for travel time [9]. Therefore, travel time estimation using GPS data has been carried out [10–12]. Over the past few decades, a great number of models for travel time estimation have been developed, including models based on mathematical statistics and models based on artificial intelligence technology.

(1) Models based on mathematical statistics. These models provide interpretable parameters and a simple model structure [7, 13]; e.g., a piecewise truncated quadratic speed trajectory to estimate travel time was proposed by Sun [14]. The speed value can be selected between the highest and lowest, which was selected as the basis of the running condition of the vehicle. The method was more accurate when the vehicle is in the state of transition and congestion. Nevertheless, in the state of free-flow, the advantages of the proposed model were not obvious. Taken the number of single lanes, the speed limitation, and the instantaneous speed as independent variables, a multiple linear regression model based on the floating car was raised by Bobba [15].

The model was applicable during peak and off-peak periods. Yet the road section studied was the section between two signalized intersections (exclude signalized intersections), which was different from most current researches. Choosing different parameters, a linear regression model and a multiple linear regression model were developed by Faria [16]. The multiple linear regression model was more accurate, but the accuracy of the model is limited, only 60%. Using the GPS data with lower frequency, two mathematical models were proposed by Sanaullah [17]. The two models were based on the number of map matched points, connectivity of links, and spatial and temporal travel time components of the link, respectively. The experimental results indicate that vehicle penetration rates, data sampling frequencies, vehicle coverage on the links, and time window lengths all influence the accuracy of link travel time estimation. Zhan [18] used GPS-OD data from New York taxis to estimate travel time of road network segments. The impact of the single lane of the road on the driving vehicle was taken into account. When the road section was wider, the number of lanes may be more, and the single lane may not portray the fineness of the road network. Using the same data, a Bayesian model to estimate short-term travel time was presented by Zhan [19]. However, to reduce the modeling complexity, several assumptions were posed. Because of easy to implementation and low computational effort, models based on mathematical statistics are widely used. However, the accuracy is generally low.

(2) Models based on artificial intelligence technology. These models do not assume any particular model structure of the data but treat it as unknown. Some successful models include Fuzzy reasoning [20], machine learning [21], and the hybrid model [22, 23]. Such as, a three-layer Artificial Neural Network (ANN) model was presented to estimate link travel time by Zheng [24]. In the proposed model, individual probe vehicle's positions, link IDs, timestamps, and speed were used as input information. Compared with Hellinga's model, the ANN model performed quite well under different traffic conditions. However, the ANN model was applied to estimate travel time based on one car with GPS. Using the sparse and large-scale GPS trajectories, Tang [25] presented a tensor-based context-aware approach to estimate personalized travel time. The model was comprised of map matching, travel time tensor construction, context-ware feature extraction, and travel time tensor factorization. The proposed model considers the spatial correlation between different road segments, the deviation between different drivers, the fine-grain temporal correlation between different time slots, and the coarse-grain temporal correlation between recent and historical traffic conditions. A bus travel time prediction model based on SVM was proposed by Reddy [26]. The model used V-Support vector regression as a linear kernel function and used the data collected by public bus equipped with a GPS system to validate. The result showed that accuracy of the model was significantly improved under the condition of high variance. Although these models need large amounts of computation, the high accuracy drives scholars to shift their research focus on artificial intelligence technology method.

In summary, a wide range of models has been developed for travel time estimation. Although these models have their own advantages, the number of independent variables selected is limited, and the influence of traffic flow parameters on travel time has not been thoroughly considered.

In recent years, data mining and machine learning have gradually come into sight. The development of traffic information acquisition technology (such as data of GPS trajectories) has provided us with a large amount of traffic data, which offer an opportunity to develop a more accurate travel time estimation based on data mining. Compared with traditional parametric models, data mining algorithm can be deeply explored implicit relationships between variables. In view of this, the paper introduces a new data mining technique called Random Forests for travel time estimation. The influence of variables on travel time can be deeply excavated through Random Forests.

The rest of the paper is structured as followed. The next section will give the methodology of Random Forests to build a travel time estimation model followed by Section 3, which describes the data used in this paper. Results and discussions are presented in Section 4. Finally, the conclusions are outlined in Section 5.

2. Methodology of Random Forests

Random Forests is an integrated learning algorithm based on decision tree proposed by Breiman in 2001 [27]. Random Forests is a high-precision algorithm in machine learning, which can overcome the shortcomings of a single prediction or classification model.

2.1. Theory. Random Forests is a combination model consisting of a set of regression decision trees. Equation (1) shows the definition of Random Forests [28].

$$\{h(x, \theta_t),\ t = 1, 2 \ldots T\} \tag{1}$$

where $h(x, \theta_t)$ is a tree-structured classifier and $\{\theta_t\}$ is independent identically distributed random vectors. x is the independent variable. θ_t is the independent distributed random variable. T represents the number of decision trees.

Use the idea of ensemble learning to take the average of each decision tree as a regression prediction result, which is shown in

$$\overline{h} = \frac{1}{T}\sum_{i=1}^{T} \{h(x, \theta_t)\} \tag{2}$$

where $h(x, \theta_t)$ is output based on x and θ.

In order to overcome the problem that the decision tree model is not high in accuracy and is prone to overfitting, the idea of bagging and stochastic subspace was introduced in Random Forests [28, 29].

(1) Bagging. Bagging is a Bootstrap sampling technique proposed in 1996 [28]. Assuming that S is the original sample and N is the number of samples in S. The probability that each sample in S is not extracted is $(1 - 1/N)^N$.

