Appl Intell (2018) 48:2429–2440
https://doi.org/10.1007/s10489-017-1095-9

A hybrid short-term traffic flow forecasting model based on time series multifractal characteristics

Hong Zhang[1,2] · Xiaoming Wang[1] · Jie Cao[2] · Minan Tang[3,4] · Yirong Guo[1]

Published online: 4 December 2017

Abstract Short-term traffic flow forecasting is a key problem in the area of intelligent transportation systems (ITS). Timely and accurate traffic state prediction is also the prerequisite of realizing proactive traffic control and dynamic traffic assignment effectively. In this paper, a new hybrid model for short-term traffic flow forecasting, which is built based on multifractal characteristics of traffic flow time series, is proposed. The hybrid model decomposes traffic flow series into four different components, namely a periodic part, a trend part, a stationary part and a volatility part, to unearth the traffic features hidden behind the data. Four parts are treated and modeled separately by using different methods, such as spectral analysis, time series and statistical volatility analysis, to further explore the underlying traffic patterns and improve forecasting accuracy. Performance of the proposed hybrid model is investigated with traffic flow data from freeway I-694 EB in the Twin Cities. The experimental results indicate that the proposed model outperforms in capturing nonlinear volatility and improving forecasting accuracy than traditional forecasting methods, especially for the multi-step ahead forecasting. Compared with the ARIMA-GARCH model, it gets an improvement of 8.23% in RMSE for one-step ahead forecasting and 10.69% for ten-step ahead forecasting. It is better than the hybrid model newly proposed in literature (Zhang et al. Transp Res Part C: Emerg Technol 43(1):65–78 2014) and gets an improvement of 1.27% in forecasting accuracy.

Keywords Hybrid model · Traffic flow · Multifractal characteristics · Short-term forecasting · Periodic regression · Volatility analysis

1 Introduction

Short-term traffic flow prediction is the foundation of traffic control and guidance. The performance of ITS largely depends on the accuracy of real-time traffic information prediction. The accurate and timely traffic state information not only allows travelers to make better travel decisions, but also benefit transportation management [1]. Because of its importance, traffic flow forecasting has generated great interest among researchers. It is always a hotspot among researchers and many forecasting methods have been proposed in literature. The existing algorithms about short-term traffic flow forecasting can be classified into three categories: (1) linear forecasting methods, including historical average method [2], time series method [3], state space method [4] and so on; (2) nonlinear forecasting methods, mainly including support vector method, neural network method [5, 6], chaotic theory method [7], nonparametric regression method [8, 9], statistical volatility analysis, pattern recognition method [10] and so on; (3) hybrid methods [11–13], which combine two or more single method and build a combined model to make full use of advantages of

✉ Hong Zhang
zhanghong@lut.cn

[1] College of Electrical & Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

[2] College of Computer & Communication, Lanzhou University of Technology, Lanzhou 730050, China

[3] College of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

[4] College of Mechanical and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China

each method and improve forecasting accuracy. Generally, linear forecasting methods provide a simple estimate for traffic state in the future and have low computing complexity and simple operation. But it can't reflect uncertainty of traffic flow variation, especially for the forecasting within 5 minutes. Nonlinear forecasting methods have higher accuracy, but they need a large amount of data to train the model, and the computing is very complex. Hybrid forecasting methods can take advantages of multiple models and provide the more comprehensive analysis and more accurate results for traffic flow forecasting. Therefore, hybrid forecasting methods will be the development trend of short-term traffic flow forecasting.

This paper focuses on developing a hybrid method that can make full use of the multi-mode information of traffic data and establish suitable models for each mode featured individually to improve accuracy of short-term traffic flow forecasting. The novelty and originality of this paper is not only the specific techniques and methods, such as spectral analysis, time series, statistical analysis theory and segmentation modeling, but also the demonstration that the forecasting model should take the dynamic characteristics and multifractal features of traffic flow into consideration. The most important contribution is that this paper provides the new idea and methodology on how to construct an appropriate forecasting model that combines multi-features analysis and multi-mode modeling and how to identify and optimize model parameters efficiently.

2 Related works

In order to mine multi-mode information behind the traffic data, many researchers have proposed hybrid models to forecast short-term traffic flow. Wei [14] first decomposed passenger data into a set of intrinsic mode function components, and used empirical mode decomposition and back-propagation neural networks to predict short-term passenger flow in metro systems. Wang [15] proposed a traffic speed forecasting hybrid model by using a wavelet function, phase space reconstruction, and support vector machine regression theory. Chen [16] compared performance of prediction models by using highway original traffic data or residual data with intraday trend removed. The experimental results indicated that the performance improved significantly with intraday trend removed. Based on spectral analysis and statistical volatility theory, Zhang [17] proposed a hybrid traffic flow forecasting model which decomposed traffic flow data from Houston into three patterns and modeled these patterns separately. The results showed promising abilities in improving the accuracy of freeway traffic flow forecasting. Pedro Lopez-Garcia [18] combined genetic algorithm (GA) and cross entropy (CE) method to predict congestion of the I5 freeway in California. The results proved that the hybrid method was more accurate than GA or CE alone for predicting short-term traffic congestion.

These studies show that a hybrid model can really improve the forecasting accuracy. Traffic flow not only exhibits periodicity and tendency, but also reveals randomness evoked by exogenous factors such as traffic accidents, weather, assembly, roadway conditions and so on. The complexity and uncertainty of traffic itself decides that traffic flow forecast cannot be solved only by a single mathematical model. It is difficult to forecast traffic flow precisely. Generally, the more the multi-mode information forecasting method incorporates, the more accurate the forecasting results. What's more, some studies have indicated that forecasting performance can be notably improved by utilizing these periodic, trend and volatile characteristics [19]. So learning and modeling periodicity, tendency and volatility under traffic data separately will enhance our understanding for traffic characteristics and improve the forecasting accuracy and reliability. This paper focuses on developing a hybrid method that can make full use of the multi-mode information of traffic data to improve the accuracy of short-term traffic flow forecasting. This paper will analyze the periodic, trend, linear as well as nonlinear characteristics of traffic flow and give explicit equations to represent the traffic data and interpret the underlining features of dynamic traffics by using spectral analysis, time series and statistical volatility theory. It will decompose the traffic data into four different components and develop a hybrid forecasting method that will model each component separately with an adaptable method.

The remaining sections of this paper are organized as follows. The third section presents the hybrid model based on time series multifractal characteristics and analyzes related theory. The fourth section presents the modeling procedure of short-term traffic flow forecasting based on the proposed hybrid model. The fifth section evaluates the performance of the proposed hybrid method. The sixth section analyzes the results and draws conclusions.

