Xiaowei HU[1]
E-mail: xiaowei_hu@hit.edu.cn
Yongzhi XIAO[1]
E-mail: xiaoyongzhi6@163.com
Tianlin WANG[1]
E-mail: 2832854461@qq.com
Lu YANG[1]
(Corresponding author)
E-mail: 18846075246@163.com
Pengcheng TANG[2]
E-mail: tangpc@foxmail.com
[1] School of Transportation Science and Engineering
Harbin Institute of Technology
Harbin 150090, China
[2] Chinaroads Communications Science & Technology
Group CO., Ltd., Beijing 100000, China

Traffic Engineering
Original Scientific Paper
Submitted: 17 May 2021
Accepted: 15 Sep. 2021

TRAFFIC VOLUME FORECASTING MODEL OF FREEWAY TOLL STATIONS DURING HOLIDAYS – AN SVM MODEL

ABSTRACT

Support vector machine (SVM) models have good performance in predicting daily traffic volume at toll stations, however, they cannot accurately predict holiday traffic volume. Therefore, an improved SVM model is proposed in this paper. The paper takes a toll station in Heilongjiang, China as an example, and uses the daily traffic volume as the learning set. The current and previous 7-day traffic volumes are used as the dependent and independent variables for model learning, respectively. This paper found that the basic SVM model is not accurate enough to forecast the traffic volume during holidays. To improve the model accuracy, this paper first used the SVM model to forecast non-holiday traffic volumes, and proposed a prediction method using quarterly conversion coefficients combined with the SVM model to construct an improved SVM model. The result of the prediction showed that the improved SVM model in this paper was able to effectively improve accuracy, making it better than in the basic SVM and GBDT model, thus proving the feasibility of the improved SVM model.

KEYWORDS

traffic volume; forecasting; SVM; holiday; quarterly conversion factor; freeway toll station.

1. INTRODUCTION

In recent years, with China's socio-economic development, people's living standards are increasing and the number of trips is at record high. More accurate traffic volume prediction has become a prerequisite for road construction and road traffic management and control. Traffic volume forecasting can be traced back to the early 20th century. Initially, researchers used growth rate models for traffic forecasting. Their calculation is simple and convenient. However, the accuracy is low and the results are not convincing [1].

In the 1920s, 1930s and 1970s, the AR (Auto Regressive) model, the MA (Moving Average) model and the ARMA (Auto Regressive Moving Average) model were proposed [2]. They formed the basis of time series analysis and are still widely used today. Time series models are well suited for modelling systems that are not easily modelled with exact mathematics and have uncertainties. In the beginning, researchers used the MA methods for traffic and accident forecasting. However, the MA requires a large amount of historical data as support and does not reflect the trend of change well. The exponential smoothing method considers dynamics of the time series to be relatively stable, so the time series can be reasonably extrapolated homoeopathically, and the most recent past dynamics will to some extent continue into the most recent future, so larger weights are placed on the most recent dynamics. Rui et al. used Markov model to modify the exponential smoothing method used to predict road passenger traffic and proved its accuracy [3]. Bezuglov et al. used the grey theory model for speed prediction to solve the adverse effects due to weather and accidents [4]. ARMA is a combination of AR and MA models. Gu et al. established ARMA and grey time series forecasting models for comparative analysis, and concluded that the prediction accuracy of the ARMA model is higher than that of the grey time series, maintaining more than 80% [5]. However, the ARMA model can only deal with smooth series, and common time series are generally non-smooth, so they must be transformed into smooth series by differentiation before the ARMA model can be used. Therefore, the ARMA models used for traffic volume prediction of road sections and intersections were proposed [6–8].

However, the time series method mainly uses the correlation between data for forecasting, without considering the influence of external factors, and mainly highlights the role of time factors in forecasting. When the external factors change significantly, the traffic volume forecast often has a large deviation. At this time, regression models can be used for analysis. Cai et al. used grey comprehensive correlation to analyse the influencing factors of air transport passenger volume and selected the main influencing factors to establish a multiple regression model of air passenger volume [9]. The local linear regression model was applied to short-term traffic forecasting and its accuracy was proved to be higher than the general nonparametric regression model [10].

Since the 1980s, with the development of computer technology, more and more new methods have been used in traffic prediction. The most representative are neural network models and support vector machine (SVM) models. With the continuous rise of machine learning in recent years, neural network learning has been applied to many aspects of traffic: Yun analysed the relationship between data characteristics and prediction accuracy of neural network models in traffic volume prediction [11]. Dia and Ishak et al. used neural networks to predict short-term traffic speeds [12]. Huang et al. proposed a prediction neural network model for predicting travel speed under severe weather conditions [13]. Alkheder et al. used an artificial neural network (ANN) to predict the injury severity of 5,973 traffic accidents that occurred in Abu Dhabi during the six years from 2008 to 2013 [14]. Song et al. introduced an adaptive variation operator in the particle swarm algorithm to address the shortcomings of BP neural network prediction in terms of local minima and slow convergence speed. The improved particle swarm algorithm was proposed to optimise the BP neural network for short-time traffic flow prediction, which was experimentally demonstrated to have better nonlinear fitting ability and higher prediction accuracy [15]. Tan et al. used a combined model of ARIMA and artificial neural network for traffic flow prediction [16]. However, the amount of data required for neural network learning is extremely large, and data acquisition becomes a problem, while the theoretical support for neural network learning is relatively insufficient. In contrast, the SVM model requires less data and has sufficient theoretical support, which has become the focus of research in recent years.

Vladimir introduced the basic theory of the SVM model [17]. Researchers in the field of transportation have studied many aspects of the SVM, including traffic flow speed, accident duration and congestion prediction, etc. [18–20]. Yang et al. proposed a short-time traffic flow prediction model based on support vector machines. The experimental data results showed that the prediction accuracy and generalisation ability of the SVM prediction model were better than the BP neural network model [21]. Zhang et al. validated the SVM model and explored the relationship between its prediction accuracy and independent variables using actual data from Dalian, China. To improve the model prediction accuracy, many scholars have improved the SVM model [22]. Feng et al. proposed a short-term traffic prediction algorithm based on adaptive multicore support vector machine (AMSVM) with spatial and temporal correlation, and the results showed that the algorithm outperformed existing algorithms [23]. Lippi et al. proposed a seasonal support vector machine model that was applied to congestion traffic volume prediction [24]. Sun et al. used a combination of wavelet model and SVM model to predict different types of passenger traffic in rail transportation [25].