If $N \longrightarrow \infty$, then

$$\lim_{N \longrightarrow \infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e} \approx 0.368 \tag{3}$$

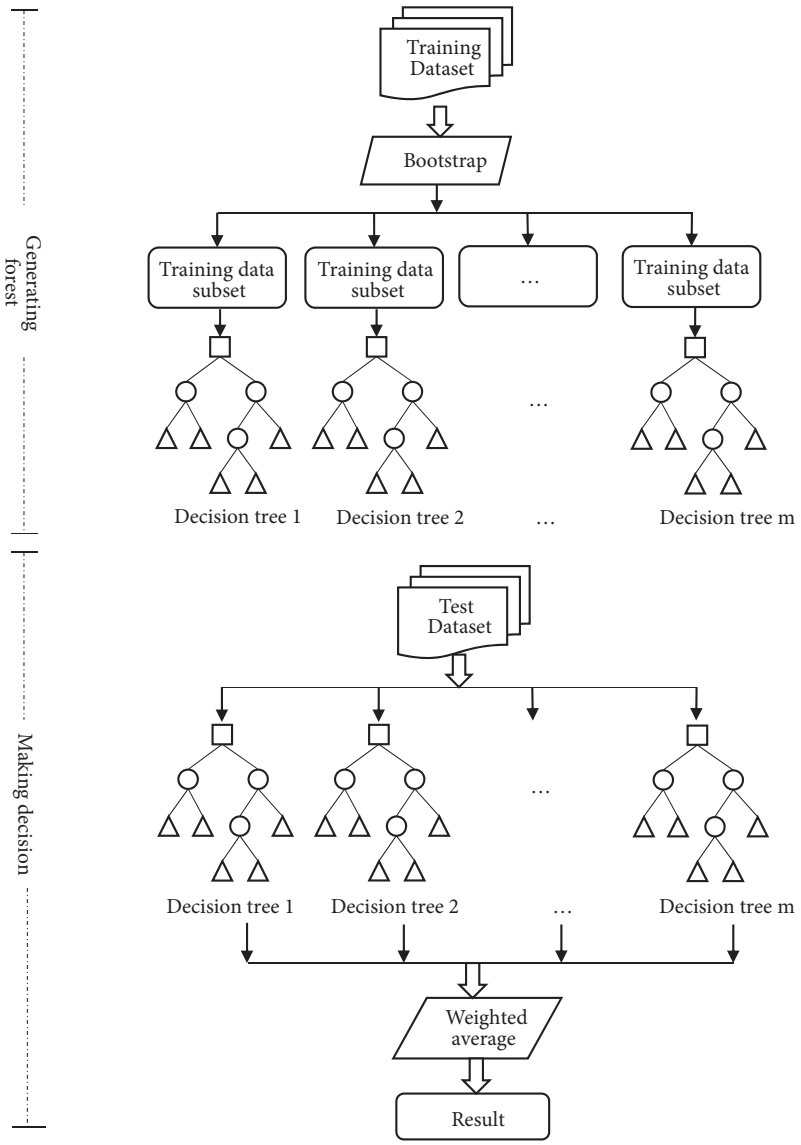

Figure 1: The process of establishing a Random Forests.

Equation (3) indicates that about 36.8% of the samples are not extracted each time, which is called OOB (Out-of-Bag) data.

(2) Stochastic Subspace. In the process of constructing the regression decision tree, each split node randomly extracts the feature subspace from the total feature space as the candidate feature set of the node and selects the optimal feature for splitting. The method ensures that the feature subsets are not only different among trees, but also the independence and diversity of the tree, and further improve the randomness in node splitting of Random Forests. Determining the stochastic subspace is to choose the number of explanatory variables to be checked for the splitting process.

In Random Forests, the final predictive performance of the model is determined by the number of trees in the forest (T) and the number of explanatory variables to be checked for the splitting process (m).

Figure 1 is the flowchart of the classifier and the flow of training and testing phases, which shows the process of establishing a Random Forests.

2.2. Generalization Error. Generalization error reflects the ability of the model to predict data outside the training set and is an important indicator for judging the quality of the model.

Definition 1. It is assumed that the training sets are extracted from the independent and identically distributed random vectors (X, Y), and the formed training sets are independent of each other. Then the mean squared error of the output $h(X)$ is $E_{X,Y}(Y - h(X))^2$.

In Random Forests, when there are enough regression decision trees and $h_t(X) = h(X, \theta_t)$, according to the large number theorem, Theorem 2 can be obtained.

Theorem 2. *When $t \longrightarrow \infty$, mean square generalization error converges on*

$$E_{X,Y}\left(Y - \overline{h}\left(X, \theta_t\right)\right)^2 \longrightarrow E_{X,Y}\left(Y - E_\theta\left(X, \theta\right)\right)^2 = PE^* \quad (4)$$

where θ_t is a random variable of the $t - th$ regression decision subtree. E_θ is a mathematical expectation. PE^* is generalization error of Random Forests.

Theorem 2 shows that, with the increase of the regression decision subtree t, Random Forests gradually converges, and generalization error will eventually tend to a limit value. Although Random Forests has been proven not prone to overfitting in mathematics [27], in the actual application process, the parameters of the Random Forests are optimized by experiments to further avoid overfitting.

3. Data

3.1. Traffic Simulation Software. To collect enough data for training and testing, traffic simulation software is used. Traffic simulation software is widely used in the study of traffic planning and traffic flow. The microscopic traffic simulation software can describe the road network and simulate the traffic flow through different models. Many types of traffic simulation software can be used to collect travel time data [30, 31].

(1) VISSIM. VISSIM is a microscopic traffic simulation software developed by PTV of Germany, which is a simulation system based on traffic behavior model. It uses a discrete, random, microscopic model with a time step of 0.1s. The longitudinal movement of the vehicle adopts the psychophysical car-following model proposed by Professor Wiedemann, and the lane-changing behavior of the vehicle adopts a rule-based algorithm. After an open COM interface, VISSIM has a good secondary development capability.

(2) CORSIM. CORSIM is developed by the US Federal Highway Administration (FHWA) and consists of two models, FRESIM and NETSIM. FRESIM is mainly used for the simulation of highways and expressways, while NETSIM is used for the simulation of urban road networks. It has lane change and car-following model simulation module and simulates the state of traffic flow in the road network with 1s simulation step. The software has functions such as analog timing, dynamic filter control, and cooperative filtering control. However, CORSIM lacks an allocation algorithm and it is difficult to evaluate the traffic volume transfer caused by ramp control, accidents, and travel information.

(3) PARAMICS. Developed by British Quadstone, PARAMICS can be applied to traffic simulation at different levels, from a single road network to a large-scale urban road network. PARAMICS supports multiuser parallel computing with a powerful application interface. However, PARAMICS lacks a model of mixed traffic and complex traffic flow.

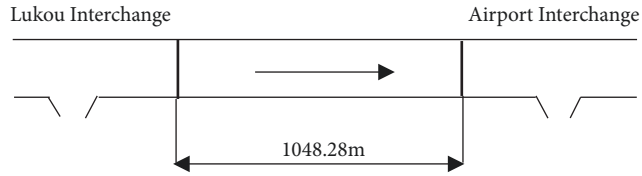

Figure 2: The study area.

