Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 9263605, 11 pages
https://doi.org/10.1155/2020/9263605

Research Article

Vehicle Fuel Consumption Prediction Method Based on Driving Behavior Data Collected from Smartphones

Ying Yao,[1] Xiaohua Zhao,[1] Chang Liu,[1] Jian Rong,[1] Yunlong Zhang,[2] Zhenning Dong,[3] and Yuelong Su[3]

[1]*Beijing Key Laboratory of Traffic Engineering and the College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China*
[2]*Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843, USA*
[3]*Joint Laboratory for Future Transport and Urban Computing of Amap, AutoNavi Software Co., Ltd., Beijing 100102, China*

Correspondence should be addressed to Xiaohua Zhao; zhaoxiaohua@bjut.edu.cn

Received 26 December 2019; Revised 8 February 2020; Accepted 15 February 2020; Published 23 March 2020

Guest Editor: Feng Chen

Transportation is an important factor that affects energy consumption, and driving behavior is one of the main factors affecting vehicle fuel consumption. The purpose of this paper is to improve fuel consumption monitoring databases based on mobile phone data. Based on the mobile phone terminals and on-board diagnostic system (OBD) installed in taxis, driving behavior data and fuel consumption data are extracted, respectively. By matching the driving behavior data collected by a mobile phone with the fuel consumption data collected by OBD, the correlation between driving behavior and fuel consumption is explored, so that vehicle fuel consumption could be predicted based on mobile phone data. The fuel consumption prediction models are built using back propagation (BP) neural network, support vector regression (SVR), and random forests. The results show that the average speed, average speed except for idle (ASEI), average acceleration, average deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage are important indicators for fuel consumption evaluation. All three models could predict fuel consumption accurately, with an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application. This method lays a foundation for monitoring database improvement and fine management of urban transportation fuel consumption.

1. Introduction

Vehicle energy consumption and pollutant emissions are key problems for the healthy and sustainable development of urban transportation. With the continuous growth of car ownership in China, the energy consumption of its private cars increased 4.2 times, from 13.12 to 68.34 million tons of standard coal, from 2005 to 2015. Based on growth of the population, GDP, and the proportion of secondary and tertiary industries of China, the trend of future transportation energy consumption can be predicted. The energy consumption of private cars will continue to increase before 2020, when it is expected to reach 117.38 million tons of standard coal [1]. Therefore, reducing energy consumption has become one of the most important challenges in the transportation field.

Among many factors that affect the energy consumption of vehicles, driving behavior plays an important role. Research conducted by Ford Motor Company [2] shows that improvement of driving behavior could improve fuel economy by 25% in the short term. Providing drivers with continuous eco-driving feedback in the long term could lead to a 10% reduction in fuel consumption. Hiraoka et al. [3] studied the influence of ecological driving behavior on fuel consumption and found that giving feedback on fuel consumption information to drivers could improve fuel economy by 10%. In addition, the eco-driving instructions given to drivers could improve the fuel economy by approximately 15%. Ahn and Rakha [4] analyzed the influence of drivers' route choice on vehicle fuel consumption, and the results indicated that energy consumption and exhaust emissions

are significantly reduced by minimizing high-emission driving behavior. Thus, it is important to study the correlation between driving behavior and energy consumption and to use driving behavior to predict energy consumption.

At present, there is a significant volume of research on prediction models of energy consumption based on driving behavior. Hu et al. [5] conducted some real vehicle tests and a questionnaire survey to study the influence of driving style on the fuel consumption of electric vehicles on urban roads and constructed a prediction model for the fuel consumption of electric vehicles. Xu et al. [6] constructed two kinds of truck fuel consumption prediction models using driving behavior data obtained from the Internet of vehicles. The dynamic relationship between truck fuel consumption and truck drivers' driving behavior was described using an energy consumption index, and a generalized regression neural network model was established to predict truck fuel consumption. Zhao et al. [7] built a fuel consumption prediction model of urban road sections based on driving behavior by applying a machine learning algorithm, and the model could intuitively show the distribution characteristics of fuel consumption in basic sections of the Beijing expressway.

Data sources supporting the studies of fuel consumption prediction are mostly based on the data collected from the main controller of the vehicle, and an on-board diagnostic system (OBD) in conjunction with a questionnaire. The controller and OBD are limited by the equipment installation cost and drivers' installation willingness, so can only realize small-scale data management for small areas and with high uncertainty. The data collection form of a questionnaire also lacks flexibility, and it is difficult to guarantee the quality of the data.

With the rapid development of mobile terminal technology, the application of mobile phone sensors has been promoted. Mobile phone terminals have been used in the collection of driving behavior data and for the warning of dangerous driving. Johnson and Trivedi [8] proposed a system using dynamic time warping (DTW) and smartphone-based sensor fusion to detect nonaggressive and aggressive driving behavior, which gave audible feedback when it detected aggressive driving. Guido et al. [9] used the vehicle tracking data from smartphone sensors to estimate the safety performance of driving (including the deceleration rate to avoid crashes and the time to collision), and the crash risks in south-bound and north-bound lanes were analyzed. The application of the mobile phone terminal in driving safety has played an important role in the evaluation of vehicle fuel consumption. Because driving behavior data collected by mobile terminals are more detailed and easier to popularize, they lay a foundation for enriching urban road fuel consumption databases.

At present, the fuel consumption and emission data monitored by the statistical monitoring platform for the Beijing Municipal Transportation Administration are mostly based on OBD devices. The data collection objects are mainly taxi drivers, bus drivers, and truck drivers and do not cover all transportation enterprises. The mobile phone terminal provides a possibility for a larger scale of data collection. Fuel consumption cannot be directly collected by mobile phone terminals, but it could be predicted accurately by exploring the correlation between mobile phone and OBD data. At the same time, the driving behavior data collected by the mobile phone are influenced by the types, placement, shaking (caused by vehicle vibration), and drivers' usage of the phone, resulting in the instability of the driving behavior data, so a lot of calibration work needs to be done on the data. By constructing a fuel consumption prediction model, the application of mobile phone data could be used to calculate the fuel consumption of vehicles, which saves the installation cost of OBD equipment and provides a theoretical basis for traffic management departments to more accurately monitor urban traffic fuel consumption.