Related studies have shown that SVM model has higher prediction accuracy than other models. According to previous studies, it is found that SVM model has better performance in daily traffic prediction, however, its prediction accuracy tends to be lower when it comes to holidays, which is also due to the basic characteristics of the SVM model. Therefore, an improved SVM model is constructed in this paper to improve the daily traffic prediction accuracy.

2. DATA COLLECTION AND ANALYSIS

2.1 Data collection and pre-processing

The data in this paper come from a toll station in Heilongjiang province, China, which records vehicle arrival data from 1 January 2018 to 31 December 2019, including vehicle type, entry time, body colour, cost, and so on. We cleaned up the involved duplicate, missing and space value data. In general, the missing data mechanism includes three cases such as all random missing, partially random missing and non-random but not counted missing. The traffic volume data involved in this case includes two cases of full random loss and partial random missing, where the random loss part needs to be supplemented by the number, and in this paper, the missing part is supplemented by the missing data according to the multiple calculations of variable complementary data. The data are shown in *Table 1*. We find that the number of passenger cars is much higher than the number of lorries during holidays.

2.2 Data analysis

After pre-processing the data, we started to analyse the data characteristics. Firstly, we conducted a comparative analysis of quarterly, monthly, weekly and daily traffic volume changes in 2018 and 2019. Then, the initial indicator system of impact factors was constructed with reference to the time series method and available data. Finally, the importance of each influencing factor was determined by the random forest method, the influencing factors with low importance were eliminated, and the SVM model independent variables were determined. The specific process is shown in *Figure 1*.

We performed statistical analysis of the toll station data for 2018 and 2019. The statistics obtained the quarterly, monthly, weekly, and daily traffic volume trends for two years. From *Figure 2a*, we can see that the traffic volume varies greatly from quarter to quarter each year, while the trends of traffic volume in different seasons in both years are similar. The traffic volume in the first and second quarters of 2019 is higher than the traffic volume in the first and second quarters of 2018, but its traffic volume in the third and fourth quarters of 2019 is lower than the traffic volume in the third and fourth quarters of 2018 due to the impact of highway construction in August and September 2019. Similarly, in *Figure 2b-2d*, we can see that its monthly, weekly and daily traffic volume trends in 2018 and 2019 are basically the same.

To further support this view, we used the R software to calculate the correlation coefficient of daily traffic data in 2018 and 2019; the result was 0.024. According to the correlation coefficient value domain level, the two are in a very weak or uncorrelated state. After analysis we found that the main reason for this is that the road construction in August and September 2019 led to its traffic volume of 0, which greatly reduced the correlation between the two. Therefore, we calculated the correlation coefficient of daily traffic volume in January–July 2018 and 2019, and the result was 0.498, which indicates a moderate correlation according to the correlation

Table 1 – Sample of vehicle arrival table

Arrival time	License plate number	Passenger and cargo category	Vehicle type	Number of axes	Fee amount	Total weight	Entrance Station
2018-01-01 07:00:09	HEI******	Lorry	5	6	14	31600	Acheng Station
2018-01-01 07:01:18	JI******	Passenger car	1	2	5	2500	Acheng Station
2018-01-01 07:02:41	HEI******	Passenger car	1	2	5	1500	Acheng Station
2018-01-01 07:02:57	HEI******	Passenger car	1	2	5	1500	Acheng Station
2018-01-01 07:03:11	HEI******	Passenger car	1	2	5	1500	Acheng Station
2018-01-01 07:03:33	HEI******	Passenger car	1	2	0	1500	Yagou Station
2018-01-01 07:03:57	HEI******	Passenger car	1	2	5	1500	Acheng Station
2018-01-01 07:05:32	HEI******	Passenger car	1	2	15	1500	Harbin Station
2018-01-01 07:06:04	MENG******	Passenger car	1	2	5	1600	Acheng Station
2018-01-01 07:10:31	HEI******	Lorry	5	6	10	17100	Acheng Station
......	......	......	......	......	......	......	......