(4) SimTraffic. SimTraffic is originally developed as transportation software for signal optimization timing and traffic model building. With the development of traffic simulation technology, SimTraffic has gradually developed into mature and fully functional microscopic traffic simulation software. It adds ramps, roundabouts, and highway modeling tools based on the original functions. Nevertheless, SimTraffic does not have a dedicated lane, as well as bus and car parking spots.

Through the above description, it can be found that the traffic simulation software has its own advantages and disadvantages. However, VISSIM has the ability to propose separate file output parameters, high traffic description accuracy, and simulated traffic has diverse characteristics. Therefore, data produced by VISSIM simulation software are used to verify the proposed travel time estimation model in this paper.

3.2. The Source of Data

3.2.1. Selection of the Simulation Section. Nanjing Airport freeway between the Airport Interchange and Lukou Interchange with the length of 1048.28m and 4 lanes in one direction is selected as the research area. Time detectors are set at both ends of the selected freeway section. The route diagram is presented in Figure 2.

3.2.2. Determination of the Simulation Parameters

(1) Simulation Traffic. The VISSIM simulation software is calibrated according to actual hourly traffic flow in Nanjing Airport freeway from Nanjing to Airport investigated by airport toll station at 9:00-15:00 on August 22, 2017. Since the real traffic flow does not include congestion, in order to cover the state of free-flow, transition, and congestion in the freeway, the traffic flow increased 600Veh/h from the real measured value of the previous period during 15: 00-17: 00, which reflect the state of congestion. Only increasing the number of vehicles does not necessarily result in congestion. However, based on the state of transition, the authors guarantee that all variables are constant and continue to increase the traffic flow to characterize the state of congestion. The input traffic flow is shown in Table 1.

(2) Vehicle Type. The user-defined taxi type is 1, and the vehicle color is blue; the truck type is 2, and the vehicle color is yellow; the bus type is 3, and the vehicle color is blue; the

Table 1: The input traffic flow.

time segments (s)	9:00-9:30	9:30-10:00	10:00-10:30	10:30-11:00
Simulation time segments (s)	0-1800	1800-3600	3600-5400	5400-7200
traffic flow (veh/h)	800	1200	1600	2000
time segments (s)	11:00-11:30	11:30-12:00	12:00-12:30	12:30-13:00
Simulation time segments (s)	7200-9000	9000-10800	10800-12600	12600-14400
traffic flow (veh/h)	2400	2600	2800	3000
time segments (s)	13:00-13:30	13:30-14:00	14:00-14:30	14:30-15:00
Simulation time segments (s)	1400-16200	16200-18000	18000-19800	19800-21600
traffic flow (veh/h)	3600	4200	4800	5400
time segments (s)	15:00-15:30	15:30-16:00	16:00-16:30	16:30-17:00
Simulation time segments (s)	21600-23400	23400-25200	25200-27000	27000-28800
traffic flow (veh/h)	6000	6600	7200	7800

car type is 4, and the vehicle color is red. The user-defined taxi is chosen as the floating car in this paper.

(3) Speed Distribution. On the freeway, the expected speed of car, truck, and bus is 120,100 and 100 km/h. The speed distribution of cars, trucks, buses, and taxis is shown in Figure 3.

(4) Vehicle Proportion. Through investigation, the vehicle proportion on the airport freeway section is car: truck: bus: taxi =0.42:0.12:0.26:0.2.

(5) Time Detector. In the freeway section, time detectors are set up to collect travel time of the individual floating car and travel time of the traffic flow. The mean and median values of travel time are calculated by the collection travel time of an individual floating car.

3.3. Design of Experimental Scheme. In the process of experiment, the dynamic changing process of the freeway traffic flow was simulated by changing the input traffic flow, including the state of free-flow, transition, and congestion.

Using different random seed number, the experiment simulated 133 times and the simulation time was 28800s. At last, 133 sets of data were obtained, representing 133 days' data of 9:00-17:00.

Travel time was obtained at the sampling interval of 300 seconds. At the same time, travel time of the floating cars was acquired at the sampling interval of 1 second.

3.4. Variables of the Model

3.4.1. Traffic State Parameter. In the Highway Capacity Manual [32], traffic state of the freeway was divided into six levels (namely A to F) according to the average speed and density. As we all know, speed, density, and traffic flow are three basic parameters, which are interrelated. If values for two of these parameters are known, the third can be computed. The standard of traffic state classification of a freeway is shown in Table 2.

In this paper, traffic state parameter refer to the standard of traffic state classification of a freeway, let $x = 1\ to\ 6$ for representing the traffic state A to F of the freeway respectively. The paper combined existing traffic state levels and described the freeway at a lower level. Therefore, traffic state of the freeway was divided into three categories. The state of free-flow includes level A and B, namely $x_f = 1, 2$; the state of transition includes level C and D, namely $x_t = 3, 4$; the state of congestion includes level E and F, namely $x_c = 5, 6$, which is presented in Table 3. The traffic parameter is $X = \{x_f, x_t, x_c\}$.