3 The hybrid model based on time series multifractal characteristics

Traffic system is a complex nonlinear system. Generally, a traffic flow series contains periodic, trend, linear and nonlinear components caused by many factors, such as specific road conditions, regular traffic demand, traffic regulations and irregular components affected by exogenous factors such as traffic incidents, weather and so on [20]. As Fig. 1 shows, a hybrid model combining multiple modeling and analysis method is built to forecast traffic. This hybrid model first decomposes the traffic flow series into

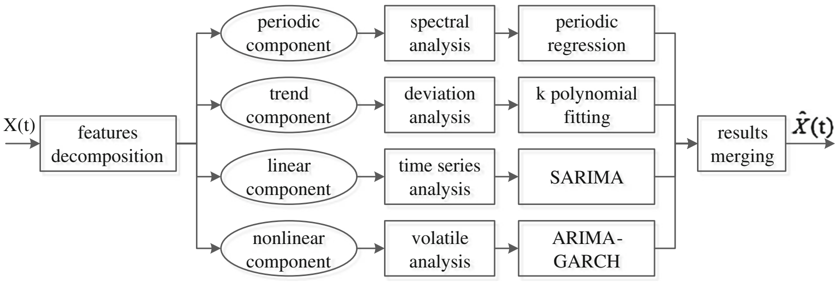


Fig. 1 Hybrid model based on time series multifractal characteristics

four components: a periodic part, a trend part, a linear part and a nonlinear part shown in (1),

$$X(t) = P(t) + T(t) + L(t) + N(t) + \varepsilon(t) \quad (1)$$

where X(t) is the observed traffic counts during time interval t, P(t) is the periodic component expressed as regression of the present on periodic sines and cosines, T(t) is the trend component, L(t) is the linear component and N(t) is the nonlinear component, $\varepsilon(t)$ is the error term. Then, it analyzes each component with a different analysis method separately and finally merges the results.

3.1 Analysis and prediction of periodic component

Spectral analysis, a better frequency domain approach, is effective at capturing intraday periodicity hidden in time series data [21]. As the regressions on periodic sines and cosines don't depend on the past values, spectral analysis gives more apparent insights into the basic structure of traffic flow data than time domain methods. Therefore we adopt spectral analysis to uncover periodic variations over time in traffic flow and use a group of sine and cosine functions to fit this periodicity, which is described by (2).

$$P(t) = R\cos(2\pi f t+\theta) = A\cos(2\pi f t)+B\sin(2\pi f t) \quad (2)$$

where t is the time index, f is a frequency index defined as cycles per unit time, $A = R\cos(\theta)$ and $B = -R\sin(\theta)$, θ is the phase determining the start point of this function and R is a representing the height of the function, in which $R = \sqrt{A^2 + B^2}$.

So any periodicity in time series can be fitted by a linear combination of a group of sine and cosine functions with multiple frequencies, amplitudes and phases. And the regression of the present traffic flow on a periodic can be described in (3), which is a generalization for (2).

$$P(t) = \sum_{j=1}^{m} \left[A_j\cos(2\pi f_j t) + B_j\sin(2\pi f_j t)\right] \quad (3)$$

where j represents the index for individual periodic series, m is the total number of periodic elements in the series. For a certain frequency f, least square error method is used to solve values of A and B.

3.2 Analysis and prediction of trend component

The trend component represents long-term variation characteristics of traffic flow time series. Identifying and removing this trend component can improve forecasting performance. In this study, large trends hidden in the traffic flow series are investigated, such as the seasonal tendency or annual tendency. For large trends in traffic flow series, local k polynomial fitting based on accumulated deviation analysis, which is similar to multifractal detrended fluctuation analysis (MFDFA) method [22], is adopted to uncover the trends component.

First, calculate the accumulated deviation y(i) of traffic flow series x(t) with (3)

$$y(i) = \sum_{k=1}^{i}(x(k) - \bar{x}), \quad (i = 1, 2, ..., n, \bar{x} = \sum_{t=1}^{n} x(t)) \quad (4)$$

where n is the length of traffic flow series x(t).

Then, the accumulated deviation y(i) is divided into $2n_s(n_s = n/s)$ non-overlapping segments with equal length, one time starting from the start of x(t) and the other time from the end of x(t). For sub series of these segments, such as v, k order polynomial is used to fit them and obtain the local trend component with (4). Each local trend component is calculated and the whole trend component of traffic flow series x(t) is obtained.

$$Q_v(i) = a_0 + a_1 i + a_2 i^2 + ... + a_k i^k (k = 1, 2, ...) \quad (5)$$

3.3 Analysis and prediction of linear component

After removing the periodic and trend component in the traffic flow series, the remaining part describes the variation of each day's specific traffic conditions. This variation mainly correlates with the previous traffic conditions in non-peak hours and the stochastic volatility in peak hours. During non-peak period, traffic flow shows some free state and the main variation is caused by interactions of flow at different time lags. As one of the most popular forecasting methods and well defined theoretical foundation and effectiveness in forecasting regular time series data, seasonal autoregressive integrated moving average (SARIMA) model is chosen to capture the variation regulation in non-peak hours and predict future values on regression of past values [23]. The mathematical representation of a *SARIMA*(p, d, q)(P, D, Q)$_s$ model is defined by (5)

$$\varphi_{\mathrm{p}}(\mathrm{B})\varphi_{\mathrm{p}}(\mathrm{B}^{\mathrm{s}})(1-\mathrm{B})^{\mathrm{d}}(1-\mathrm{B}^{\mathrm{s}})^{\mathrm{D}}\mathrm{x}_{\mathrm{t}} = \Theta_Q(\mathrm{B}^{\mathrm{s}})\theta_{\mathrm{q}}(\mathrm{B})\mathrm{e}_{\mathrm{t}} \tag{6}$$

where p and P denote the order of non-seasonal and seasonal autoregressive part respectively, q and Q denote order of the non-seasonal and seasonal lagged error part respectively, d and D denote the number of non-seasonal and seasonal difference respectively, s denotes the number of seasons, x_t is a non-stationary seasonal time series, e_t is a Gaussian white noise series with zero mean and constant variance, B is a delay operator.

3.4 Analysis and prediction of nonlinear component

It is obvious that traffic flow in peak hours changes fiercely and takes on statistical characteristics that the variance is larger and unstationary. Generalized autoregressive conditional heteroscedasticity (GARCH) model can reflect this real situation on traffic flow that the variance of traffic flow usually will last for a certain periods and that a large variance will follow another large variance and a small variance will follow another small variance. Compared with traditional time series model used to forecast traffic flow, GARCH model relaxes the assumption in (5) that e_t simply satisfies the white noise properties and treats the variance of traffic flow data as the conditional heteroscedasticity [24]. GARCH model has the potential to capture these non-linear patterns and is employed to predict the nonlinear volatile component of traffic flow series. The expression of e_t is

$$\begin{aligned} &e_t = \sqrt{h_t}\varepsilon_t, \\ &h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i e_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}, \\ &\varepsilon_t \sim N(0, 1) \end{aligned} \tag{7}$$

where h_t is the conditional variance of traffic flow series, ε_t is a sequence assumed to follow a standard normal distribution, and

$$\begin{aligned} &p \geq 0, q > 0, \alpha_0 > 0, \\ &\alpha_i \geq 0, i = 1, 2, \cdots, q, \beta_{\mathrm{i}} \geq 0, i = 1, 2, \cdots, p, \textstyle\sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i) < 1 \end{aligned}$$