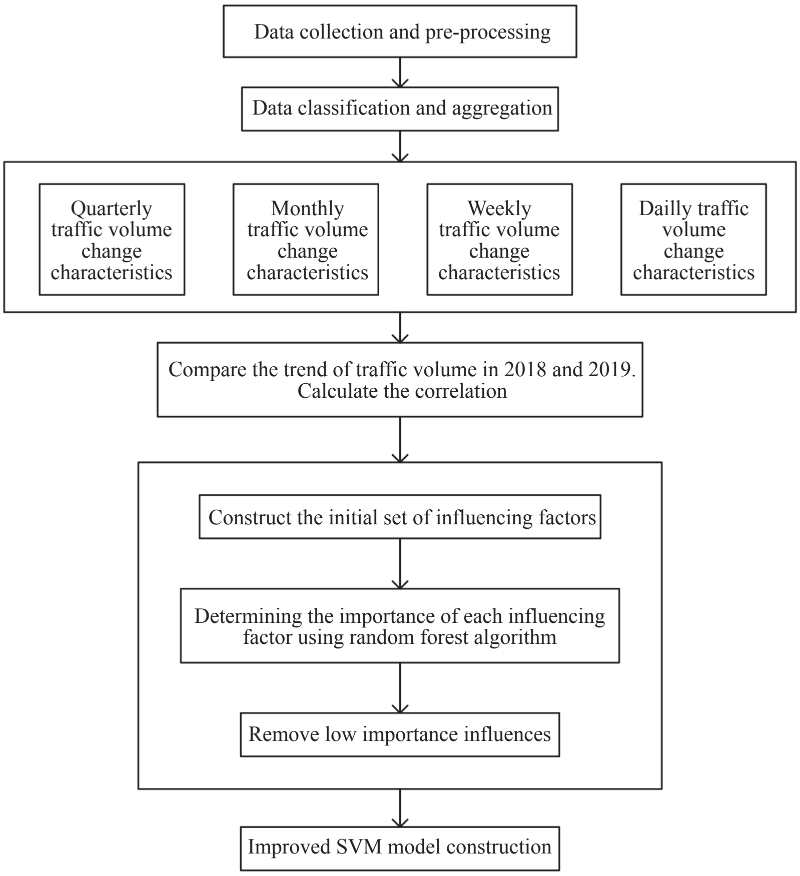
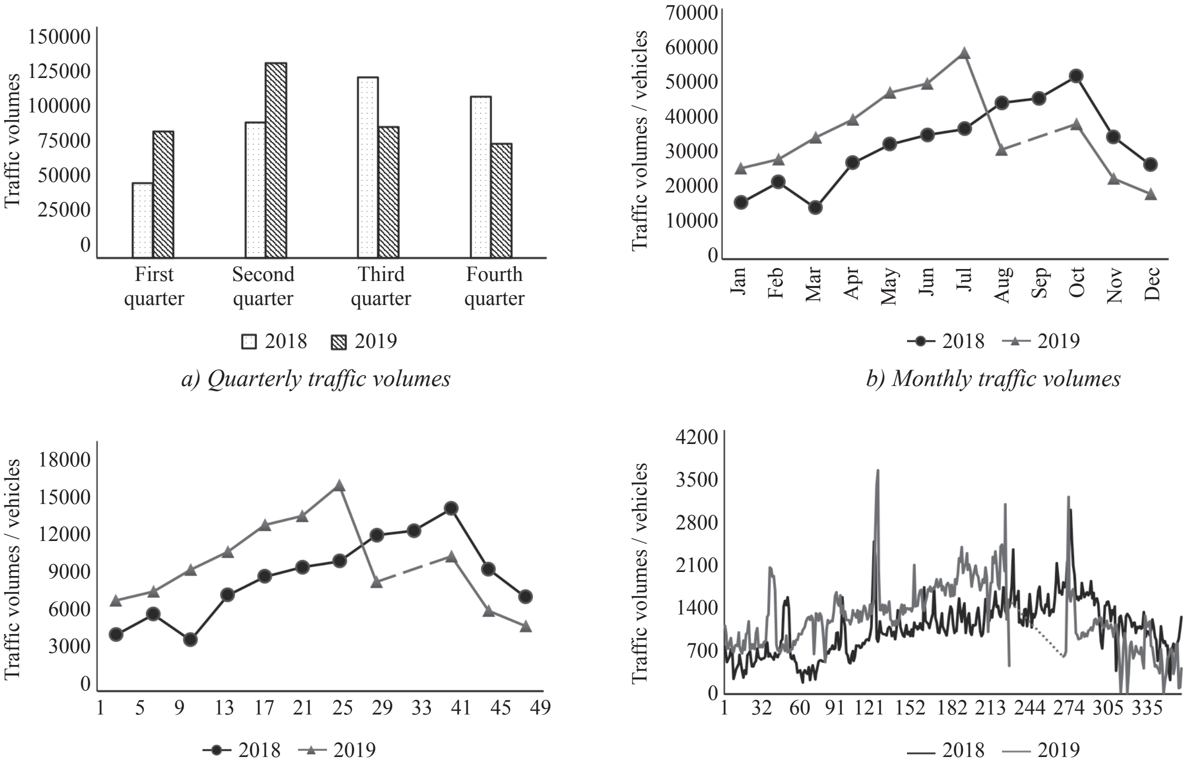

Data collection and pre-processing
Data classification and aggregation
Quarterly traffic volume change characteristics
Monthly traffic volume change characteristics
Weekly traffic volume change characteristics
Dailly traffic volume change characteristics
Compare the trend of traffic volume in 2018 and 2019. Calculate the correlation
Construct the initial set of influencing factors
Determining the importance of each influencing factor using random forest algorithm
Remove low importance influences
Improved SVM model construction

Figure 1 – Data collection and analysis



a) Quarterly traffic volumes

b) Monthly traffic volumes

c) Weekly traffic volumes

d) Quarterly traffic volumes

Figure 2 – Comparison of traffic volumes in 2018 and 2019

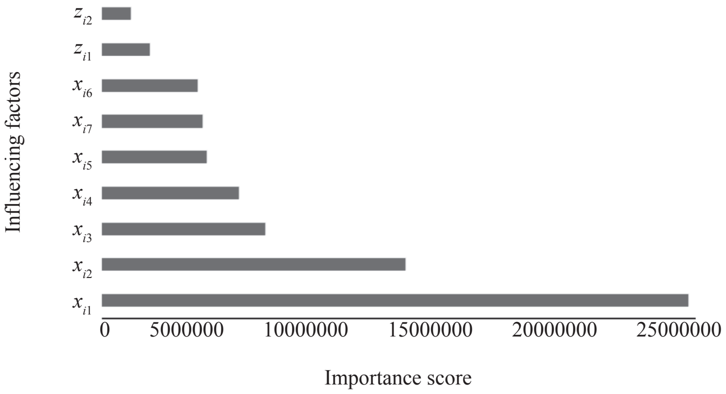


Figure 3 – Importance of influencing factors

coefficient level. It indicates that the distribution trend of daily traffic volume between the two years is close without the influence of road construction.

Referring to the time series method and combining the available data, we use the traffic volume of the seven days before the forecast day, the day of the month in which the forecast day is located and the day of the week in which the forecast day is located as input variables. The forecast day traffic volume is used as the output variable of traffic volume forecast. The initial indicator system is established.

$$y_i = g(x_{i1},x_{i2},x_{i3},x_{i4},x_{i5},x_{i6},x_{i7},z_{i1},z_{i2}) \quad (1)$$

where i represents the forecast day as the i-th day of the year; x_{ij} represents the traffic volume on the j-th day before the forecast day, j=1, 2…,7; z_{i1} represents the forecast day as the day of the month; z_{i2} represents the forecast day as the day of the week; y_i represents the real traffic volume on forecast day.

The initial index system used to build the prediction model is prone to over-fitting and may not always result in optimal prediction accuracy. Therefore, the influencing factors need to be analysed and screened to eliminate low impact indicators. After screening the initial feature indicators, the optimal combination of indicators obtained will be directly used as the input variables of the later model to improve the prediction accuracy of support vector machine modelling.