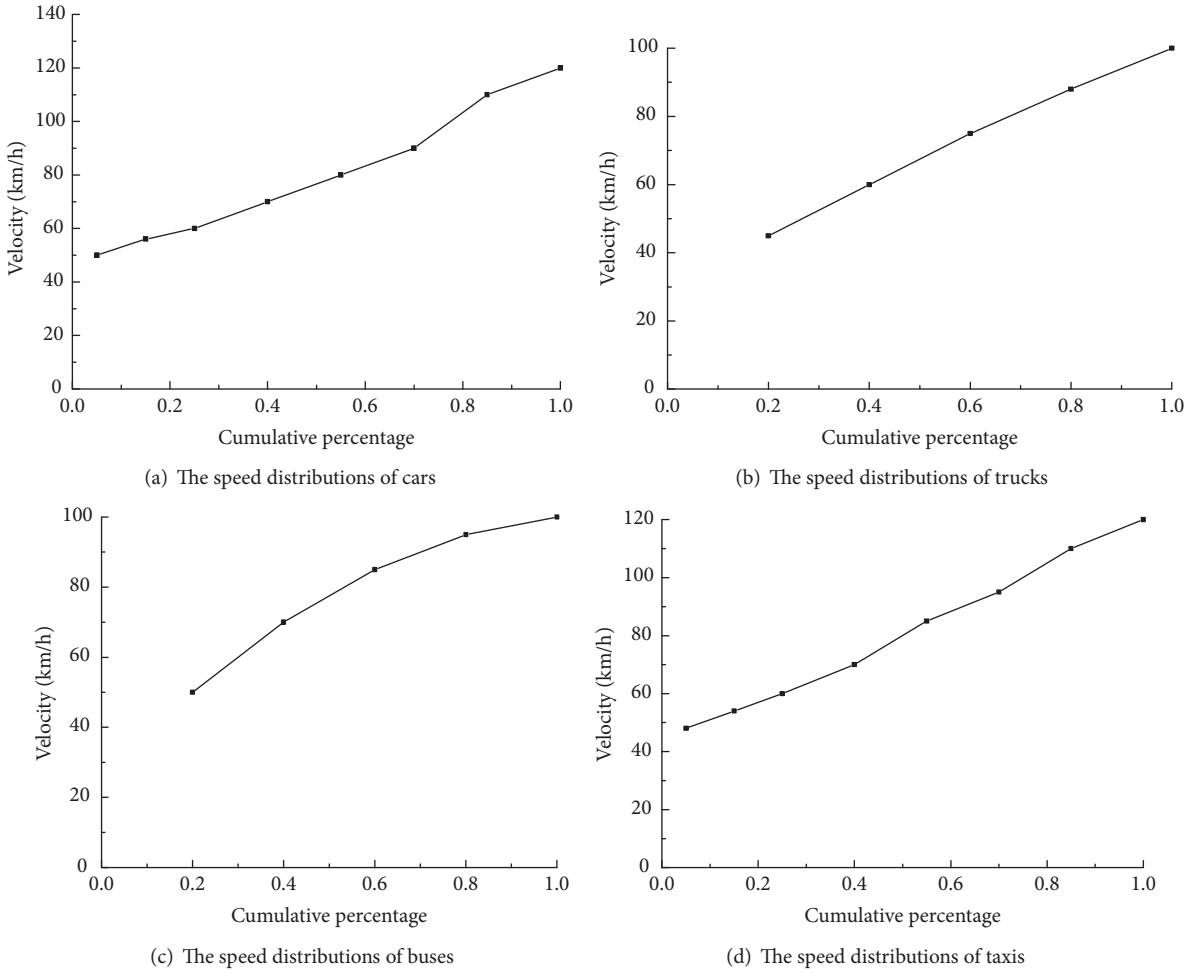

Figure 3: The speed distributions.

Table 2: The standard of traffic state classification of a freeway [32].

Traffic state	Density Range	Design speed (mi/h)			
		75		65	
	(pc/mi/ln)	Speed (mi/h)	Traffic flow (pc/h/ln)	Speed (mi/h)	traffic flow (pc/h/ln)
A	11	75	820	65	710
B	18	74.8	1350	65	1170
C	26	70.6	1830	64.6	1680
D	35	62.2	2170	59.7	2090
E	45	53.3	2400	52.2	2350
F	>45	<53.3	>2400	<52.2	>2350

Note: in order to keep the data neat, the unit used pc/mi/ln and mi/h in Table 2. Table 2 can change into pc/km/ln and km/h by 1mi = 1.609km. The number in Table 2 is the maximum value of each level.

3.4.2. Travel Time Calculation of the Floating Car. On any freeway section Δx_n, there are s floating cars within the time interval t_i to t_{i+1}, which is shown in Figure 4.

Assuming that travel time of each floating car on the freeway section is t_g (g = 1, 2, 3 . . . s), since the mean value expressed as $\overline{t}$ and the median value denoted by t_{men} can represent the general level of the whole data. The travel time of the floating car is being calculated by the mean value and median value respectively.

3.4.3. Variables of the Model. In the process of VISSIM simulation, 14 traffic variables can be obtained, that is, number of floating car N_f, occupancy of floating car R_f, number of vehicle N_{all}, occupancy of vehicle R_{all}, density of

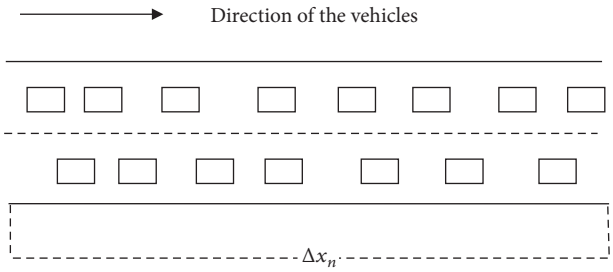

Figure 4: The distribution of floating cars on the freeway.

Table 3: The table of traffic state parameter.

traffic state	traffic state parameter	traffic state of the paper
A	1	free-flow
B	2	
C	3	transition
D	4	
E	5	congestion
F	6	

floating car K_f, speed of floating car V_f, traffic flow of floating car Q_f, density of vehicle K_{all}, speed of vehicle V_{all}, traffic flow of vehicle Q_{all}, ratio of floating car $Ratio_f$, travel time of floating car (mean value $\bar{t}_f$ and median value t_{men}), and traffic state parameter X.

4. Results and Discussions

Using the data obtained from VISSIM and the variables discussed above, the Random Forests model for travel time estimation was established.

SPM 8.2 data mining software developed by Salford Systems was used to establish the Random Forests model [33], although Random Forests can use the OOB error to evaluate the model. However, in order to compare with other models, in this paper 133 sets of data were trained and validated different models in two scenarios. Total data of 132 simulations were selected as training data, and the 5th simulation data were selected as test data, which used to compare with the mathematical statistics model. Meanwhile, in order to contrast with machine learning model, data of 133 days were divided into two data sets, in which 27-133 days of data were used as training data sets and 1-26 days of data were used as test data sets.

Mean Square Error (MSE), Mean Absolute Deviation (MAD), and Relative Error (RE) were selected as evaluation criteria.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \quad (5)$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (6)$$

$$RE = \frac{y_i - \hat{y}_i}{y_i} \quad (7)$$

where in (5)-(7) n is the total number of samples, y_i is the real value of travel time, and $\hat{y}_i$ is the estimation value of travel time.