4 Short-term traffic flow forecasting based on the hybrid model

4.1 Data descriptions and preprocessing

The traffic flow data studied in this paper comes from the Minnesota Department of Transportation which has been collecting and publishing daily volume and occupancy data for the Twin Cities' freeways. We select the data from six stations located on I-694 EB to investigate the accuracy and reliability of the model for forecasting freeway traffic flow. Figure 2 presents the location of I-694 EB and the yellow segment is the selected stations, of which the IDs are 163, 165, 166, 168, 171, and 173. Every station has three detectors. So data from 18 detectors is used to study. The IDs of these detectors are 530, 531, 532, 538, 539, 540, 532, 543, 544, 549, 550, 551, 741, 742, 743, 747, 748, and 749. Measurements take place every 30s. The sample time is from August 1, 2016 to August 31, 2016. Sample data has been aggregated into five-minute intervals. Partial missing data on day 20 and 21 is replaced by data calculated with moving average method.

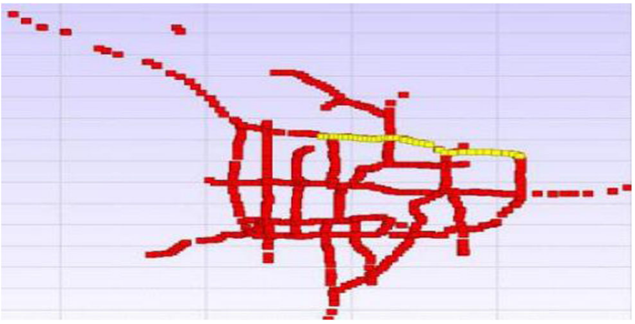
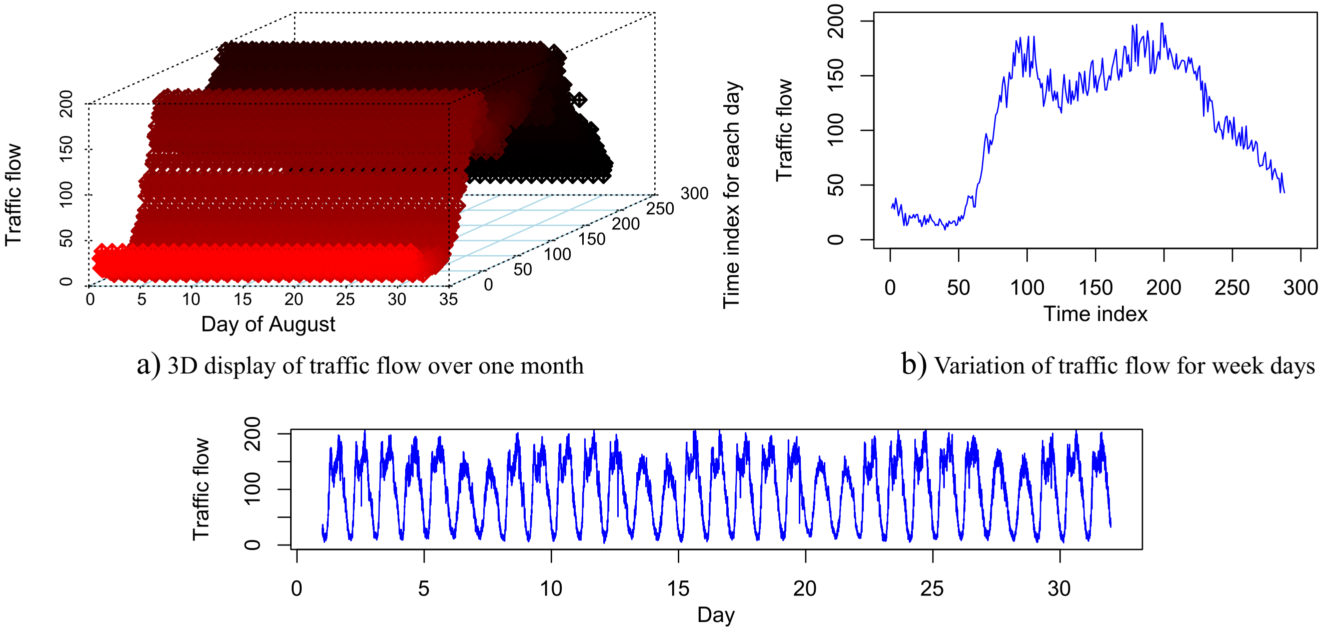


Fig. 2 Distribution of stations

Figure 3 presents the characteristics of traffic flow data on I-694 EB in the Twin Cities from August 1, 2016 to August 31, 2016. From chart a), it can be seen that traffic flow has similar patterns over different days and different times of a day and reaches peaks at the same time index. Chart c) shows that traffic flow on weekdays and weekend days takes on different features. The former is the M shape and presents obvious morning/afternoon peaks. But the latter has only one peak and the peak value is smaller than that of weekdays. So we mainly study the variation of traffic flow on weekdays. Because there are 8 weekend days in August 2016, 22 weekdays of traffic flow data is used to forecast the traffic flow of the last weekday. Chart b) describes the traffic flow series of one weekday. It shows the periodogram reaches the largest value at time index 100 and the second largest value at time index 195. That is to say, it reaches morning peak at time about 8:15 am and afternoon peak at time about 4:20 pm.

4.2 Spectral analysis and periodic prediction

Cyclical regression method based on spectral analysis is used to estimate the periodical patterns of traffic flow data. In this paper, all parameters are computed by statistical software R, which contains lots of packages we can use directly, and the least square error method is utilized to optimize



a) 3D display of traffic flow over one month

b) Variation of traffic flow for week days

c) Traffic flow series over one month

Fig. 3 Traffic flow from August 1, 2016 to August 31, 2016

the parameters' values. So we use harmonic function (harmonic(x, m)) in package TSA to create a matrix of m pairs of harmonic functions for fitting a harmonic trend (cosine-sine trend) with the response being x, a traffic flow series. Then, linear regression lm is used to fit the created matrix consisting of a group pairs of cos ($2k\pi t$) and sin ($2k\pi t$) ($k = 1, 2, \cdots, m$). The estimated parameters for the regression model are listed in Table 1. The intercept, which is 105.52, approximates the mean value of traffic flow, which is 103.69. Parameters of each sine and cosine represent the amplitude of each cyclical component. The value of R-squared is 0.96, a better goodness of fit P-value is smaller than 2.2e-16. So all of these estimated parameters are statistically significant. Figure 4 describes the original data, the periodic component data and the fitted chart for five days of traffic flow series. It can be seen that the model fits the original data very well.