R software is a language and operating environment for statistical analysis and drawing. R is a free and open-source software belonging to the GNU system. It is an excellent tool for statistical calculations and statistical graphics [26]. Random forest algorithm is an important Bagging-based ensemble learning method, which can be used for classification, regression and other problems. We use the random forest algorithm to determine the importance of each influencing factor in random forest regression. The package 'randomForset' in R software is integrated according to the random forest algorithm, which is simple and convenient. To import this package, we only need to enter the number of sample predictors at each split node (mtry) and the number of trees (ntree) to get the fitting of the regression prediction model and the importance of influencing factors [27]. It is estimated that the fitting of the model is highest when mtry is 3 and ntree is 1000. Of course, as the value of ntree increases, the fitting of the model will still increase, but the growth rate is very slow.

As can be seen from *Figure 3*, the top two influencing factors in order of importance are x_{i1} and x_{i2}whose importance ratings are much higher than those of other influencing factors. Influencing factors x_{i3},x_{i4},x_{i5},x_{i7} and x_{i6} are in the 3rd–7th positions, their ratings are more relatively stable. Influence factors z_{i1} and z_{i2} were rated low, so they were excluded. Therefore, the daily traffic volume for 7 days before the forecast day is used as the input to the SVM model, and the forecast day traffic volume is used as the output. Based on this, the learning of the SVM model and traffic volume forecasting are performed.

3. MODEL CONSTRUCTION

3.1 Introduction of SVM model

Support vector machine (SVM) is a machine learning method based on statistical learning theory, which is mainly used to deal with classification and regression problems and can be extended to fields and disciplines such as prediction and comprehensive evaluation, etc. The solution of the

SVM always achieves the global optimal solution without being limited to local minima and shows strong resistance to overfitting problems and high generalisation performance. Unlike neural network algorithms, the SVM has very strong theoretical support. The mechanism of the SVM is to find an optimal classification hyperplane that satisfies the classification requirements, so that the hyperplane can maximise the blank area on both sides of the hyperplane while ensuring the classification accuracy. Theoretically, support vector machines can achieve optimal classification of linearly separable data.

Support Vector Regression (SVR) is an extension and application of the SVM to the regression estimation problem. For nonlinear regression problems, the basic idea is to introduce a kernel function to transform the problem into a linear regression problem in a high-dimensional space (Hilbert space) to construct a decision function. The basic principle of applying the SVR for traffic volume regression forecasting is as follows.

Given a sample (x_1,y_1), (x_2,y_2), …, (x_l,y_l), $(x_i \in X \in R^n, y_i \in Y \subset R)$, SVR uses a nonlinear mapping φ to map x into a high-dimensional feature space H. A linear approximation is performed in this space to find the mapping function so that we can get a better approximation for the given data sample. According to the statistical learning theory [28], we can obtain the following functions.

$$f(x) = w\varphi(x) + b \tag{2}$$

Regression can be defined as a risk minimisation problem for a loss function. The optimal regression function is the minimum and regularised universal function Q under certain constraints.

$$Q = \underbrace{\min}_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)) \tag{3}$$

where w is a standard vector. The first term is named the regularisation term to flatten the function and improve the generalisation ability of the function; the second term is named the empirical risk generic function, which can be determined by different loss functions; C is used to balance the relationship between the structural and empirical risks (C>0). ε is the width of the interval band, which can be selected according to actual needs and L_ε is the ε-insensitive loss function.

$$L_\varepsilon(y_i, f(x_i)) = \max(0, |y_i - f(x_i)| - \varepsilon) \tag{4}$$

Introduce the slack variables ξ_i(>0) and $\hat{\xi}_i$(>0), then

$$\begin{aligned} Q = &\underbrace{\min}_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) \\ \text{s.t.} \quad & f(x_i) - y_i \le \varepsilon + \xi_i \\ & y_i - f(x_i) \le \varepsilon + \hat{\xi}_i \\ & \xi_i \ge 0, \ \hat{\xi}_i \ge 0, \ i = 1,2,\dots,l \end{aligned} \tag{5}$$

By introducing the Lagrange multipliers $\mu_i \ge 0$, $\hat{\mu}_i \ge 0$, $\alpha_i \ge 0$, $\hat{\alpha}_i \ge 0$, the Lagrange function is obtained:

$$\begin{aligned} &L(w,b,\alpha,\hat{\alpha},\xi_i,\hat{\xi}_i,\mu,\hat{\mu}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) \\ &-\sum_{i=1}^{l}\mu_i\xi_i + \sum_{i=1}^{l}\alpha_i(f(x_i) - y_i - \varepsilon - \xi_i) \\ &+\sum_{i=1}^{l}\hat{\alpha}_i(y_i - f(x_i) - \varepsilon - \hat{\xi}_i) \end{aligned} \tag{6}$$

Let $L(w,b,\alpha, \hat{\alpha},\xi_i,\hat{\xi}_i,\mu,\hat{\mu})$ find the partial derivative of $w,b,\xi_i,\hat{\xi}_i$. Make its partial lead to 0. Then by bringing it into the original formula, we can get the dual problem of the SVR:

$$w = \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i)\varphi(x_i) \tag{7}$$

$$\begin{aligned} Q = &\underbrace{\max}_{\alpha,\hat{\alpha}} \sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon\sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) \\ &-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)K(x_i,x_j) \\ \text{s.t.} \quad &\sum_{j=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0 \\ &0 \le \hat{\alpha}_i,\alpha_i \le \frac{c}{l} \end{aligned} \tag{8}$$

The above process needs to meet the Karush–Kuhn–Tucker (KKT) conditions, namely:

$$\begin{cases} \alpha_i(f(x_i) - y_i - \varepsilon - \xi_i) = 0 \\ \hat{\alpha}_i(y_i - f(x_i) - \varepsilon - \hat{\xi}_i) = 0 \\ \alpha_i\hat{\alpha}_i = 0, \ \ \xi_i,\hat{\xi}_i = 0 \\ (C - \alpha_i)\xi_i = 0, \ (C - \hat{\alpha}_i)\hat{\xi}_i = 0 \end{cases} \tag{9}$$

Taking w into *Equation 2*, the SVR can be expressed as:

$$f(x) = \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i)k(x_i,x) + b \tag{10}$$

where $k(x_i,x)$ is the kernel function and is equal to the inner product of $\varphi(x_i)$ and $\varphi(x)$.