4.1. Parameter Determination. There are 2 parameters to be ascertained in Random Forests, namely, the number of trees in the forest (T) and the number of explanatory variables to be checked for the splitting process (m).

(1) The Number of Trees in the Forest (T). In Random Forests, decision trees are not pruned. It is demonstrated that increasing the number of trees would not increase the precision but brings the computational burden. However insufficient trees are generated, the calculated variables importance may not be accurate enough. The number of trees in the forest was determined by 10-fold cross-validation. Table 4 shows the 10-fold cross-validation errors with a different number of trees in the forest.

It can be seen from Table 4 that the number of trees was 500 and 600 with the same minimum test error. The fewer the trees are, the smaller the computational burden is; therefore, the number of trees was 500 in the Random Forests model.

(2) The Number of Explanatory Variables to Be Checked for the Splitting Process (m). In Random Forests, only a subset of independent variables is checked to find the best splits, which makes the forest development more efficient. Beriman [27] has shown that randomly selecting a subset of independent variables to find the best split makes the process faster and leads to accurate results. Ghasri [34] adopted the square root of the number of independent variables in each split. There are other methods, such as twice the square root and half the square root of the number of independent variables in each split. When the number of explanatory variables to be checked for the splitting process is determined by 10-fold cross-validation error, the effect is not obvious. Random Forests can use OOB (Out-of-Bag) error estimates as unbiased estimates of generalization error without running a cross-validation procedure to measure the Random Forests model [28, 34–36]. Therefore, the OOB error was used to determine the number of explanatory variables to be checked for the splitting process. In this paper, the OOB errors are obtained using a different number of explanatory variables to be checked for the splitting process, respectively, and finally, choose the number of independent variables with the smallest value of OOB error. Table 5 shows the OOB errors with a different number of independent variables in each split.

Finally, the number of explanatory variables to be checked for the splitting process was 3 in the Random Forests model.

4.2. Variable Importance. Using the training data to train the model, the order of variables importance can be gained.

Table 4: 10-fold cross-validation errors with a different number of trees in the forest.

trees	100		200		300	
Data sets	Learn	Test	Learn	Test	Learn	Test
RMSE	2.2149	2.6889	2.2153	2.6857	2.2179	2.6789
MAD	0.9228	1.1243	0.9189	1.2112	0.9163	1.1165
trees	400		500		600	
Data sets	Learn	Test	Learn	Test	Learn	Test
RMSE	2.2126	2.6752	2.2144	2.6703	2.2144	2.6703
MAD	0.9141	1.1163	0.9152	1.1151	0.9152	1.1151

Table 5: The OOB errors with a different number of independent variables in each split.

m	1	2	3	4	5	6	7
OOB errors	10.283	7.219	6.342	6.936	6.591	6.474	7.230
m	8	9	10	11	12	13	14
OOB errors	7.306	7.235	7.300	7.260	7.282	7.395	7.302

Table 6: Variable importance.

Variable	Variable importance
mean travel time of floating car $\overline{t}_f$	58.98
density of vehicle K_{all}	35.58
traffic state parameter X	29.95
median travel time of floating car t_{menf}	15.72
density of floating car K_f	5.39
speed of vehicle V_{all}	4.74
occupancy of vehicle R_{all}	2.75
speed of floating car V_f	1.68
traffic flow of vehicle Q_{all}	0.22
number of vehicle N_{all}	0.22
occupancy of floating car R_f	0.21
traffic flow of floating car Q_f	0.15
number of floating car N_f	0.07
ratio of floating car $Ratio_f$	0.05

Variable importance explains the influence of independent variables on the dependent variable. The higher the value of the variable importance is, the stronger the influence on the model is. Variable importance is shown in Table 6.

It can be seen from Table 6 that mean travel time of floating car $\overline{t}_f$, density of vehicle K_{all}, traffic state parameter X, and median travel time of floating car t_{menf} are the important factors, which are much greater than other variables. It is indicated that travel time of traffic flow is closely related to travel time of floating car, density of vehicle and traffic state parameter.

4.3. Filtering Feature Variables. As can be seen from Table 6, in the established Random Forests model, much variable importance has low values, such as the ratio of floating car, indicating that there are some redundant variables in the model and the variables need to be screened. The literature [37] uses the variable importance obtained by the model to filter the feature variables, which is presented as follows:

(1) Create a Random Forests model using a set of feature variables containing n variables and rank the variable importance of the n feature variables in descending order.

(2) Delete the variable with the lowest variable importance among the n feature variables, and get the feature variable set containing n-1 variables.

(3) Create a Random Forests model using a set of feature variables containing n-1 variables and rank the variable importance of the n-1 feature variables in descending order.

(4) Delete the variable with the lowest variable importance among the n-1 feature variables, and get the feature variable set containing n-2 variables.

(5) Repeat steps (3) and (4) until there is one remaining feature variable.

(6) The Random Forests models are established containing n, n-$1, n$-$2 \cdots 1$ variables. OOB errors are ranked in order, and the Random Forests model with the smallest OOB error and the feature variable set is selected.

According to the method described above, a set of feature variables and a Random Forests model containing 7 feature variables are obtained. The training result of the model is shown in Figure 5 and the variable importance is shown in Table 7.

From Figure 5, reducing the number of feature variables does not reduce the performance of the model; however, the OOB error decreases from 6.342 to 5.586. It can be observed from Table 7 that mean travel time of floating car $\overline{t}_f$, traffic state parameter X, density of vehicle K_{all}, and median travel time of floating car t_{menf} are still the most important factors, which is much greater than the other three variables.

In all the 7 variables, $\overline{t}_f$ and t_{menf} enter the model at the same time, but the variable importance of $\overline{t}_f$ is much larger

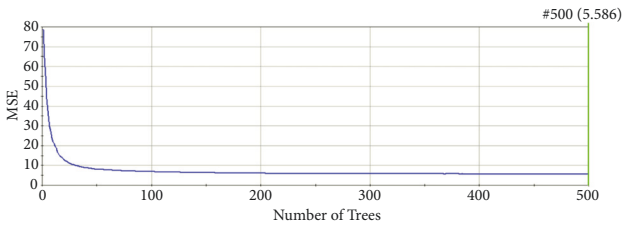

Figure 5: The result of Random Forests.