4.3 Accumulated deviation analysis and trend prediction

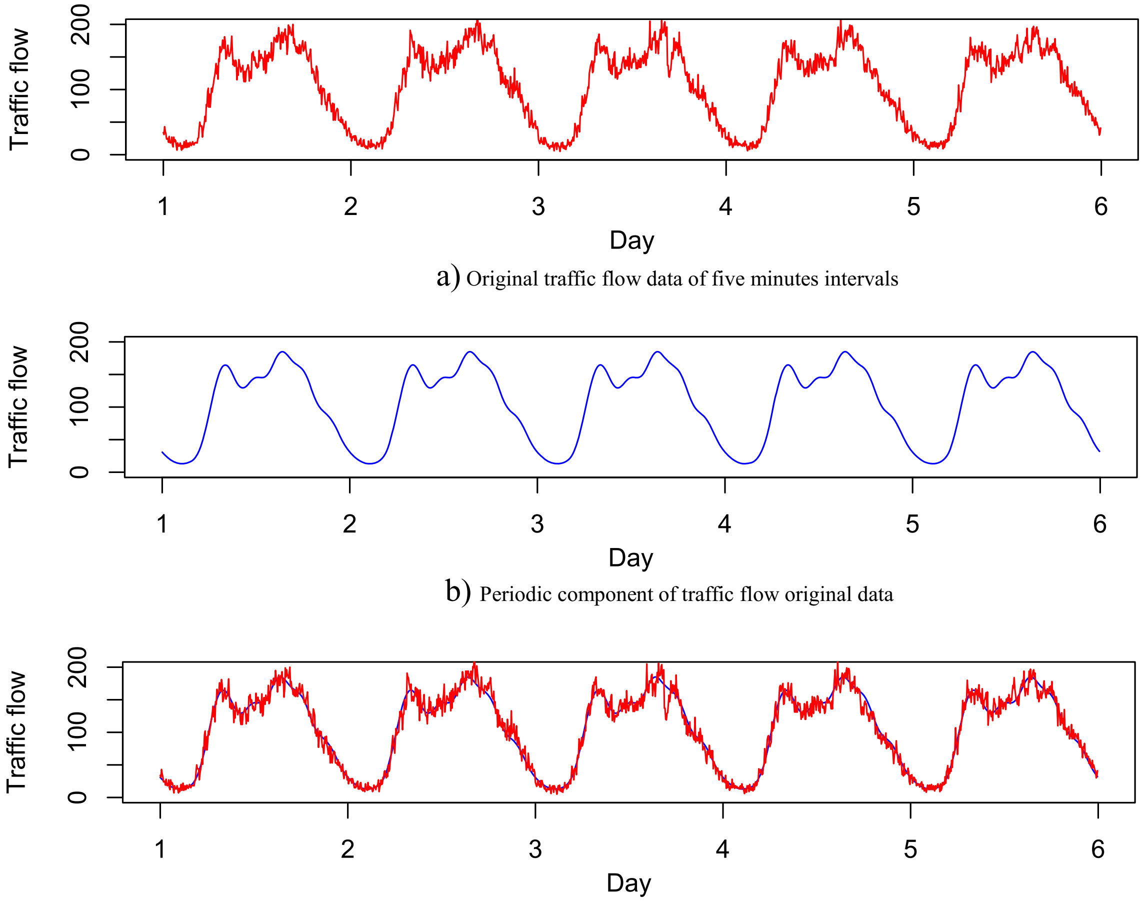
After removing periodic component, we calculate the trend component of residual data using (3) and (4) with k equaling 2 in (4). Figure 5 shows the trend component of residual data for 22 weekdays after removing periodic part. It indicates that the difference of probable minimum and maximum values is very small and less than 8. This trend may be caused by stochastic volatility of traffic flow variation. Thus, the trend component of this sample traffic flow data can be ignored.

4.4 Analysis and modeling for the residuals

Figure 6 presents the residuals of traffic flow data removed periodic and trend component. From chart a), it can be seen that residuals still show some same pattern that daily traffic counts take on states of first stationarity and then volatility. For one weekday of residuals, values of traffic counts before 5:40 am (time index 67) are relatively stationary and small, but they begin to vibrate and become more complex after 5:40 am which are shown in chart b). So different modeling and predicting method is used to capture these features The residuals between 0:00 am and 5:40 am are predicted with SARIMA model and remaining residuals are predicted with ARIMA-GARCH model showed in chart b).

Table 1 Estimated parameters

Intercept	Cos(2*pi*t)		Cos(4*pi*t)		Cos(6*pi*t)		Cos(8*pi*t)		Cos(10*pi*t)		Cos(12*pi*t)		Cos(14*pi*t)		R^2	P-value
	Sin(2*pi*t)		Sin(4*pi*t)		Sin(6*pi*t)		Sin(8*pi*t)		Sin(10*pi*t)		Sin(12*pi*t)		Sin(14*pi*t)			
105.52	−64.15	−33.37	−16.81	−15.39	16.64	−5.00	0.17	5.55	−6.72	−4.24	−0.71	−3.48	−1.84	3.14	0.96	<2.2e-16



a) Original traffic flow data of five minutes intervals

b) Periodic component of traffic flow original data

c) Fitting chart of original data and periodic component

Fig. 4 Spectral analysis and periodic regression

4.4.1 SARIMA modeling and analysis

The appropriate order of SARIMA is determined based on two factors, which are AIC (Akaike information criterion) and statistical significance of parameters. The best model should be the one that has smaller value of AIC:

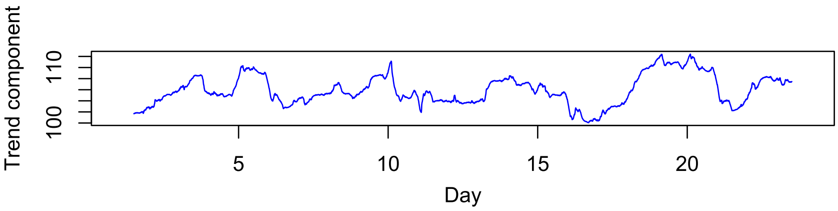
$$AIC = -2\log(L) + 2m \tag{8}$$

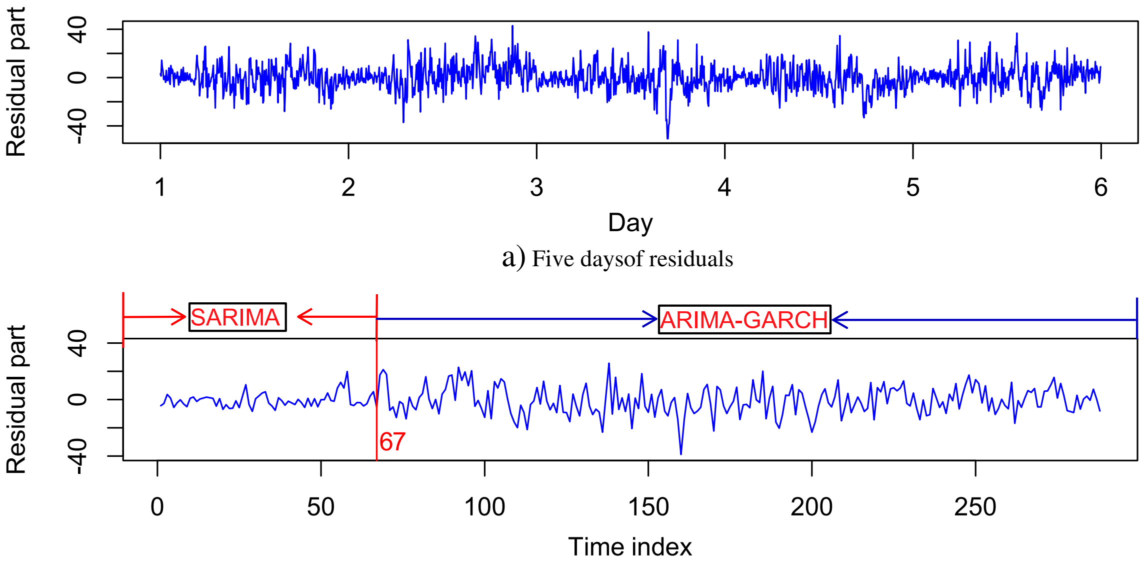
where L is the likelihood of data for the specific model and m is the number of parameters selected for this model, and for which all the parameters are statistical significance with 95% high confidence level. All parameters are estimated through the maximum likelihood method carried out by the statistical software R. Based on the two criteria, the SARIMA(2,1,1) $(0,0,1)_{67}$ model fits the non-peak hours' traffic flow data best. The fitted model is:

$$\begin{aligned}&(1-0.27B+0.014B^2)(1-B)x_t\\&\quad=(1-0.956B)(1-0.179B^{67})a_t, \sigma_a^2=23.52\end{aligned} \tag{9}$$

The standard errors are 0.06, 0.059, 0.025 and 0.053. Figure 7 shows the result of model checking. Chart a) shows that the standardized residuals evenly distribute near the zero and that the maximum value is less than 3 Chart b)

Fig. 5 One month of trend component



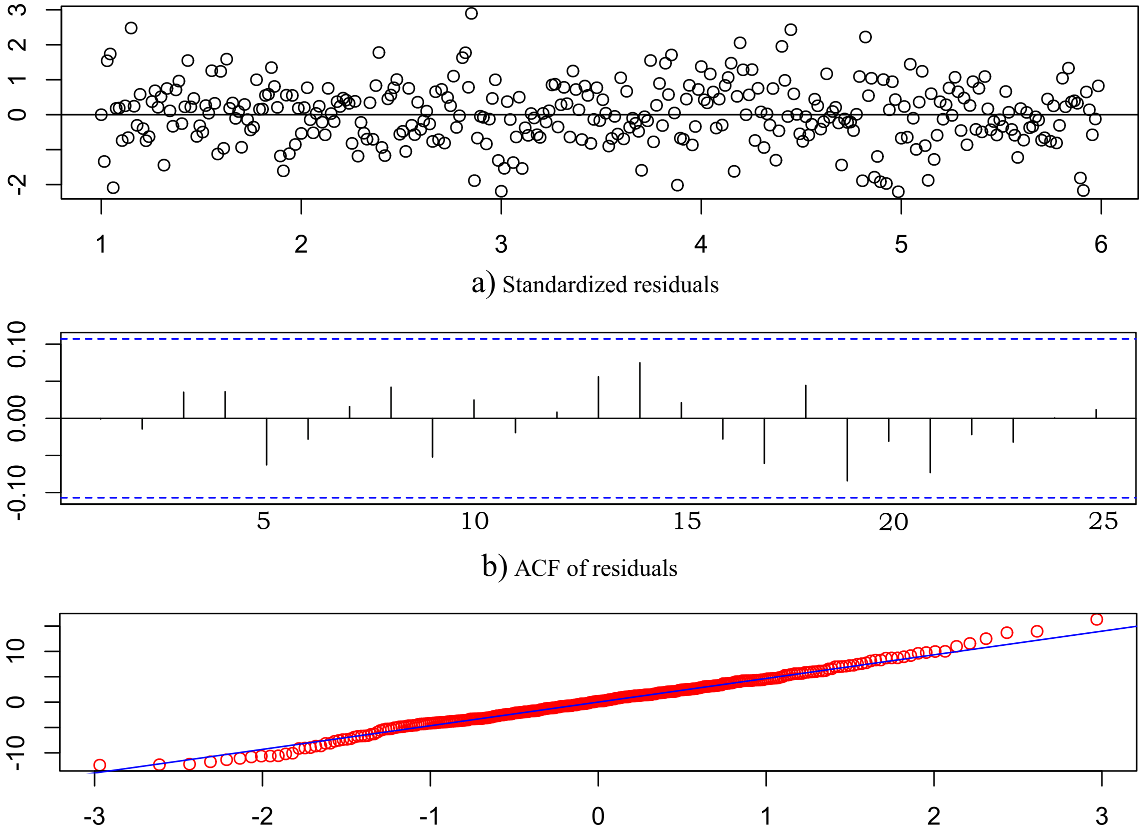


a) Five daysof residuals

b) One day of residuals

Fig. 6 Residuals removed periodic and trend component

indicates that no significant correlations exist in the residual series. Chart c) indicates that the distribution of the error series of SARIMA model satisfies normality except for only a few points at Q-Q plot two ends. All these features show that the SARIMA model is correct and fits residuals in non-peak periods very well.



a) Standardized residuals

b) ACF of residuals

c) Normal Q-Q plot of residuals

Fig. 7 Results of model checking

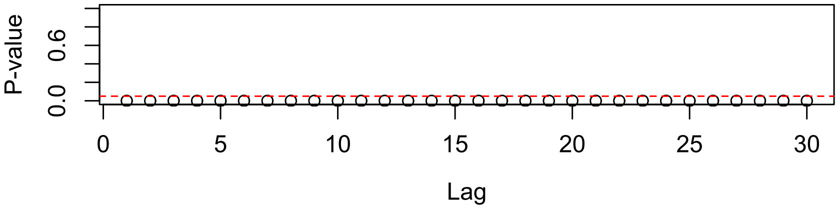


Fig. 8 P-value for squared values of residuals

4.4.2 Volatility modeling and analysis

Before applying the volatility model, it is necessary to detect the ARCH effect. Only the ARCH effect exists, the volatility model is needed. The Ljung-Box statistics of the squared values of residuals are used to test the validity. As showed in Fig. 8, p-value is about 2.2e-16 and far less than 0.05, which proves strong serial correlations exist within the residuals part. Thus, the volatility model is necessary.