3.2 SVM basic model prediction

In this section, we use the traffic volume data from 1 January 2018 to 30 September 2018 as the learning set, and the traffic volume data from 1 October 2018 to 31 December 2018 as the validation set. The traffic volume data from 1 January to 31 July 2019 is used as the test set. The forecasting process is mainly divided into the following steps:

First, select sample data and perform noise reduction on the data to construct the learning set. According to the analysis in Section 2, it is clear the importance of the traffic volume in the previous 7 days is relatively high. Therefore, the traffic volume data x_{i1},x_{i2},x_{i3},x_{i4},x_{i5},x_{i6} and x_{i7} before the current day of 2018 January to July were used as independent variables x_i, and the traffic volume of the current day y_i was used as the dependent variable. Second, based on the validation set, penalty factors, loss functions, etc. are determined. In this study, a linear kernel is used. After experimental analysis, when the penalty factor C value is less than 1, the model accuracy becomes more accurate with the increase of C value, and when the C value is larger than 1, the increase of its accuracy is extremely small, so C=1.0 is chosen. Third, the optimisation problem is constructed and solved for the sample and predicted values as input values. Fourth, the optimal prediction learning set function is obtained and the existing samples are used to predict the daily traffic volume Q_s at toll stations from January to July 2019. Fifth, the prediction error is calculated. The regression model evaluation index R^2, mean absolute error (MAE), mean relative error (MAPE), root mean square error (RMSE) and mean square relative error (MASPE) are used for evaluation.

We used the python software to program the basic SVM model to predict the validation set. After repeated adjustments, the accuracy of the validation set of the basic SVM model is 0.7762. Then predictions on the test set were made, with the prediction results shown in *Figure 5b*, and the model's regression evaluation index R^2 is 0.7582.

After removing the outliers, the SVM model is tested for the normality of the residuals. The result is p=0.2841, which satisfies the normality assumption of residuals. *Figure 4a* is the normal P-P diagram of the SVM model.

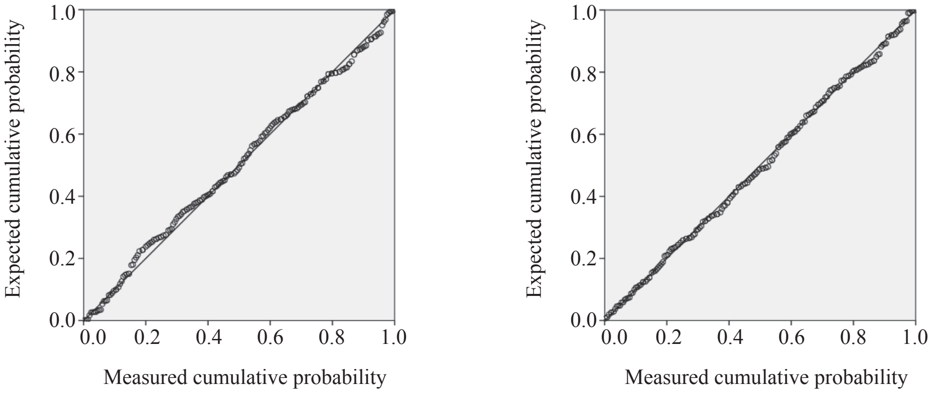
From *Figure 5b*, we can see that the SVM is relatively accurate in forecasting daily traffic volume at toll stations, which confirms the feasibility of the SVM model. But we can also see that the prediction accuracy of the SVM model is relatively poor at certain times when traffic volume peaks, so we have to find these time points where the SVM model fits poorly. We found that the locations where the fit is poor are usually located at holidays, which is consistent with the characteristic of the SVM model that it cannot accurately predict traffic volumes at anomalies.

Therefore, we started to think about the impact of holidays on the prediction accuracy of the SVM model. Considering that holidays not only affect the traffic travel behaviour on the day of the holiday, but also affect the traffic travel behaviour before and after the holiday, and in addition, the independent variable used for the SVM prediction is the traffic volume 7 days before the prediction day. So, the traffic volume on the day before the holiday, the 8 days after the holiday and during the holiday are rounded off. The SVM model is then applied to forecast the traffic volume on non-holiday days. The holidays involved from January to July 2019 include New Year's Day, Spring Festival, Qingming Festival, Labour Day, and Dragon Boat Festival, totalling 62 days. After removing these holidays, the R^2 is 0.8435. The SVM model is more accurate in forecasting traffic volume at toll stations, and the fitting effect is improved by 8.53%. Therefore, improving the prediction accuracy of holiday traffic volume becomes a key issue affecting the overall prediction accuracy of the SVM model.

3.3 The improved SVM model

Holiday traffic forecasting is a major obstacle to forecast accuracy. To improve the overall forecasting accuracy, the most important thing is to improve the holiday traffic volume forecasting accuracy. In this section, we propose a method of combining the SVM model and conversion coefficient.

First, from the previous subsection, we can see that the prediction accuracy of the non-holiday SVM model meets the prediction requirements. Therefore, we think it is in line with the requirements to predict the traffic volume on holidays based on the traffic volume 7 days before the holiday without considering the holiday factor. Similarly, it is also accurate to forecast the traffic volume after the holiday by using the holiday traffic volume without considering the holiday factor as the independent variable. This method is defined as Q_{fj}, j=0,1,2…, with j=1 representing the first day of the holiday.



a) Normal P-P diagram of SVM model

b) Normal P-P diagram of improved SVM model

Figure 4 – Normality test of residuals

Second, we propose a conversion factor that corrects the traffic volume Q_{fj} obtained from the forecast. Previous studies have used the ratio of the daily traffic volume within the travel time of each holiday in the base year to the average traffic volume of the day of the week to which it belongs in that year as the basis for correction. However, since the toll station studied in this paper is located on the territory of the Heilongjiang province in China, combined with the analysis of the quarterly change characteristics of the traffic volume in Section II of this paper, it can be seen that the distribution of traffic volume corresponding to different quarters varies greatly. So, correcting the holidays belonging to different quarters according to a uniform weekday ratio will affect the prediction accuracy. Therefore, this paper uses the ratio of different holidays to the average traffic volume of the day of the week within the respective quarter to which they belong as the conversion factor. The conversion factor within the holiday travel time is calculated as follows.

$$\theta_{jk} = \frac{q_h}{q_{ka}} \tag{11}$$

where θ_{jk} represents the traffic conversion coefficient of different holidays, k represents the quarter of the year in which the holiday belongs to, j represents the j-th day of the holiday, q_h represents the daily traffic volume of the holiday in the previous year, q_{ka} represents the average traffic volume on Sundays in the quarter of the previous year except for holidays.