This study proposes a vehicle fuel consumption prediction method based on Global Positioning System (GPS) data collected from a smartphone. Taxi drivers participated in this experiment. By matching the driving behavior data of the mobile phone and the fuel consumption data of the OBD terminal, the driving behavior indexes that affect fuel consumption were screened, and the fuel consumption prediction models were constructed using machine learning algorithms. The prediction model of drivers' individual fuel consumption based on mobile phone data could not only further improve the real-time monitoring database of fuel consumption with strong error tolerance but also provide technical support for macro control of urban transportation energy consumption and effectiveness evaluation of the transportation energy policy.

2. Method

2.1. Analysis Framework. Since mobile phones cannot obtain the data of vehicles' fuel consumption directly, the driving behavior data collected from mobile phones and the fuel consumption collected from OBD were matched, and the fuel consumption prediction model was built. In the process of model construction, the data collected from mobile phones and OBD were both applied. After the model was built, larger-scale traffic fuel consumption was able to be predicted using only the driving behavior data collected from the mobile phones. The framework of model construction is shown in Figure 1. The steps of fuel consumption prediction are as follows:

(1) Data collection: natural driving behavior data of multiple drivers were collected based on GPS, linear accelerometer, gyroscope, and other sensors of mobile phones. At the same time, the real-time vehicle fuel consumption data were collected by the OBD terminal installed in the vehicle simultaneously.

(2) Index extraction: the data of mobile phones and OBD terminals were combined based on time. By comparing the consistency and difference of driving behavior data of the two terminals, the indexes for predicting vehicle fuel consumption based on mobile phone data were extracted.

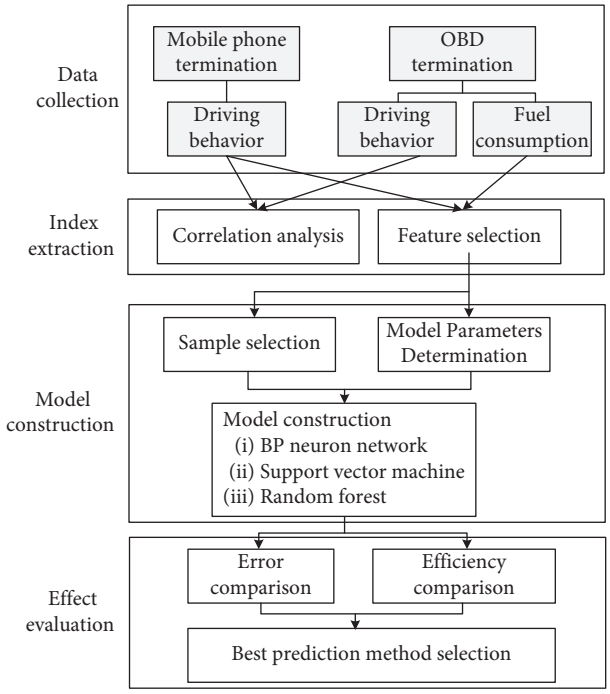

Figure 1: The framework of model construction. OBD means on-board diagnostic system and BP represents back propagation.

(3) Model construction: the training set and test set were selected randomly, and the fuel consumption prediction models were built using a back propagation (BP) neural network, a support vector machine, and a random forest.

(4) Effect evaluation: by building the fuel consumption prediction models several times and comparing the accuracy and efficiency of the three prediction models using different methods, the best method to predict vehicle fuel consumption based on mobile terminals is proposed.

2.2. Prediction Model. BP neural networks, support vector regression (SVR), and random forests are several common prediction methods with high accuracy and operation efficiency. This study built three types of prediction models, compared the difference in the prediction results, and finally we chose the best model for fuel consumption prediction.

2.2.1. BP Neural Network. An artificial neural network (ANN) is an operation model that mimics the process of neurons transmitting perceptual information to the human brain. This method has the characteristics of self-learning and high efficiency when processing nonlinear, unstructured, and large sample data. The error back propagation algorithm (BP neural network) [10] is one of the most widely used supervised learning algorithms in artificial neural networks. After the weights of the network are randomly selected, the BP neural network uses the back propagation method to update weights to minimize loss, and finally the connection weights of the network are determined. The structure of the vehicle fuel consumption prediction model based on a BP neural network is shown in Figure 2.

After screening the prediction indexes of fuel consumption, n indexes are determined as input variables. There are 5 neurons in the hidden layer, and the output y is the predicted fuel consumption. The connection weight between the input layer and the hidden layer is w_{ij}, and the connection weight between the hidden layer and the output layer is w_{jk}. First, the sample is transmitted through the input layer, and the data is converted into a nonlinear array within a certain range using the excitation function. Then, the nonlinear array reaches the output layer through weighting and outputs the results. If the error between the output fuel consumption and the actual fuel consumption exceeds the set of the expected error, the weight coefficient is corrected by back propagation. The network is repeat trained until the error is within the expected error, and the vehicle fuel consumption prediction model based on BP neural network is finally built.

2.2.2. Support Vector Regression (SVR). As a supervised machine learning algorithm, support vector machines are mainly applied to classification problems and regression problems [11]. The support vector machine algorithm transforms nonlinear problems into linear problems in high-dimensional space by constructing kernel functions, which gives the problem a geometrical explanation. The structure of the vehicle fuel consumption prediction model based on SVR is shown in Figure 3.

For a given set of samples $\{X_i, y_i\}$, $i = 1, 2, \ldots, m$, X is the n-dimensional input vector (including n driving behavior indicators) and y is the corresponding fuel consumption. The input vector is mapped to high-dimensional space, and the output y can be calculated as follows:

$$f(X) = w \cdot \varphi(X) + b, \tag{1}$$

where w is the weight vector, $\varphi(\cdot)$ is the mapping function that maps the input vector to the high-dimensional feature space, and b is the bias term.