Table 7: Variable importance after filtering feature variables.

Variable	Variable importance
mean travel time of floating car $\overline{t}_f$	113.69
traffic state parameter X	45.07
density of vehicle K_{all}	41.38
median travel time of floating car t_{menf}	13.46
speed of vehicle V_{all}	4.88
density of floating car K_f	1.17
speed of floating car V_f	0.94

than t_{menf}. In the previous research, due to the correlation of variables, the two parameters were generally not included in the model at the same time. However, the mean value and median value can represent the general level of the whole data with different statistical significance. Therefore, the Random Forests model uses both $\overline{t}_f$ and t_{menf} as variables to take full advantage of different variables.

Density of vehicle K_{all} and density of floating car K_f have different effects on the Random Forests model. Density is the most important parameter of traffic flow, and it is an evaluation index of traffic demand. The higher the density is, the slower the speed is and the longer the travel time is. Density is an important indicator that affects travel time. K_{all} is much more important than K_f, which is simple to understand. The paper uses travel time of the floating car to calculate travel time of traffic flow but K_{all} represents the condition of all vehicles, which is a more intuitive reflection on travel time of traffic flow.

Speed of vehicle V_{all} and speed of floating car V_f have an influence on the estimated travel time because speed is the most intuitive reflection of travel time. The variable importance value of V_f is the lowest (0.94), but it is also an important factor affecting travel time of traffic flow. The reason is that the travel time estimation model is based on travel time of floating car and speed of floating car is closely related to travel time of floating car.

Traffic state parameter X is the second most important influence variable in the Random Forests model. As a newly introduced parameter in this paper, X is an intuitive indicator that directly reflects traffic states.

To sum up, the paper uses travel time of floating car to reflect travel time of traffic flow. Travel time of floating car (both $\overline{t}_f$ and t_{menf}) is an indispensable factor in the Random Forests model. Speed and density are the most intuitive reflection of travel time; therefore K_{all}, K_f, V_{all}, and V_f are the selected influence variable of the Random Forests model. The variable of X is selected in the Random Forests model because it directly reflects the traffic states.

4.4. Accuracy of the Established Model. To test the accuracy of the model presented in this paper, a quadratic polynomial regression model with different states was established according to the method of [38]. Consistent with the Random Forests model, the total data of 132 simulations were selected for quadratic polynomial regression and the 5th simulation data were selected as validation data. The quadratic polynomial regression model is provided in

$$T_{flow} = \begin{cases} 0.0275\overline{T}_f^2 - 1.2203\overline{T}_f + 45.5727 & free-flow \\ 0.0058T_{ment}^2 + 0.0902T_{ment} + 37.1768 & transition \\ -0.0034\overline{T}_c^2 + 2.0308\overline{T}_c - 65.0968 & congestion \end{cases} \quad (8)$$

where T_{flow} is travel time of traffic flow. $\overline{T}_f$ is mean travel time of floating car in the state of free-flow. T_{ment} is median travel time of floating car in the state of transition. $\overline{T}_c$ is mean travel time of floating car in the state of congestion.

Equation (8) is a regression model with different states. Although there are no separate states to establish the Random Forests model, the introduced traffic state parameter X can distinguish different traffic states. The errors of different models are presented in Table 8.

Meanwhile, BP (Back Propagation) neural network model was also established by using the data of 27-133 days of data as training data sets. BP neural network [39] is a multilayer feedforward network trained by error inverse propagation algorithm, which was proposed by a team of scientists led by Rumelhart and McCelland in 1986.

The topology of BP neural network includes input layer, hidden layer, and output layer, which are divided into information forward propagation and error back propagation [40]. BP neural network model with a three-layer feedforward Perceptron algorithm is used to estimate travel time. Figure 6 is the network structure of the three-layer BP neural network model designed in this paper.

In order to a fair comparison with the Random Forests model, the input variable in the BP neural network model is the selected variables in Section 4.3, and the output variable is the estimated travel time. The function of the hidden layer is logistic in the BP neural network. The network structure is 7-6-1, that is, the number of input layer nodes is 7, the number of hidden layer nodes is 6, and the number of output layer nodes is 1. Then the model was tested using 1-26 days of data sets. The training and test errors of different models are shown in Table 9.

Figure 7 is travel time obtained by different models. Figure 8 shows the comparison between travel time of the 5th day in the test data sets (real travel time) and travel time obtained with various models.

Several conclusions can be drawn based on Tables 8-9 and Figures 7-8.

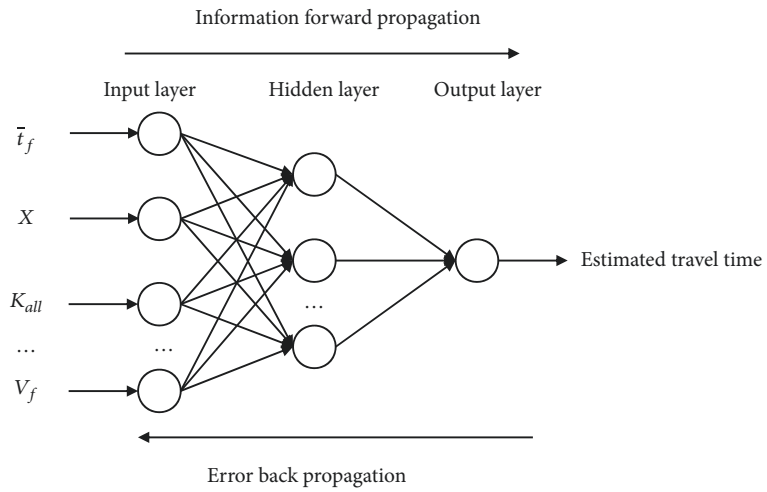

Figure 6: The network structure of the three-layer BP neural network model.

Table 8: The MAD of Random Forests model and reference [38] model.

MAD	Random Forests model	model of reference [38]
free-flow	0.5295	0.6556
transition	0.6640	1.2326
congestion	1.7573	6.8407
average	0.8047	5.5113

Table 9: The MAD of Random Forests model and BP neural network model.