Third, considering the influence of holiday traffic volume on post-holiday traffic volume, the weighted value of the holiday traffic volume calculated by the original SVM model and the traffic volume corrected by the conversion factor is used as the final forecast value. From the above analysis, we can find that the traffic volume prediction gap between the SVM basic model on the day before the holiday and the first day of the holiday is large. At this time, we can choose the model modified by the conversion coefficient to give a larger weight. With the passage of time, the impact of the holiday traffic volume as the independent variable on the SVM basic model gradually increases, and the prediction accuracy of the SVM basic model also gradually improves [29]. So, give the SVM a larger weight. As shown in *Equation 12*.

$$Q_j = \mu Q_{fj} \theta_{jk} + \sigma Q_s \tag{12}$$

where μ, σ are weight coefficients, others as above.

We propose *Equation 13* after considering the number of days affected by holidays. We choose the number 7 as the maximum value. On the one hand, there are 7 days in a week. According to previous studies, the weekly characteristics of traffic volume have obvious regularities [30]. On the other hand, in section 2.2 “data analysis”, the largest impact on the traffic volume of the day is the first 7 variables. As the number of days increases, the increase in prediction accuracy becomes smaller and smaller.

$$Q_j = \frac{7-j}{7} Q_{fj} \theta_{jk} + \frac{j}{7} Q_s \tag{13}$$

We use the python software to program the improved SVM model to predict the validation set. After repeated adjustments, the accuracy of the validation set of the improved SVM model is 0.8376. Then we make predictions on the test set. The prediction results are shown in *Figure 5b* and the model's regression evaluation index R^2 is 0.8340.

After removing the outliers, the improved SVM model is tested for the normality of the residuals. The result is p=0.5404, which satisfies the normality assumption of residuals. *Figure 4b* is the normal P-P diagram of the improved SVM model.

4. MODEL COMPARSION ANALYSIS

4.1 Introduction of the GBDT model

To verify the accuracy of the model, here we will compare the prediction results of the GBDT (Gradient Boosting Decision Tree) model with the improved SVM model.

GBDT is an integrated algorithm that integrates three algorithms: Boosting, Gradient and Decision Tree. The formation process of the GBDT is based on the weak correlation between decision trees in random forests. The concept of “lifting” is proposed. To make the lifting algorithm easier and more convenient when solving the loss function, the gradient boosting algorithm is finally proposed, namely the GBDT. GBDT is widely used in traffic volume forecasting. The GBDT models considering the neighbouring traffic condition tend to outperform the traditional simple temporal prediction model [31]. Lin and Zhou proposed a multi-feature GBDT model for toll station traffic flow prediction and then compared it with the BP neural network model to prove the effectiveness and feasibility of the GBDT model [32]. The model is as follows:

First, initialise a weak classifier:

$$f_0(x) = arg\ min_c \sum_{i=1}^{n} L(y_i,c) \tag{14}$$

Second, for m=1,2,…,M:

a) For i=1,2,…,N, calculate:

$$r_{mi} = -\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{15}$$

b) Fit a regression tree to r_{mi} to get the leaf node area R_{mj} of the mth tree, j=1,2,...,J.

c) For j=1,2,...,J, calculate:

$$c_{mi} = arg\ min_c \sum_{x_i \in R_{mj}} L(y_i,f_{m-1}(x_i)+c) \tag{16}$$

d) renew:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \tag{17}$$

In the last, get the GBDT:

$$\hat{f}(x) = f_M(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \tag{18}$$

Then, we use the same learning set, validation set and test set as the SVM model, and use the python software to predict traffic volume.

4.2 Model comparison analysis

The SVM basic model has a high accuracy for non-holiday traffic prediction, so the non-holiday traffic is predicted by the SVM basic model with the holidays removed. However, the SVM basic model has a large error for holiday traffic volume forecast. So, we propose a prediction method that combines the correction coefficient with the basic SVM model. To illustrate the importance of the top influencing factors x_{i1} (x1 SVM) and x_{i2} (x2 SVM), we choose each of two factors and both two factors (x12 SVM) to predict non-holiday and holiday traffic, whose results is shown in *Figure 5a*. From the figure we can find that using the two top influencing factors can improve prediction accuracy effectively.

The prediction results of the SVM model and the improved SVM model are shown in *Figure 5b*. From the figure, we can find that the improved SVM model has a smoother traffic prediction curve for non-holiday traffic and is close to the true value. Holiday traffic is often the peak of the daily traffic variation curve. The holiday traffic prediction results of the basic SVM model tend to be flatter, which causes larger errors. The improved SVM model is a weighted average of the basic SVM model and the SVM model with the effect of holidays removed, and the results are closer to the true value. Of course, we also find that there are two outliers, point 95 and point 150, which may be caused by the inaccuracy of the seasonal variation coefficient due to the small number of statistics.

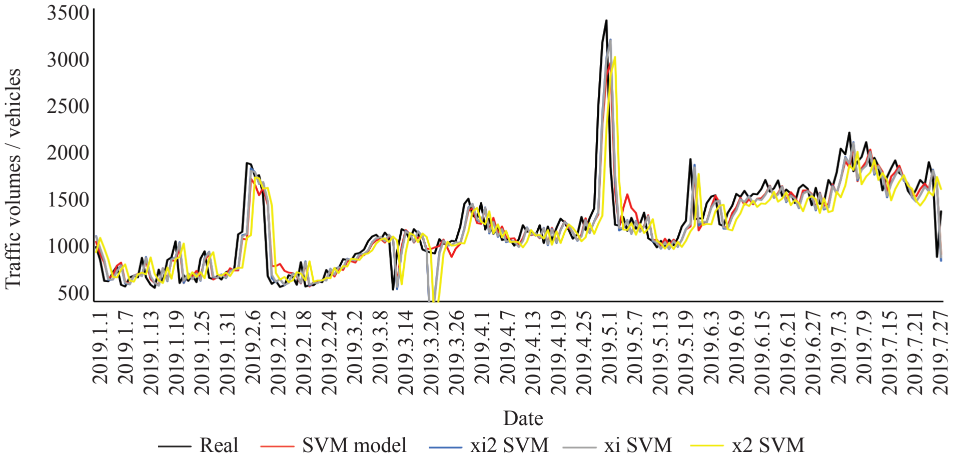
To verify the validity of the improved SVM model, it is necessary to construct an evaluation index system for the validity of the model prediction results. Common model evaluation indexes include regression coefficient of determination (R^2), mean absolute error (*MAE*), mean relative error (*MAPE*), root mean square error (*RMSE*) and mean square relative error (*MSPE*). The calculation formula of each index is as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