According to the criterion of parameters significance and errors minimization, ARIMA(3,1,1)- GARCH(1,1) is used to fit residuals from 5:40 am to 12 pm. The fitted model is:

$$(1\text{-}B + 0.092B^2 + 0.014B^3)x_t = (1 - 0.703B)a_t,$$
$$a_t = \sigma_t\varepsilon_t, \varepsilon_t \sim N(0, 1),$$
$$\sigma_t = 6.586 + 0.044a_{t-1}^2 + 0.901\sigma_{t-1}^2 \tag{10}$$

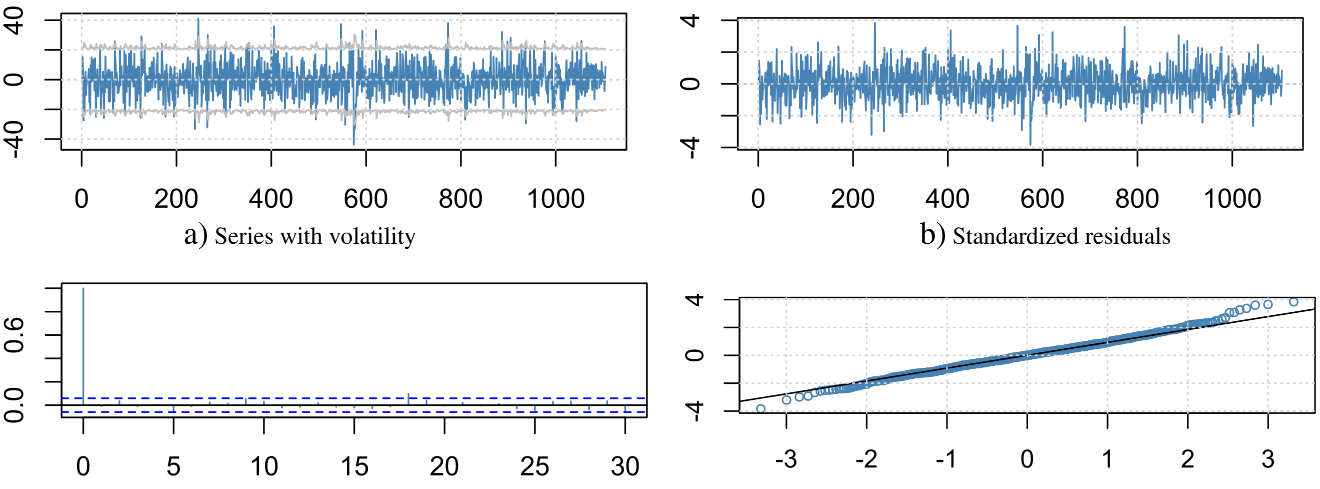
All estimated parameters are statistically significant with 95% confidence degree. Figure 9 shows the result of modeling and checking. From chart a), it is shown that the volatility well follows the residual series and model can be used to predict changing interval of traffic flow. The interval prediction is more practical than point prediction. The distribution of standardized residuals and the values of ACF indicate that ARMA(3,1)-GARCH(1,1) fits residual series very well. Chart d) shows that standardized residuals of model basically satisfy normality. Only a few front end of points deviate the line. It is normal because all used traffic flow data is the raw material without abnormal handling and outliers may exist.

5 Model evaluation

Two datasets are used to test the performance of the hybrid model. One dataset (dataset1) is the dataset described in Section 4.1. 22 days of data is used to train the model and the last day of data is used to evaluate the model. The other dataset (dataset2), one weekday of traffic flow data from another highway in the Twin Cities, is applied to test the robustness of model to further check the performance of the model.

After removing periodic and trend component based on spectral analysis and accumulated deviation analysis, residuals of each day were divided into two parts taking time index 67 as the boundary, and SARIMA and ARIMA-GARCH were applied to model and forecast. Figures 10 and 11 separately describe rollback forecasting results of one-step ahead on dataset1 and dataset2 with the proposed hybrid model. Blue solid line stands for real traffic flow values, and red points represent the predicted traffic flow values.

Two figures show that the proposed model accurately predicts traffic flow values with small errors in most situations, especially in non-peak periods. In rush hours, when a



a) Series with volatility

b) Standardized residuals

c) ACF of residuals

d) Normal Q-Q plot of standardized residuals

Fig. 9 Modeling and checking of volatile part

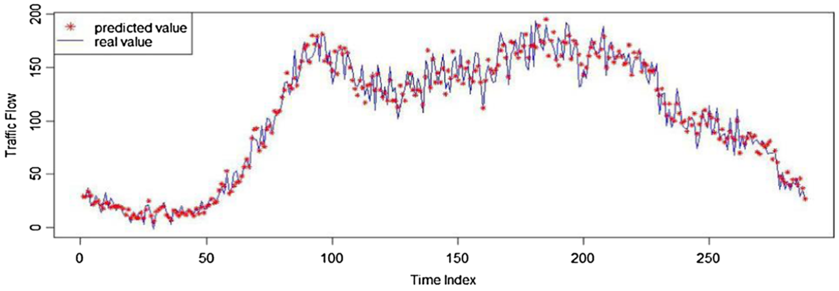


Fig. 10 Forecasting results on dataset1

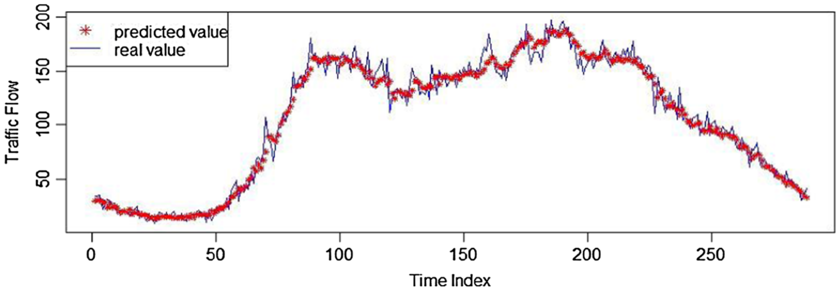


Fig. 11 Forecasting results on dataset2

Table 2 RMSE and MAE values of one-step ahead forecasting

	RMSE	MAE
ARIMA	10.35	8.09
ARIMA-GARCH	10.33	8.09
Hybrid method(dataset1)	9.48	7.29
Hybrid method(dataset2)	9.72	7.43

Table 3 RMSE for ARIMA, ARIMA-GARCH and hybrid model

	Numbers of forecasting steps ahead									
	1	2	3	4	5	6	7	8	9	10
ARIMA	10.35	11.32	12.37	13.42	14.53	15.60	16.66	17.57	18.61	19.65
ARIMA-GARCH	10.33	10.35	10.37	10.37	10.40	10.41	10.43	10.45	10.46	10.48
Hybrid method	9.48	9.50	9.49	9.47	9.46	9.41	9.40	9.37	9.36	9.36