Data set	Traffic state	Random Forests model	BP Neural Network model
Training Data	free-flow	0.5264	0.5414
	transition	0.5768	0.6003
	congestion	1.2856	1.8827
	average	0.8043	0.8987
Test Data	free-flow	0.5405	0.5898
	transition	0.6640	0.8375
	congestion	1.3024	1.9009
	average	0.8139	0.9076

Firstly, it can be seen from Table 8 that the accuracy of the Random Forests model is much greater than that of the quadratic polynomial regression model. In addition to travel time (mean and median) of floating car, the proposed model has selected another six variables, which indicate that Random Forests are not sensitive to the interaction between variables. Therefore, the Random Forests model can choose a richer impact variable.

Secondly, Table 9 shows the comparison between the Random Forests model and the BP neural network model; it is found that the error of the Random Forests model is generally less than the BP neural network model in both training data sets and test data sets. The reason may be different from the machine learning algorithm as black boxes (such as BP neural network and SVM); Random Forests has capabilities of data mining. The relationship between variables can be deeply exploited through Random Forests.

Thirdly, it revealed that, in Table 9, when traffic flow is operating in the state of free-flow with high speed, travel time obtained by the two models is close to the real value in both the two data sets. While in the state of transition and congestion with the lower speed, error of the proposed model is obviously less than the BP neural network model in both two data sets. It is shown that the model proposed in this paper has more advantages in the state of transition and congestion.

Fourthly, as indicated in Figures 7-8 that travel time obtained in this paper is consistent with the real travel time, which indicates that the proposed Random Forests model is effective.

Table 10: The variables entering the model after eliminating multicollinearity.

traffic state	variables
free-flow	mean travel time of floating car $\overline{T}_f$, traffic state parameter X_f, occupancy of vehicle R_{allf}, number of floating car N_f
transition	median travel time of floating car T_{ment}, traffic state parameter X_t, occupancy of vehicle R_{allt}, number of floating car N_t
congestion	mean travel time of floating car $\overline{T}_c$, traffic state parameter X_c, occupancy of vehicle R_{allc}, number of floating car N_c

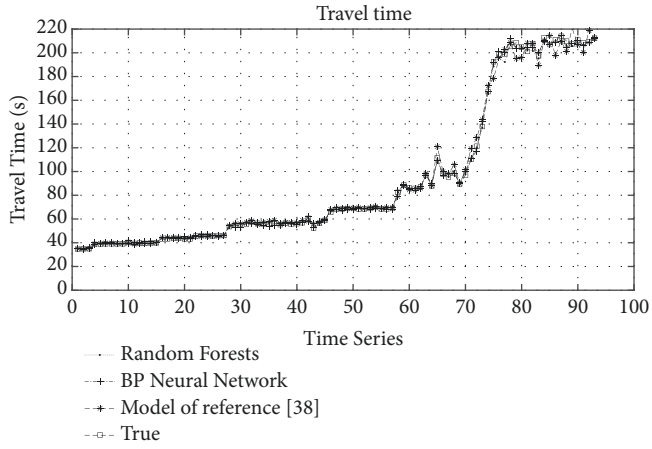

Figure 7: The estimated travel time.

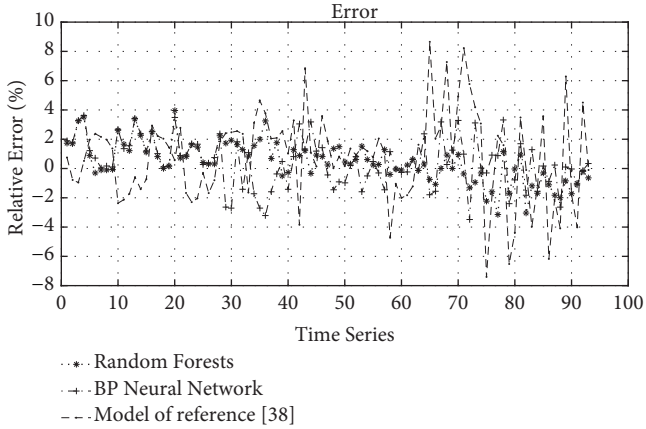

Figure 8: The relative error of the models.

In addition, the paper established a multiple linear regression model for different states using the 14 variables mentioned in Section 3.4.3. Before establishing a regression model, the multicollinearity between independent variables is tested firstly. Multicollinearity (collinearity for short) proposed by Freund [41] refers to a precise correlation or a high degree of correlation between the variables in the linear regression model, which makes the model difficult to estimate accurately. After eliminating multicollinearity, the variables entering the model are shown in Table 10.

It can be seen from Table 10 that, due to the multicollinearity of variables, when the regression model is established, only 4 variables are selected, and some of the variables that affect travel time are ignored, such as speed, density, etc. However, Random Forests is not affected by multicollinearity of variables. The relationship between travel time and variables can be deeply excavated through Random Forests.

5. Conclusion

In this paper, Random Forests is proposed for travel time estimation, which is a hotspot algorithm in machine learning and can deeply excavate the complex relationships between variables. The proposed model is established with 7 variables, namely, mean travel time of floating car $\overline{t}_f$, traffic state parameter X, density of vehicle K_{all}, median travel time of floating car t_{menf}, speed of vehicle V_{all}, density of floating car K_f, and speed of floating car V_f. Using different random seed number, the experiment simulates 133 times with VISSIM simulation software. Total data of 132 simulations are selected as training data, and the 5^{th} simulation data are selected as test data, which used to compare with the quadratic polynomial regression model. Meanwhile, data of 133 days are divided into two data sets, in which 27-133 days of data are used as training data sets and 1-26 days of data are used as test data sets in order to contrast with the BP neural network model. Comparison results show that the Random Forests model is more accurate than the quadratic polynomial regression model and the BP neural network model. The included variables are more abundant in the Random Forests model.

However, data are obtained by VISSIM, which limited the diversity of data. In future research, the variables of weather, characters of drivers, and other variables which affect travel time will be considered in the Random Forests model.

Data Availability

The data of the study was simulated by VISSIM and can be obtained up request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study is supported by the National Natural Science Foundation of China under Grants nos. 51478114 and 51778136.

References

[1] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research Part B: Methodological*, vol. 53, pp. 64–81, 2013.

[2] J. F. Xi, Z. H. Zhao, W. Li, and Q. Wang, "A traffic accident causation analysis method based on AHP-apriori," *Procedia Engineering*, vol. 137, pp. 680–687, 2016.