By adding a convex optimization problem and relaxation factor, the support vector regression problem can be converted into the following equivalent solution:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \widehat{\xi}_i\right)$$

$$\text{s.t.} \quad \begin{cases} f(X_i) - w^T \cdot \varphi(X_i) - b \le \varepsilon + \xi_i \\ w^T \cdot \varphi(X_i) + b - f(X_i) \le \varepsilon + \widehat{\xi}_i \\ \widehat{\xi}_i, \xi_i \ge 0 \\ i = 1, 2, \ldots, n \\ C > 0, \end{cases} \tag{2}$$

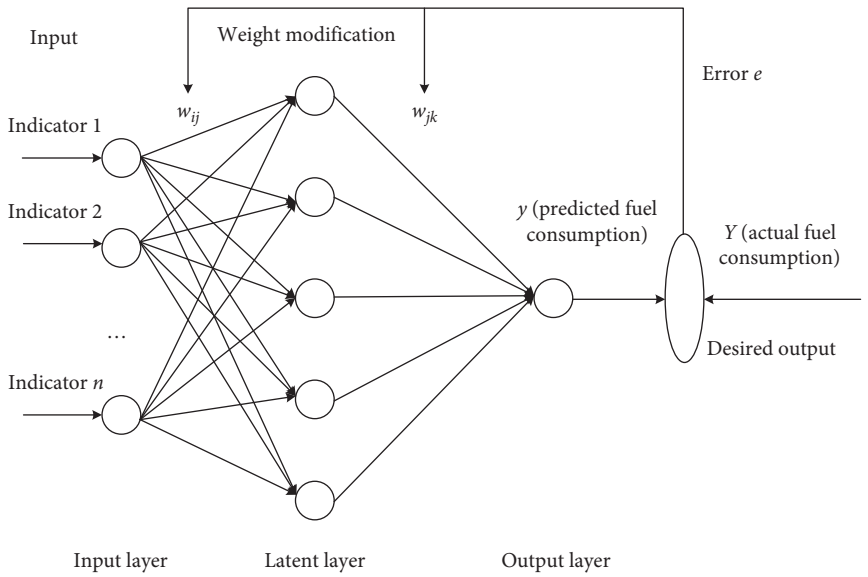

Figure 2: The structure of the vehicle fuel consumption prediction model based on the BP neural network.

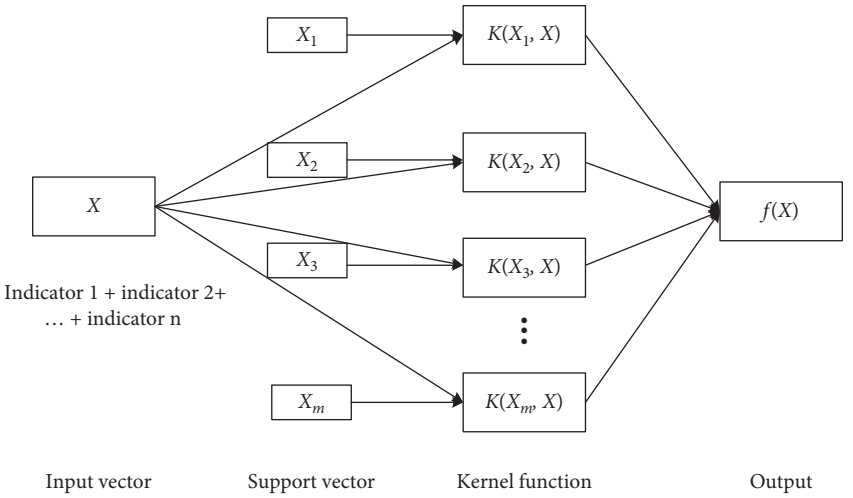

Figure 3: The structure of vehicle fuel consumption prediction model based on support vector regression (SVR).

where $\widehat{\xi}_i, \xi_i$ are the slack variables; C is the penalty-factor, which reflects the importance of outlier points; and ε is the insensitive loss function coefficients, which can ignore the error of the true value within a certain range and affect the final number of support vectors.

Three parameters, namely, ε, C, and the kernel function, should be determined when using the SVR algorithm to predict vehicle fuel consumption. The input vector is the n indicators required for fuel consumption prediction, and the output is vehicle fuel consumption. ε and C are determined by dividing them into several small cells according to certain rules. The model error corresponding to the variable value of each cell is calculated, and the variable values of the small cells with the minimum error are selected. The radial basis function (RBF) has a better performance in the application of SVR [12, 13]. Therefore, the kernel function adopted in this study is RBF, and the calculation method is as follows:

$$K(X, X') = e^{-||X - X'||^2/\sigma^2} \quad (3)$$

where σ is the hyperparameter of the RBF kernel, which is able to determine the range characteristics of input data and the correlation extent between support vectors.

2.2.3. Random Forest. A random forest (RF) is an effective classification method for prediction and classification [14]. A random forest is composed of a large number of decision trees. On the basis of decision trees, random processes are added to the row and column vectors, so as to avoid the potential overfitting problem of decision trees. For each tree, the training sample is sampled with replacement, and the

out-of-bag (OOB) data in each tree accounts for approximately 37% of the total data. The main calculation steps of the random forest regression algorithm are as follows:

First of all, k groups of training sample sets were selected by sampling with replacement. Secondly, m features were randomly selected from n features in each training sample set as splitting nodes, and k decision trees were generated. The node splitting of each decision tree adopted the principle of minimum mean square error, which minimizes the sum of mean square deviations of two groups of datasets after splitting. Finally, the predicted vehicle fuel consumption was obtained by averaging the predicted value of k decision trees. The structure of the vehicle fuel consumption prediction model based on the random forest is shown in Figure 4.

The three models have their advantages and disadvantages on the basis of different datasets. This study constructed three kinds of fuel consumption prediction models, and the most suitable and efficient model was chosen to predict fuel consumption.