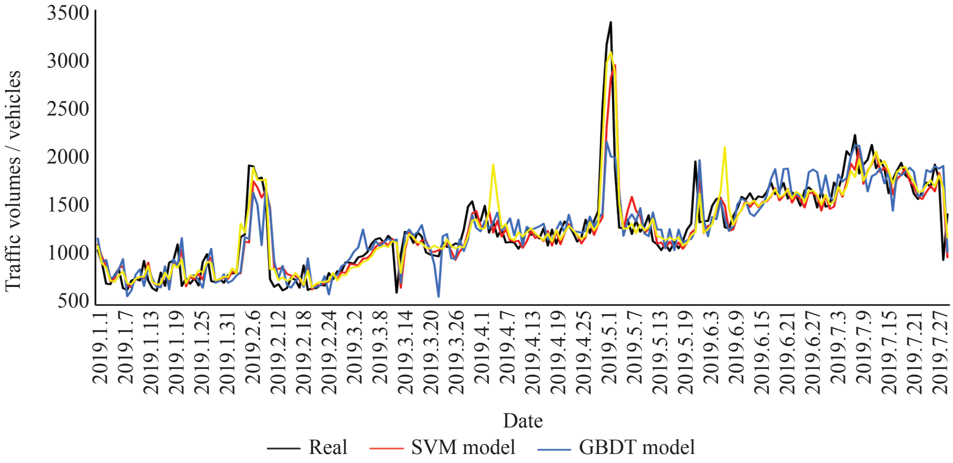
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{20}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{22}$$



a) Prediction results with different factors



b) Prediction results

Figure 5 – Improved SVM model

$$MSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2} \tag{23}$$

where y_i is the actual traffic volume; $\hat{y}_i$ is the predicted traffic volume; n is the total number of predicted days.

The prediction accuracy of the improved SVM is compared with factors x_{i1}, x_{i2} and it shows that their R^2 is 0.7307, 0.4854 and 0.7316.

The comparison of the prediction accuracy of the SVM model, GBDT model and the improved SVM model is shown in *Table 2*. From the table, we can see that the R^2 of the basic SVM model and GBDT model are 0.7582 and 0.6858, and the R^2 of the improved SVM model is 0.8340. In comparison, the prediction accuracy of the improved SVM model is 7.58% higher than that of the basic SVM

Table 2 – Prediction accuracy comparison

Evaluation indexes	SVM model	GBDT model	Improved SVM model
R^2	0.7582	0.6858	0.8340
MAE	147.47	167.98	128.06
MAPE [%]	11.53	13.34	10.26
RMSE	232.15	264.62	192.37
MSPE [%]	17.99	20.63	15.47

model and 14.82% higher than the GBDT model. In addition, the evaluation indexes MAE, MAPE, RMSE and MSPE of the improved SVM model are also higher than those of the basic SVM model and the GBDT model.

5. CONCLUSION

To improve the accuracy of daily traffic forecasting of a toll station, an improved SVM model is proposed in this study. First, the data from 1 January 2018 to 31 July 2019 at a toll station in Heilongjiang Province, China, were processed and analysed. After determining that the two years of daily traffic data are correlated, the input variables of the SVM model are filtered using the random forest algorithm with reference to the time series method and the available data. Model construction was then performed: when predicting non-holiday traffic volumes, the model was chosen to remove the effect of holiday traffic volumes. When predicting holiday traffic volumes, we proposed the improved SVM model which combines the seasonal correction factor, the SVM model that ignores the effect of holidays and the basic SVM prediction model. The result showed that the prediction accuracy of the improved SVM model is 0.8340, which is 7.58% higher than that of the basic SVM model for predicting the daily traffic volume at toll stations. Finally, the improved SVM model was compared with the SVM model and GBDT model to verify the superiority of the improved SVM model.

It is well known that traffic forecasting is a prerequisite for road construction and traffic management and control. Convenient and efficient traffic forecasting can bring real economic benefits to society and a more comfortable environment for human travel. In this paper, an improved SVM model is proposed for the daily traffic volume at toll stations. However, this model is not only applicable to toll station daily traffic forecasting, but its application can be extended to hourly traffic forecasting, roadway traffic forecasting, intersection traffic forecasting, etc.

Of course, there are certain shortcomings in this paper. The improved SVM model uses a relatively simple method of determining weight coefficients in predicting holiday traffic in this paper, and thus 2 outliers that deviate from the true value appear. With the advent of the era of big data, correlations between data are more easily uncovered. It is worth being considered how to determine the weight coefficients for traffic volume prediction by a large amount of historical data.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (91846301); Heilongjiang Philosophy and Social Science Research Planning Project (20GLC204); The Project of Jilin Provincial Department of Transportation (2019-1-11).