Table 4 MAE for ARIMA, ARIMA-GARCH and hybrid model

	Numbers of forecasting steps ahead									
	1	2	3	4	5	6	7	8	9	10
ARIMA	8.09	8.68	9.38	10.03	10.76	11.71	12.56	13.46	14.52	15.53
ARIMA-GARCH	8.09	8.11	8.14	8.15	8.17	8.18	8.20	8.23	8.25	8.27
Hybrid method	7.29	7.27	7.24	7.21	7.18	7.13	7.09	7.05	7.00	6.99

Table 5 RMSE for model of literature [17] and our model

	Numbers of forecasting steps ahead									
	1	2	3	4	5	6	7	8	9	10
Literature [17]	23.43	25.90	28.00	29.76	30.96	32.11	32.94	33.30	33.88	34.65
Our method	9.48	9.50	9.49	9.47	9.46	9.41	9.40	9.37	9.36	9.36

larger volatility occurs, some points deviate from real values both on two figures. Compared with Fig. 10, deviation on Fig. 11 is more obvious and the predicted values are more stationary and agglomerative. But on the whole, the predicted values well follow the real values. The errors are listed in Table 2 (definition of RMSE and MAE sees (10) and (11)). Table 2 shows the robustness and performance of our model are both better than that of ARIMA and ARIMA-GARCH. Errors on dataset2 are a little larger than errors on dataset1 with RMSE value 9.72 and MAE value 7.43. But they are smaller than those of ARIMA and ARIMA-GARCH.

Two measurements of effectiveness are applied to evaluate the model forecasting performance, which are root mean squared error (RMSE) and mean absolute error (MAE). The equations are showed as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i) - \hat{x}(i))^2} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x(i) - \hat{x}(i)\right| \tag{12}$$

where $x(i)$ is real value, $\hat{x}(i)$ is predicted value, and n is the total number of forecasting data.

This study compares forecasting errors from one-step ahead to ten-step ahead of ARIMA, ARIMA-GARCH and hybrid model to investigate multi-step ahead traffic flow forecasting performance. As mentioned above, the orders of the ARIMA model are selected based on criterions that AIC is the least and all parameters are statistical significance. The parameters of the model are estimated through the maximum likelihood method by utilizing statistical software R. So model ARIMA(2,1,3) and ARIMA(2,1,3)-GARCH(1,1) is employed to fit traffic flow data. Tables 3 and 4 respectively list the RMSE and MAE values for the ARIMA, ARIMA-GARCH and the hybrid method from one-step to ten-step ahead forecasting.

Results indicate that the hybrid model performs better than the ARIMA and ARIMA-GARCH. For one-step ahead forecasting, RMSE for ARIMA is close to RMSE for ARIMA-GARCH, but RMSE for hybrid method is the smallest with an improvement of 8.41% and 8.23% in forecasting accuracy. What's more, the experimental results indicate that while forecasting steps increase, RMSEs for ARIMA increase rapidly with 19.65 of ten-step ahead forecasting, RMSEs for ARIMA-GARCH increase gradually with 10.48 of ten-step ahead forecasting, but RMSEs for hybrid model decrease gradually with 9.36 of ten-step ahead forecasting. It represents that the hybrid model shows obvious advantages compared with the other two models with an improvement of 52.37% and 10.69% in forecasting accuracy of ten-step ahead. This illustrates that the hybrid model performs well in improving forecasting accuracy and can capture the longer traffic flow variation trend. In order to further examine the forecasting performance, Table 4 shows MAE of three models. Same results are obtained. For one-step ahead forecasting, hybrid mode gets an improvement of 9.89% compared with the first two models. As forecasting steps increases, MAEs for ARIMA increase rapidly and MAEs for ARIMA-GARCH increase gradually, but MAEs for hybrid model decrease gradually Compared with two models, our model obtains improvements of 54.99% and 15.48% in forecasting accuracy of ten-steps ahead.

In order to have a further investigation to the performance of our proposed model, we compared it with the model proposed in literature [17], which is a hybrid short-term traffic flow forecasting model based on spectral analysis and statistical volatility. Using ARIMA-GARCH model as a benchmark model, literature [17] obtains an improvement of 3.63% in forecasting accuracy of one step ahead, whereas the improvement of our proposed model is 8.23%. The average RMSEs for our model are far less than those for model of literature [17], which are shown in Table 5. All this illustrates that the performance of our hybrid model is better and it can be used for short-term traffic flow forecasting. All these investigations illustrate that both forecasting accuracy and portability of our hybrid model are better. And it can be used to forecast short-term traffic flow or act as a supplementary means of traffic analysis and prediction.

6 Conclusions

This paper proposes a new hybrid model for traffic flow forecasting. Firstly, this model decomposed traffic flow into four different components: a periodic part, a trend part, a stationary part, and a volatility part. Then, for a different part, a different method is used to fit the traffic flow data and perform the final prediction. We adopted two datasets on weekday traffic flow from freeway I-694 EB in the Twin Cities to evaluate the forecasting performance of proposed hybrid model. The results indicate that the hybrid model performs well in improving forecasting accuracy. For multi-step ahead forecasting, the improvement is more obvious. So following conclusions are drawn.

- Treating the periodicity, tendency, stationarity, and volatility individually and establishing a different model for a different part can highly improve forecasting accuracy in terms of RMSE and MAE.
- Removing periodic and trend part from traffic flow data and segmentation modeling for the residual part can improve the forecasting accuracy of the model. It may be that periodicity and tendency present long term characteristics of traffic flow and that segmentation modeling has better goodness of fit.

- By synthetically considering characteristics of traffic flow, hybrid models provide better insights into the basic structures of traffic flow data and interpret the regularity of traffic flow changing hidden in the data.
- ARIMA model can't present non-linear characteristic of traffic flow and is not suitable for multi-step ahead forecasting. ARIMA-GARCH model can efficiently capture traffic volatility and improve forecasting accuracy of multi-step ahead.

In summary, hybrid models outperform single models and they are the development direction of studying traffic flow forecasting. Compared with classical mode, such as ARIMA and ARIAMA-GARCH, or new model proposed in literature the hybrid model performs well. For one-step ahead forecasting, our method gets improvements of 8 and 9 percentage points in RMSE and MAE compared with ARIMA and ARIMA-GARCH For ten-step ahead forecasting, it gets improvements of about 50 percentage points compared with ARIMA and 10 to 15 percentage points compared with ARIMA-GARCH in RMAE and MAE. Compares with the model newly proposed in literature [17], it gets an improvement of 1.27% in forecasting accuracy of one-step ahead Therefore, the hybrid model proposed in this paper achieves satisfactory performance and can be used to forecast the short-term traffic flow accurately no matter one-step or multi-step ahead. When other factors, such as speed, weather condition, time intervals and so on, are considered, how about the performance of the proposed model? This is the next research mission we should do.

Acknowledgments The authors are grateful to the anonymous reviewers for their comments, which will help to improve this paper.

Funding This work was supported by National Natural Science Foundation of China [Grant No.61663021]; Scientific Research Project in Universities of Gansu [Grant No. 2015B-031]; Science and Technology Support Program of Gansu [Grant No.1304GKCA023].