3. Data Source and Index Extraction

3.1. Data Source. Experimental data were collected from OBD terminals installed in taxis and mobile phone terminals, and the sampling interval was 1 s. The data types that were collected are shown in Table 1.

The experiment was conducted in August 2017, and 20 drivers participated in the experiment to collect natural driving data for 15 days. All the taxicabs were Elantra with a 4-cylinder, 1.6-liter engine and were certified by the national-level-IV emission standard. On-board diagnostics (OBD) were installed in each taxicab during the experiment to collect driving behavior and fuel consumption data. The OBD devices have been widely used in Beijing taxi companies for over five years for monitoring the fuel consumption and emission data by the statistical monitoring platform for the Beijing Municipal Transportation Administration. The instantaneous fuel consumption of vehicles collected from OBD is calculated by relevant parameters such as engine load rate, engine speed, peak air inflow, and fuel correction factor. By comparing the fuel consumption calculated by OBD with the fuel consumption collected by CAN bus (calculated by fuel injection pulse width), the error of the instantaneous fuel consumption was within ±3% and the error of average fuel consumption per 100 km was 0.74% [15]. Meanwhile, drivers were asked to install software on their own mobile phone and keep the software running while driving to collect GPS data. The software is based on the android system and is specifically developed to collect GPS data from mobile phone sensors and calculate driving behavior. The two types of data were collected and uploaded to the cloud simultaneously.

Before the experiment, a mobile phone holder was given to each driver and held in the same position in the vehicle. The screen of the phone was placed perpendicular to the horizontal line. Mobile phone direction sensors were applied to the test to ensure that the location of the mobile phone is fixed and unified, and drivers were required to keep their phones in place while driving.

Although both the OBD and the mobile phone have GPS and drivers are required to place the mobile phone in a fixed position during driving, the output results of GPS data collected from OBD and mobile phone are different, which may be caused by the shaking of mobile phone when the vehicle vibrates or the differences of mobile phone type. In the actual driving process, mobile phone shaking and type differences are inevitable. Therefore, this study assumes that the construction of the fuel consumption prediction model could reduce the influence of data error collected by mobile phones and predict the fuel consumption accurately without the OBD device.

3.2. Index Extraction. By matching the data collected from the OBD and mobile phone terminals, the daily driving behavior of each driver and the corresponding fuel consumption could be obtained. There are many driving behavior factors that affect the fuel consumption of vehicles [16]. Seven indicators which could be calculated by mobile phone data were selected to predict fuel consumption. The types and definitions of the indicators are shown in Table 2. The acceleration condition is defined as acceleration greater than 0.1 m/s^2, the deceleration condition is defined as acceleration less than −0.1 m/s^2, and the condition of cruising is defined as the absolute value of acceleration less than 0.1 m/s^2. By averaging the driving behaviors of 20 drivers over 15 days, a total of 300 sets of data can be obtained.

Although road conditions, weather, and other factors also have a great influence on fuel consumption, they are not considered in this study. The main objective of this study is to evaluate the daily eco-driving level of taxi drivers, so as to help the traffic management department to monitor and improve the eco-driving skills of taxi drivers, and eco-driving training courses could be provided to drivers with poor eco-driving skills to reduce fuel consumption. Therefore, it is necessary to estimate the daily average fuel consumption (L/100 km) for taxi drivers. Since each taxi driver drives a different route each day, it is difficult to count all road types throughout the day. Although ignoring the influence of road conditions and other factors resulted in a decrease in the prediction accuracy of fuel consumption, the method adopted in this study is more applicable to a wider range of conditions and could estimate the daily ecological driving level of the drivers. The method also provides the feasibility demonstration for the future refined fuel consumption prediction. In the future research, the fuel consumption prediction results of drivers under different road conditions (such as ramps, curves, and intersections) would be analyzed and compared, so as to improve the accuracy of vehicle fuel consumption prediction.

Pearson correlation analysis was adopted to verify the correlation between driving behavior data from OBD and mobile phone terminals, and the results are shown in Table 3. It can be seen that, except for the cruising time

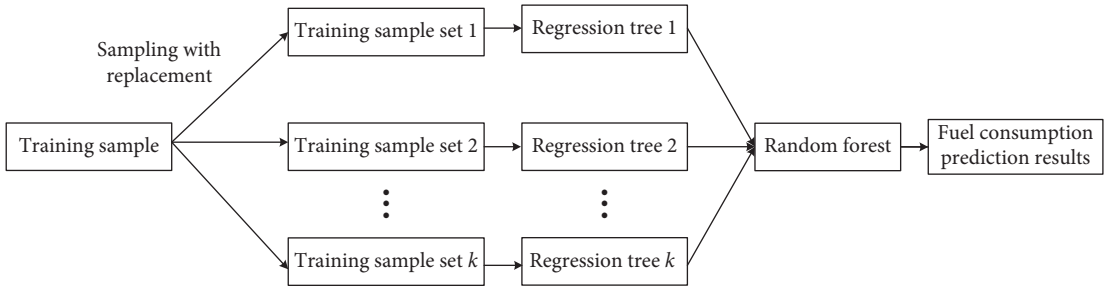

Figure 4: The structure of vehicle fuel consumption prediction model based on the random forest.

Table 1: Data types collected from the OBD terminal and mobile phone terminal.

OBD termination	Mobile phone termination
Time	Time
The global positioning system (GPS) latitude and longitude	The GPS latitude and longitude
GPS direction angle	Elevation
Speed in car dashboard	GPS speed
GPS speed	X-direction acceleration
Revolutions per minute (RPM)	Y-direction acceleration
Torque	Z-direction acceleration
State of air condition	X-direction angular acceleration
Oxygen sensor state	Y-direction angular acceleration
The instantaneous fuel consumption	Z-direction angular acceleration

Table 2: Related indexes to predict fuel consumption.