[3] T. Yi and B. M. Williams, "Dynamic traffic flow model for travel time estimation," *Transportation Research Record*, vol. 2526, pp. 70–78, 2015.

[4] W.-H. Lee, S.-S. Tseng, and S.-H. Tsai, "A knowledge based real-time travel time prediction system for urban network," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4239–4247, 2009.

[5] D. M. Miranda and S. V. Conceição, "The vehicle routing problem with hard time windows and stochastic travel and service time," *Expert Systems with Applications*, vol. 64, pp. 104–116, 2016.

[6] D. Woodard, G. Nogin, P. Koch, D. Racz, M. Goldszmidt, and E. Horvitz, "Predicting travel time reliability using mobile phone GPS data," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 30–44, 2017.

[7] M. Rahmani, E. Jenelius, and H. N. Koutsopoulos, "Non-parametric estimation of route travel time distributions from low-frequency floating car data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 343–362, 2015.

[8] L. Lu, J. Wang, Z. He, and C.-Y. Chan, "Real-time estimation of freeway travel time with recurrent congestion based on sparse detector data," *IET Intelligent Transport Systems*, vol. 12, no. 1, pp. 2–11, 2018.

[9] M. Rahmani, H. N. Koutsopoulos, and E. Jenelius, "Travel time estimation from sparse floating car data with consistent path inference: A fixed point approach," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 628–643, 2017.

[10] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 276–281, 2004.

[11] C. A. Quiroga and D. Bullock, "Travel time studies with global positioning and geographic information systems: an integrated methodology," *Transportation Research Part C: Emerging Technologies*, vol. 6, no. 1-2, pp. 101–127, 1998.

[12] J. Tang, F. Liu, Y. Wang, and H. Wang, "Uncovering urban human mobility from large scale taxi GPS data," *Physica A: Statistical Mechanics and Its Applications*, vol. 438, pp. 140–153, 2015.

[13] Z. Ma, H. N. Koutsopoulos, L. Ferreira, and M. Mesbah, "Estimation of trip travel time distribution using a generalized Markov chain approach," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 1–21, 2017.

[14] L. Sun, J. Yang, and H. Mahmassani, "Travel time estimation based on piecewise truncated quadratic speed trajectory," *Transportation Research Part A: Policy and Practice*, vol. 42, no. 1, pp. 173–186, 2008.

[15] R. Bobba, *Predicting Speeds on Urban Streets Using Real Time GPS Data*, University of Texas at Arlington, Arlington, Va, USA, 2002.

[16] D. Faria, "A framework to transform real-time GPS derived from transit vehicles to determine speed-flow," 2003.

[17] I. Sanaullah, M. Quddus, and M. Enoch, "Developing travel time estimation methods using sparse GPS data," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 20, no. 6, pp. 532–544, 2016.

[18] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.

[19] X. Zhan, S. V. Ukkusuri, and C. Yang, "A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data," *Automation in Construction*, vol. 72, pp. 237–246, 2016.

[20] Y. Li and M. McDonald, "Link travel time estimation using single GPS equipped probe vehicle," in *Proceedings of the 5th IEEE International Conference on Intelligent Transportation Systems, ITSC 2002*, pp. 932–937, Singapore, September 2002.

[21] A. Hofleitner, R. Herring, and A. Bayen, "Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning," *Transportation Research Part B: Methodological*, vol. 46, no. 9, pp. 1097–1122, 2012.

[22] S. Lee, B. Lee, and Y. Yang, "Estimation of link speed using pattern classification of GPS probe car data," in *Proceedings of the International Conference on Computational Science and Its Applications*, vol. 3981, pp. 495–504, 2006.

[23] B. A. Kumar, L. Vanajakshi, and S. C. Subramanian, "A hybrid model based method for bus travel time estimation," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 22, no. 5, pp. 390–406, 2017.

[24] F. Zheng and H. van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 31, pp. 145–157, 2013.

[25] K. Tang, S. Chen, and A. J. Khattak, "Personalized travel time estimation for urban road networks: A tensor-based context-aware approach," *Expert Systems with Applications*, vol. 103, pp. 118–132, 2018.

[26] K. K. Reddy, B. A. Kumar, and L. Vanajakshi, "Bus travel time prediction under high variability conditions," *Current Science*, vol. 111, no. 4, pp. 700–711, 2016.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[29] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[30] A. Louati, S. Darmoul, S. Elkosantini, and L. ben Said, "An artificial immune network to control interrupted flow at a signalized intersection," *Information Sciences*, vol. 433/434, pp. 70–95, 2018.

[31] J. Wu, M. Brackstone, and M. McDonald, "The validation of a microscopic simulation model: a methodological case study," *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 6, pp. 463–479, 2003.

[32] National Research Council, *HCM2010: Highway Capacity Manual*, Transportation Research Board, 5th edition, 2010.

[33] M. Gualtieri, C. A. Rowan, and K. TaKeaways, "The Forrester Wave ™: Big Data Predictive Analytics Solutions, Q1," *Forrester Research*, 2013.

[34] M. Ghasri, T. Hossein Rashidi, and S. T. Waller, "Developing a disaggregate travel demand system of models using data mining techniques," *Transportation Research Part A: Policy and Practice*, vol. 105, pp. 138–153, 2017.

[35] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour and Society*, vol. 14, pp. 1–10, 2019.

[36] J. Bao, P. Liu, X. Qin, and H. Zhou, "Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data," *Accident Analysis & Prevention*, vol. 120, pp. 281–294, 2018.

[37] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[38] J. W. Li, *Estimation and Prediction of Link Travel Time for Urban Trunk and Secondary Stress*, Jilin University, 2012.

[39] Y.-K. Liu, F. Xie, C.-L. Xie, M.-J. Peng, G.-H. Wu, and H. Xia, "Prediction of time series of NPP operating parameters using dynamic model based on BP neural network," *Annals of Nuclear Energy*, vol. 85, pp. 566–575, 2015.

[40] X. Yu, J. Han, L. Shi, Y. Wang, and Y. Zhao, "Application of a BP neural network in predicting destroyed floor depth caused by underground pressure," *Environmental Earth Sciences*, vol. 76, no. 15, Article ID 535, 2017.

[41] R. J. Freund and R. C. Littell, *SAS System for Regression*, SAS Publishing, 3rd edition, 2000.