胡晓伟， 博士[1]
邮箱: xiaowei_hu@hit.edu.cn
肖永智，硕士[1]
邮箱: xiaoyongzhi6@163.com
王天麟， 硕士[1]
邮箱: 2832854461@qq.com
杨璐，硕士[1],（通讯作者）
邮箱: 18846075246@163.com
唐鹏程，高级工程师[2]
邮箱: tangpc@foxmail.com
[1] 哈尔滨工业大学，交通科学与工程学院
中国黑龙江省哈尔滨市南岗区黄河路73号
邮编: 150090
[2] 中交远洲交通科技集团有限公司,
中国河北省石家庄高新区槐安东路351号
邮编: 10000

高速公路收费站节假日交通量预测模型：一种支持向量机模型

摘要

支持向量机*(SVM)*模型在预测收费站日交通量方面有较好的性能，但无法准确预测节假日交通量。因此，本文提出了一种改进的*SVM*模型。本文以黑龙江某收费站为例，以日交通量作为学习集，以当日交通量作为因变量，以前*7*天交通量作为自变量用于模型学习。本文发现基本的 *SVM* 模型不足以准确预测假期期间的交通量。为提高模型精度，本文首先采用*SVM*模型对非节假日交通量进行预测，并提出一种利用季度换算系数的预测方法，结合基础的*SVM*模型构建改进的*SVM*模型。预测结果表明，本文改进的*SVM*模型能够有效提高精度，精度优于基础*SVM*和*GBD*T模型，证明了*SVM*改进模型的可行性。

关键词

交通量；预测；支持向量机；假期；
季度换算系数；高速公路收费站

REFERENCES

[1] Shao CF. [*Traffic Planning*]. Beijing: China Railway Publishing; 2004. Chinese.

[2] Box, et al. *Time series analysis: Forecasting and control*. Hoboken: John Wiley & Sons Publishing; 2015.

[3] Rui, et al. [Prediction method of highway passenger transportation volume based on exponential smoothing method and Markov model]. *Journal of Traffic and Transportation Engineering*. 2013;13(04): 87-93. Chinese.

[4] Bezuglov A, Comert G. Short-term freeway traffic parameter prediction: Application of grey system theory models. *Expert Systems with Applications*. 2016;62: 284-292. doi: 10.1016/j.eswa.2016.06.032.

[5] Gu Y, Han Y, Fang X. [Research on passenger flow prediction method of bus hubs based on ARMA model]. *Journal of Transport Information and Safety*. 2011;29(02): 5-9. Chinese.

[6] Nihan NL, Holmesland KO. Use of the box and Jenkins time series technique in traffic forecasting. *Transportation*. 1980;9(2): 125-143. doi: 10.1007/BF00167127.

[7] Williams BM, Hoel LA. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering*. 2003;129(6): 664-672. doi: 10.1061/(ASCE)0733-947X(2003)129:6(664).

[8] Giraka O, Selvaraj VK. Short-term prediction of intersection turning volume using seasonal ARIMA model. *Transportation Letters*. 2020;12(7): 483-490. doi: 10.1080/19427867.2019.1645476.

[9] Cai WT, Peng Y, Chen QJ. [Air transport passenger volume forecasting based on multiple regression model]. *Aeronautical Computing Technique*. 2019;49(04): 50-53+58. Chinese.

[10] Hongyu Sun, et al. Use of local linear regression model for short-term traffic forecasting. *Transportation Research Record*. 2013;1836(1): 143-150. doi: 10.3141/1836-18.

[11] Yun SY, et al. A Performance evaluation of neural network models in traffic volume forecasting. *Mathematical and Computer Modelling*. 1998;27(9): 293-310. doi: 10.1016/S0895-7177(98)00065-X.

[12] Dia H. An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research*. 200;131(2): 253-261. doi: 10.1016/S0377-2217(00)00125-9.

[13] Huang SH. *An application of neural network on traffic speed prediction under adverse weather condition*. State of Wisconsin: The University of Wisconsin—Madison Publishing; 2003.

[14] Alkheder S, Taamneh M, Taamneh S. Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*. 2017; 36(1): 100-108. doi: 10.1002/for.2425.

[15] Song L, Lijun L, Man Z. Prediction for short-term traffic flow based on modified PSO optimized BP neural network. *Systems Engineering-Theory & Practice*. 2012;32(9): 2045-2049. doi: 10.12011/1000-6788(2012)9-2045.

[16] Tan M-C, Feng L-B, Xu J-M. Traffic flow prediction based on hybrid ARIMA and ANN model. *China Journal of Highway and Transport*. 2007;4(86): 118-121.

[17] Vapnik V. *The nature of statistical learning theory.* Berlin: Springer Science & Business Media Publishing; 2013.

[18] Yao B, et al. Short-term traffic speed prediction for an urban corridor. *Computer-Aided Civil and Infrastructure Engineering*. 2017; 32(2): 154-169.

[19] Vanajakshi L, Rilett LR. A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed. *IEEE Intelligent Vehicles Symposium*. 2004;26(3): 194-199. doi: 10.1109/IVS.2004.1336380.

[20] Wang LL, Ngan HYT, Yung NHC. Automatic incident classification for large-scale traffic data by adaptive boosting SVM. *Information Sciences*. 2018;467: 59-73. doi: 10.1016/j.ins.2018.07.044.

[21] Yang Z-S, Wang Y, Guan Q. Short-time traffic flow prediction method based on support vector machine method. *Journal of Jilin University (Engineering and Technology Edition*). 2006;06: 881-884.

[22] Zhang MH, et al. Accurate multisteps traffic flow prediction based on SVM. *Mathematical Problems in Engineering*. 2013; 11-23. doi: 10.1155/2013/418303.

[23] Feng X, et al. Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*. 2018;20(6): 2001-2013. doi: 10.1109/TITS.2018.2854913.

[24] Lippi M, Bertini M, Frasconi P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*. 2013;14(2): 871-882. doi: 10.1109/TITS.2013.2247040.

[25] Sun Y, Leng B, Guan W. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Neurocomputing*. 2015;166: 109-121. doi: 10.1016/j.neucom.2015.03.085.

[26] Kabacoff RI. *R in action: Data analysis and graphics with R. Getting started.* New York: Simon and Schuster Publishing; 2015.

[27] Team RC. *R: A Language and Environment for Statistical Computing*. 2013.

[28] Li H. [*Statistical learning methods. Support vector machines*]. Tsinghua University Press; 2012. Chinese.

[29] He SJ, et al. Weight analysis of each influence factor of the green tide disaster based on SVM. *China Environmental Science*. 2015;11: 3431-3436.

[30] Wang W, Guo XC. [*Traffic Engineering*]. Nanjing: Southeast University Press Publishing; 2000. Chinese.

[31] Yang S, et al. Ensemble learning for short-term traffic prediction based on gradient boosting machine. *Journal of Sensors*. 2017; 2017.

[32] Lin P, Zhou N. Short-term traffic flow forecast of toll station based on multi-feature GBDT model. *Journal of Guangxi University (Natural Science Edition)*. 2018;43(03): 1192-1199.