Indicators	Definition	Unit
Average speed V_{mean}	$V_{\text{mean}} = (1/T)\sum_{i=1}^{T} v_i$ where v_i is the speed of i second and T is the total driving time of one day	km/h
Average speed except for idle (ASEI) V'_{mean}	$V'_{\text{mean}} = (1/T')\sum_{i=1}^{T'} v_i$ where T' is the driving time of one day except idle	km/h
Average acceleration $\overline{a}_+$	$\overline{a}_+ = (1/t_a)\sum_{i=1}^{t_a} a_i$ where a_i is the acceleration of i second and t_a is the driving time of acceleration per day	m/s^2
Average deceleration $\overline{a}_-$	$\overline{a}_- = (1/t_d)\sum_{i=1}^{t_d} a_i$ where t_d is the driving time of deceleration per day	m/s^2
Acceleration time percentage P_a	$P_a = (t_a/T) \cdot 100\%$	%
Deceleration time percentage P_d	$P_d = (t_d/T) \cdot 100\%$	%
Cruising time percentage P_c	$P_c = (t_c/T) \cdot 100\%$ where t_c is the driving time of cruising per day	%
Fuel consumption FC	$\text{FC} = \sum_{i=1}^{T} \text{FC}_i/\text{distance}$ where FC_i is the instantaneous fuel consumption of i second and distance is the total driving distance per day	L/100 km

Table 3: Correlation analysis of driving behavior collected from OBD and mobile phone terminals.

	Pearson correlation coefficient	P value
Average speed V_{mean}	0.975	<0.001
ASEI V'_{mean}	0.936	<0.001
Average acceleration $\overline{a}_+$	0.793	<0.001
Average deceleration $\overline{a}_-$	0.670	<0.001
Acceleration time percentage P_a	0.662	<0.001
Deceleration time percentage P_d	0.662	<0.001
Cruising time percentage P_c	0.060	0.467

percentage, the other driving behavior indicators calculated by OBD and mobile phones are significantly correlated, with a correlation coefficient above 0.6. The reason for the difference in the cruising time percentage is that there are some differences in sampling accuracy between mobile phones and OBD, so the value of speed and acceleration calculated by GPS data collected from mobile phones and OBD are not exactly the same. The high correlation of multiple indicators indicates that the method of using mobile phone data to predict fuel consumption is feasible.

In order to verify the correlation between the data collected by mobile phone and the fuel consumption data collected by OBD and extract the relevant indexes for predicting fuel consumption, the relationship between different driving behavior indexes collected from mobile phones and fuel consumption collected from OBD were analyzed; the results are shown in Figure 5. As can be seen from Figure 5(a), the higher the average driving speed of the driver, the lower the fuel consumption. There is a strong correlation between average speed and fuel consumption. Since this study only considered the average driving speed of each day, the maximum average speed does not exceed 50, and the relationship between fuel consumption and speed is linear. From the perspective of instantaneous speed, fuel consumption increases when it exceeds 80 km/h, and the speed and fuel consumption are u-shaped curves [17]. The relationships between average acceleration/deceleration and fuel consumption are shown in Figures 5(b) and 5(c). The results show that a driver with heavy acceleration or deceleration during a day's driving would consume more fuel. Figure 5(d) shows the relationship between acceleration time percentage, deceleration time percentage, cruising time percentage, and fuel consumption. The results show that for a journey with lower fuel consumption, the driving time of cruising takes a larger proportion and the driver has less idle behavior, and a journey with high fuel consumption usually shows the driver as idle for a long time. Time percentage and fuel consumption also show a certain correlation, but these trends are not as obvious as the impact of the value of speed or acceleration on fuel consumption. In order to verify the influence of various driving behavior indexes on fuel consumption and select the related indexes of fuel consumption prediction, correlation analyses are examined in the following section.

Pearson correlation is a common filter-based feature selection method. By analyzing the Pearson correlation between driving behavior data collected by mobile phones and fuel consumption data collected by OBD, the key driving behavior indexes that affect vehicle fuel consumption can be selected through filtering. The results are shown in Table 4. All the driving behavior indexes are significantly correlated with fuel consumption ($P < 0.05$). Therefore, the indicators of average speed, ASEI, average acceleration, average deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage are selected to predict fuel consumption.

4. Results and Discussion

4.1. Model Training. The process of building the fuel consumption prediction model based on taxi drivers' daily driving behavior data is shown in Figure 6. On the one hand, the indicators of average speed, ASEI, average acceleration and deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage of each driver during each day were calculated using the driving behavior data collected from the mobile phone terminal. On the other hand, the instantaneous fuel consumption data of the vehicle were collected through the OBD terminal and converted into daily fuel consumption data. The two sources of data (driving behavior data and daily fuel consumption data) were matched through the collection time. Of all data collected, 75% were randomly selected as training samples and the remaining data were test samples. The fuel consumption prediction models were constructed based on the BP neural network, SVR, and random forest. To ensure the accuracy and stability of the prediction model, sample selection and model training were conducted 10 times. By comparing the difference in predicted fuel consumption and actual fuel consumption between the three models, the accuracy of using mobile phone data to predict vehicle fuel consumption was evaluated.

In the fuel consumption prediction model based on BP neural network, the "trainlm" algorithm was used for training, the logarithmic function "tansig" was used for the exciting function, and the linear function "purelin" was used for the node transfer function. The training times of the model were set as 100 times, and the convergence condition was set as the error of the model which is less than 0.001.

Based on the fuel consumption prediction model of SVR, the determination of the value of the insensitive loss function and penalty parameter was based on the exhaustive method, and the optimal value of the two coefficients was calculated by limiting the number of iterations to make the error less than a certain absolute value. The radial basis function (RBF) was taken as the kernel function of the SVR model.

Based on the fuel consumption prediction model of the random forest, 50 regression trees were set for training. The relationship between the number of regression trees and the out-of-bag error is shown in Figure 7. It can be seen that with the increase in the number of regression trees, the model error decreases, and the model is converged when there are about 50 regression trees.