References

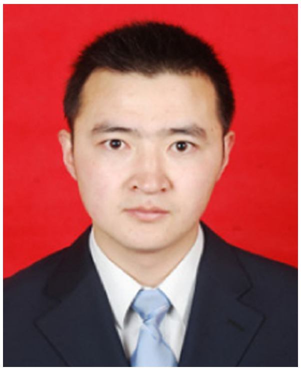
1. Wang Y, Geroliminis N, Leclercq L (2016) Recent advances in ITS, traffic flow theory, and network operations[J]. Trans Res Part C: Emerg Technol 68:507–508
2. Zhang Y, Zhang Y (2016) A comparative study of three multivariate short-term freeway traffic flow forecasting methods with missing data[J]. J Intell Transp Syst 20(3):205–218
3. Ghosh B, Basu B, Mahony MO' (2009) Multivariate short-term traffic flow forecasting using time-series analysis[J]. IEEE Trans Intell Transp Syst 10(2):246–254
4. Dong C, Xiong Z, Shao C, Zhang H (2015) A spatial-temporal-based state space approach for freeway network traffic flow modelling and prediction [J]. Trans: A Transp Sci 11(7):1–14
5. Tang J, Liu F, Zou Y, Zhang W, Wang Y (2017) An improved fuzzy neural network for traffic speed prediction considering periodic characteristic [J]. IEEE Trans Intell Transp Syst 18(9):2340–2350
6. Moretti F, Pizzuti S, Panzieri S, Annunziato M (2015) Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling[J]. Neurocomputing 167(C):3–7
7. Li Y, Jiang X, Zhu H et al (2016) Multiple measures-based chaotic time series for traffic flow prediction based on Bayesian theory [J]. Nonlinear Dyn 85(1):179–194
8. Pang X, Wang C, Huang G (2016) A short-term traffic flow forecasting method based on a three-layer k-nearest neighbor non-parametric regression algorithm[J]. J Transp Technol 6:200–206
9. Zheng Z, Su D (2014) Short-term traffic volume forecasting: a k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm [J]. Transp Res Part C: Emerg Technol 43:143–157
10. Habtemichael FG, Cetin M (2016) Short-term traffic flow rate forecasting based on identifying similar traffic patterns[J]. Transp Res Part C: Emerg Technol 66:61–78
11. Cheng A, Jiang X, Li Y et al (2016) Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method[J]. Physica A Stat Mech Appl 466:422–434
12. Hu W, Yan L, Liu K, Wand H (2016) A short-term traffic flow forecasting method based on the hybrid PSO-SVR [J]. Neural Process Lett 43:155–172
13. Wang C, Ye Z (2015) Traffic flow forecasting based on a hybrid model [J]. J Intell Transp Syst 20(4):428–437
14. Wei Y, Chen M (2012) Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks [J]. Transp Res Part C: Emerg Technol 21(1):148–162
15. Wang J, Shi Q (2013) Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet Analysis-Support Vector Machine theory [J]. Transp Res Part C: Emerg Technol 27:219–232
16. Chen C, Wang Y, Li L, Hu J, Zhang Z (2012) The retrieval of intra-day trend and its influence on traffic prediction[J]. Transp Res Part C: Emerg Technol 22:103–118
17. Zhang Y, Zhang Y, Haghani A (2014) A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model [J]. Transp Res Part C: Emerg Technol 43(1):65–78
18. Lopez-Garcia P, Onieva E, Osaba E, Masegosa AD (2016) A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy [J]. IEEE Trans Intell Transp Syst 17(2):557–569
19. Tan H, Wu Y, Shen B, Jin PJ, Ran B (2016) Short-term traffic prediction based on dynamic tensor completion [J]. IEEE Trans Intell Transp Syst 17(7):1–11
20. Blandin S, Argote J, Bayen AM, Work DB (2013) Phase transition model of non-stationary traffic flow: Definition, properties and solution method [J]. Transp Res Part B Methodol 52(2):31–55
21. Belletti F, Huo M, Litrico X, Bayen AM (2015) Prediction of traffic convective instability with spectral analysis of the Aw–Rascle–Zhang model [J]. Phys Lett A 379(38):2319–2330
22. Shang P, Lu Y, Kamae S (2008) Detecting long-range correlations of traffic time series with multifractal detrended fluctuation analysis [J]. Chaos, Solitons Fractals 36(1):82–90
23. Kumar SV, Vanajakshi L (2015) Short-term traffic flow prediction using seasonal ARIMA model with limited input data [J]. Eur Transp Res Rev 7(3):21
24. Chen C, Hu J, Meng Q, Zhang Y (2011) Short-time traffic flow prediction with ARIMA-GARCH model [J]. Intell Veh Symp 32(14):607–612

Hong Zhang was born in Gansu, P. R. China. She received the Bachelor degree in Computer Application from Lanzhou University of Technology, Lanzhou, China, in 2001, the Master degree in Communication and Information System from Lanzhou University of Technology, Lanzhou, China, in 2004. She is currently a Ph. D. student in Systems Engineering, Lanzhou University of Technology. She is also an associate professor in College of Computer & Communication, Lanzhou University of Technology. Her research interests are in the areas of intelligent transportation systems and machine learning.

Xiaoming Wang was born in Gansu, P. R. China. He received the Bachelor degree in Automatic Control from Lanzhou Jiaotong University, Lanzhou, China, in 1982. He is currently a professor as well as a doctoral supervisor in College of Electrical & Information Engineering, Lanzhou University of Technology. His research interests are in the areas of intelligent transportation systems and intelligent information processing.

Jie Cao was born in Gansu, P. R. China. She received the Bachelor degree in Automatic Control from Lanzhou University of Technology, Lanzhou, China, in 1987, the Master degree in Electrical Engineering from Xian Jiaotong University, Xian, China, in 1994. She is currently a professor as well as a doctoral supervisor in College of Computer & Communication, Lanzhou University of Technology. Her research interests are in the areas of intelligent information processing and information fusion.

Minan Tang was born in Hanzhong, Shaanxi, P. R. China. He received the Master degree in Communication and Electronic Engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2006, the Ph.D. degree in Transportation Information Engineering and Control from Lanzhou Jiaotong University, Lanzhou, China, in 2011. He is currently an associate professor in School of Automation and Electrical Engineering, Lanzhou Jiaotong University. He is also a post-doctoral fellow in Intelligent Transportation Control from Lanzhou University of Technology. His research interests are in the areas of intelligent control systems, intelligent transportation systems.



Yirong Guo was born in Gansu, P. R. China, in 1982. He received the Bachelor degree in Computer Science from Lanzhou University of Technology, Lanzhou, China, in 2006, the Master degree in Computer Science from Lanzhou University of Technology, Lanzhou, China, in 2009. He is currently a Ph. D. student in Control Theory and Control Engineering, Lanzhou University of Technology. His research interests are in the areas of intelligent information processing and intelligent transportation systems.