4.2. Evaluation Results. The fuel consumption prediction results of one training process are shown in Figure 8. The figure shows the approximation degree between the three fuel consumption prediction results (BP neural network, SVR, and random forest) and the actual fuel consumption. As can be seen from Figure 8, some points with a larger deviation are the prediction results of the BP neural network model. However, in general, the three prediction models

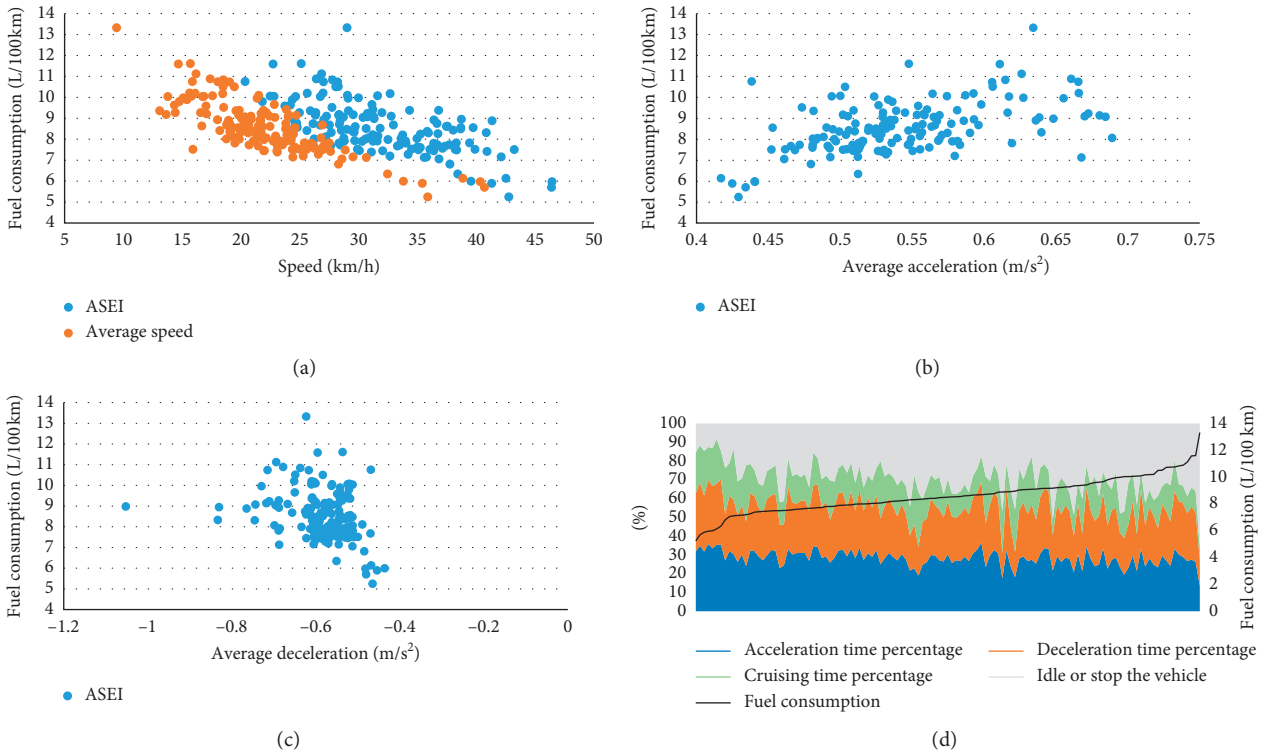

Figure 5: Fuel consumption distribution based on different driving behavior indexes. (a) The relationship between average speed and fuel consumption. (b) The relationship between average acceleration and fuel consumption. (c) The relationship between average deceleration and fuel consumption. (d) The relationship between time percentage and fuel consumption.

Table 4: Correlation analysis of driving behavior collected from mobile phone termination and fuel consumption collected from OBD.

	Fuel consumption	
	Pearson correlation coefficient	P value
Average speed V_{mean}	−0.8	<0.001
ASEI V'_{mean}	−0.659	<0.001
Average acceleration $\overline{a}_+$	0.515	<0.001
Average deceleration $\overline{a}_-$	−0.314	<0.001
Acceleration time percentage P_a	−0.363	<0.001
Deceleration time percentage P_d	−0.293	<0.001
Cruising time percentage P_c	−0.229	0.005

have a good fitting degree; the prediction results are basically distributed on both sides of $y = x$ with a high approximation degree.

In order to evaluate the accuracy and efficiency of the three fuel consumption prediction models, four indexes, namely, the root-mean-square error (RMSE), mean absolute percentage error K, R-squared, and model running time, were compared. The calculation methods of the first three of these indexes are as follows:

$$\text{RMSE} = \sqrt{\frac{\sum (f_i - y_i)^2}{n}},$$
$$K = \left| \frac{f_i - y_i}{y_i} \times 100\% \right|, \tag{4}$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2},$$

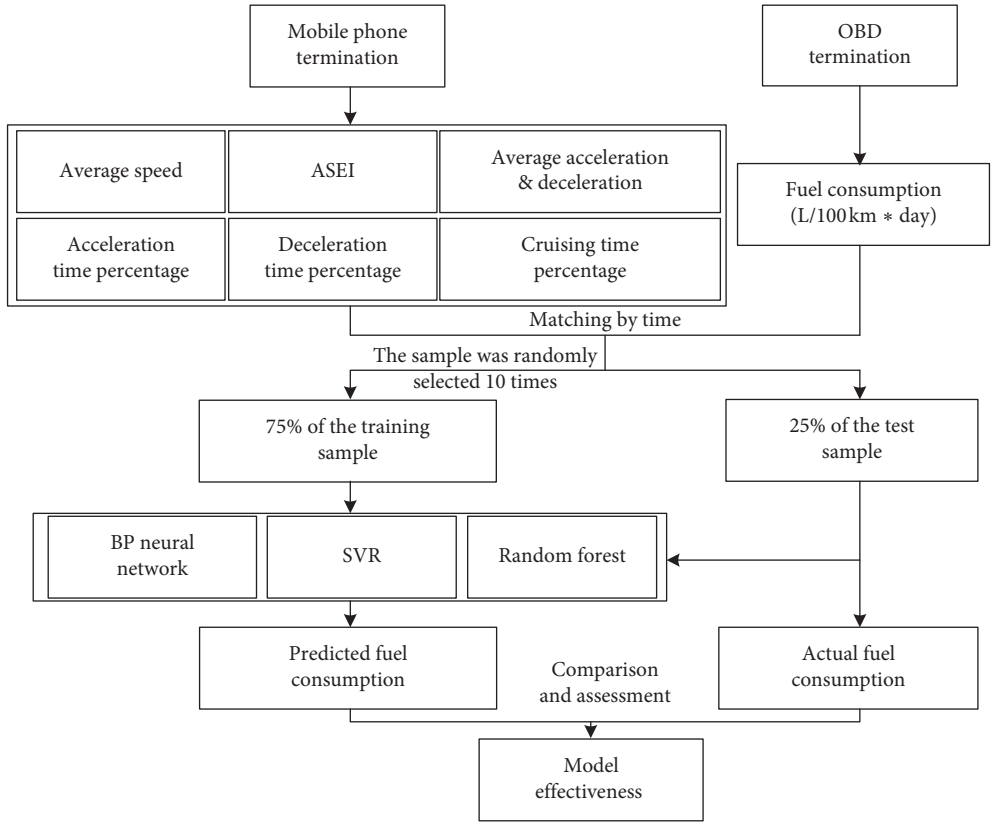

Figure 6: The process of building the fuel consumption prediction model.

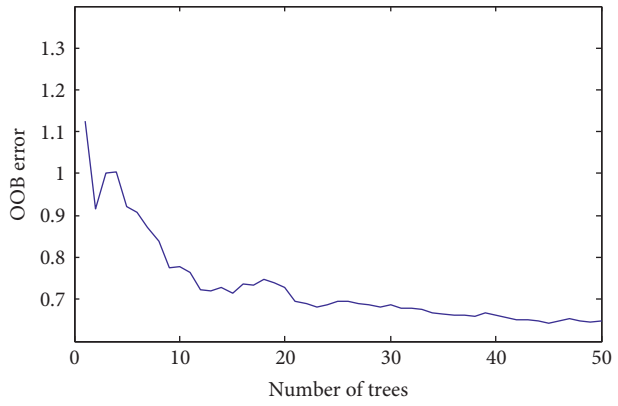

Figure 7: Out-of-bag error of the random forest model.

where f_i is the predicted fuel consumption, y_i is the actual fuel consumption, $\overline{y}$ is the average fuel consumption, and n is the number of samples.

The model evaluation results are shown in Table 5. It can be seen that the three models all show high prediction accuracy. The RMSE is 0.78–0.89 L/100 km, the absolute relative error (K) is 6.9%–7.5%, and the R-squared is greater than 0.5, indicating that the three models can accurately predict the fuel consumption of vehicles with the data collected by mobile phones. By comparing the errors and efficiency among the three models, it can be seen that the model based on the random forest has higher accuracy than BP neural network or SVR, and the running time of the random forest model is far lower than that of the BP neural network and SVR models. Therefore, the fuel consumption prediction model based on the random forest is effective and efficient for predictions based on individual driving behavior collected from mobile phones and is more suitable for practical applications to larger sample datasets.

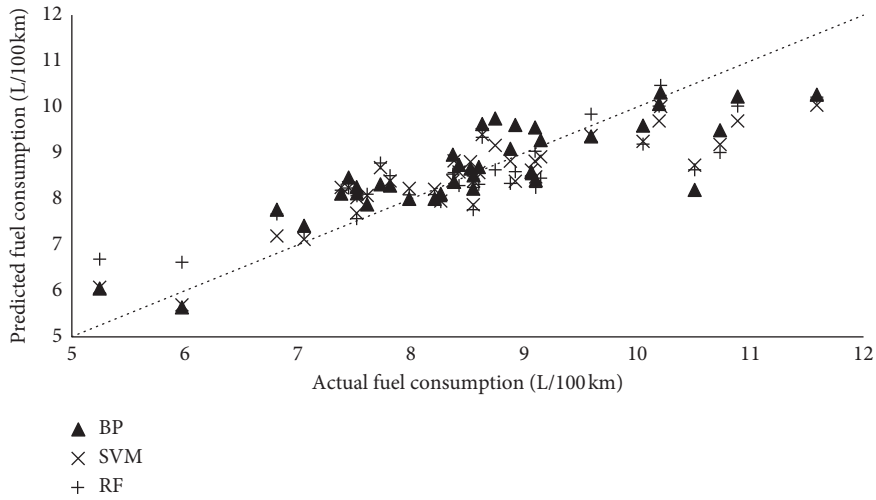

Figure 8: Fuel consumption prediction results.

Table 5: Model evaluation results.

Prediction method	Root-mean-square error (RMSE)	K	R-squared	Time (s)
BP neural network	0.872	0.075	0.547	0.724
Support vector regression	0.888	0.073	0.519	0.933
Random forest	0.783	0.069	0.635	0.140

5. Conclusion

In this study, driving behavior data and fuel consumption data of taxi drivers collected from OBD and mobile phone terminals, respectively, were matched. The correlation between driving behavior and fuel consumption was analyzed, and relevant driving behavior indicators affecting fuel consumption were extracted through the filter-based feature selection method. Using the seven selected driving behavior indicators (namely, average speed, ASEI, average acceleration, average deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage), three fuel consumption prediction models based on a BP neural network, SVR, and a random forest were constructed.

The results of model error and the run time comparison analysis show that the three models could predict fuel consumption accurately, and the random forest model had the highest accuracy and efficiency, with an RMSE of 0.783 L/100 km, mean absolute percentage error (K) of 6.9%, and model running time of 0.14 s. This finding is consistent with the research of Wickramanayake and Bandara [15], which also shows that random forest models are most effective in predicting fuel consumption based on driving behavior data. The research object of Wickramanayake and Bandara is the fuel consumption prediction of the bus, and this study focuses on the fuel consumption of the taxicabs. At the same time, the driving behavior data of this study are collected from mobile phones with higher flexibility and complexity rather than a fixed GPS device. This method could predict vehicle fuel consumption with high accuracy and efficiency based on cell phone data and provide strong support for traffic management departments to monitor the ecological levels of driving behavior of taxi drivers.

It is worth emphasizing that in the early stage of model construction, driving behavior data collected by mobile phones and fuel consumption data collected by OBD are applied. After the prediction model is built, mobile phone data can be directly used to predict the daily fuel consumption of drivers without installing OBD devices. Application of this method could change the traditional way of fuel consumption acquisition, and the use of mobile phone data to evaluate the ecological impacts of individual driving behavior could save the cost of equipment installation. At the same time, since not all taxi drivers are willing to install OBD devices in their taxicabs, this method could help increase the user data source, which could greatly improve the database size of taxi fuel consumption. Therefore, the method in this study could improve the depth, breadth, and refinement level of fuel consumption monitoring and management of taxi drivers' driving behavior, thus laying a theoretical foundation and providing technical support for the city to reduce fuel consumption.

This study aims to propose a method to predict vehicle energy consumption using mobile phone data. Although the sample size used in this study is limited, it provides a basis for larger scale and more accurate fuel consumption prediction. In future research, the collection of samples will be further expanded, and the fuel consumption under various road conditions, traffic conditions, and weather conditions would be considered. Through the data enrichment, model optimization, and improvement of the prediction indicators, the method could lay a theoretical foundation for the precise energy consumption supervision of taxi enterprises.

Meanwhile, since taxicabs are relatively homogenous, the fuel consumption prediction model in this study was fixed, taking only taxi drivers as the research object. In future study, more types of vehicles, such as buses and trucks, could be considered. Differentiated fuel consumption prediction models based on different vehicle types could be constructed to further improve the monitoring and management of urban energy consumption.

Data Availability

The driving behavior and fuel consumption data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the National Key R&D Program of China (Grant no. 2018YFB1601000), National Natural Science Foundation of China (Grant no. 61672067), Natural Science Foundation of Beijing Municipality (Grant no. 17JH0001), Joint Laboratory for Future Transport and Urban Computing of Amap, and Beijing Engineering Research Center of Urban Transport Operation Guarantee, Beijing University of Technology.

References

[1] H. Wang, "Energy consumption in transport: an assessment of changing trend, influencing factors and consumption forecast," *Journal of Chongqing University of Technology (Social Science)*, vol. 7, 2017.

[2] J. N. Barkenbus, "Eco-driving: an overlooked climate change initiative," *Energy Policy*, vol. 38, no. 2, pp. 762–769, 2010.

[3] T. Hiraoka, Y. Terakado, S. Matsumoto, and S. Yamabe, "Quantitative evaluation of eco-driving on fuel consumption based on driving simulator experiments," in *Proceedings of the 16th ITS World Congress and Exhibition on Intelligent Transport Systems and Services*, Stockholm, Sweden, September 2009.

[4] K. Ahn and H. Rakha, "The effects of route choice decisions on vehicle energy consumption and emissions," *Transportation Research Part D: Transport and Environment*, vol. 13, no. 3, pp. 151–167, 2008.

[5] K. Hu, J. Wu, and M. Liu, "Modelling of EVs energy consumption from perspective of field test data and driving style questionnaires," *Journal of System Simulation*, vol. 30, no. 11, pp. 83–91, 2018.

[6] Z. Xu, T. Wei, S. Easa, X. Zhao, and X. Qu, "Modeling relationship between truck fuel consumption and driving behavior using data from internet of vehicles," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 3, pp. 209–219, 2018.

[7] X.-h. Zhao, Y. Yao, Y.-p. Wu, C. Chen, and J. Rong, "Prediction model of driving energy consumption based on PCA and BP network," *Journal of Transportation Systems Engineering and Information Technology*, vol. 5, pp. 185–191, 2016.

[8] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1609–1615, Toronto, Canada, October 2011.

[9] G. Guido, A. Vitale, V. Astarita, F. Saccomanno, V. P. Giofré, and V. Gallelli, "Estimation of safety performance measures from smartphone sensors," *Procedia—Social and Behavioral Sciences*, vol. 54, pp. 1095–1103, 2012.

[10] W. J. Zhang, S. X. Yu, Y. F. Peng, Z. J. Cheng, and C. Wang, "Driving habits analysis on vehicle data using error back-propagation neural network algorithm," in *Computing, Control, Information and Education Engineering*, vol. 55, CRC Press, Guilin, China, 2015.

[11] H. Drucker, J. C. Chris, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, pp. 155–161, MIT Press, Cambridge, MA, USA, 1997.

[12] H.-l. Feng, "Study on prediction model of ecological security index in Chongqing city based on SVR model," *Computer Science*, vol. 40, no. 8, pp. 245–248, 2013.

[13] Z. Ramedani, M. Omid, A. Keyhani, S. Shamshirband, and B. Khoshnevisan, "Potential of radial basis function based support vector regression for global solar radiation prediction," *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 1005–1011, 2014.

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

[15] S. Wickramanayake and H. M. N. D. Bandara, "Fuel consumption prediction of fleet vehicles using machine learning: a comparative study," in *Proceedings of the 2016 Moratuwa Engineering Research Conference (MERCon)*, pp. 90–95, IEEE, Moratuwa, Sri Lanka, April 2016.

[16] M. Kuhler and D. Karstens, "Improved driving cycle for testing automotive exhaust emissions," in *Proceedings of the SAE International*, Dearborn, MI, USA, 1978.

[17] D. Yang, M. Li, and X. Ban, "Real-time on-board monitoring method of gasoline vehicle fuel consumption based on OBD system," *Journal of Automotive Safety and Energy*, vol. 7, no. 1, pp. 108–114, 2016.