Applied Intelligence (2018) 48:3827–3838
https://doi.org/10.1007/s10489-018-1181-7

A multivariate short-term traffic flow forecasting method based on wavelet analysis and seasonal time series

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

Published online: 30 April 2018

Abstract

Short-term traffic flow forecasting is a key step to achieve the performance of intelligent transportation system (ITS). Timely and accurate traffic information prediction is also the prerequisite of realizing proactive traffic control and dynamic traffic assignment effectively. Based on the fact that univariate forecasting methods have limited forecasting abilities when the data is missing or erroneous and that single models make no full use of information underline data, a new hybrid method with multivariate for short-term traffic flow forecasting is proposed. This method combines statistical analysis method with computational intelligence techniques to mine the characteristic of traffic flow as well as forecast short-term traffic state. First, the wavelet de-noising is employed to remove the noise information. Then, time series analysis is used to analyze time-varying and periodic characteristic of traffic flow. Furthermore, the seasonal auto-regressive moving average with external input (SARIMAX) is established to fit traffic flow with occupancy as exogenous variables. Finally, wavelet forecast is adopted to forecast the values of occupancy which are used as exogenous input, and a WSARIMAX is constructed to forecast traffic flow. Using the relationship of flow and occupancy at the same road section and taking traffic flow and occupancy data from freeway I-694 EB in the Twin Cities as endogenous variables and exogenous variables respectively, this paper studies the forecasting performance of the proposed method. The study results are encouraging. Compared with SARIMA newly proposed in literature, WSARIMA and SARIMAX improved method with wavelet analysis and multivariate modeling method, the proposed method gets improvements of 12.95%, 12.62% and 10.41% in forecasting accuracy of one-step ahead respectively. For ten-steps ahead forecasting, it gets improvements of 18.87%, 17.05% and 2.57% in forecasting accuracy respectively.

Keywords Hybrid method · Traffic flow · Short-term forecasting · Wavelet analysis · Multivariate

1 Introduction

Prediction of traffic information is the foundation of traffic control and guidance and plays a significant role in the transportation system domain. The key issue of ITS is the accurate and timely prediction of traffic information [1]. The prediction information for traffic state not only makes travelers acquire more information of traffic conditions and make better travel decisions to save their time and cost, as well as improve road safety by reducing traffic accidents and congestion, but also benefit the environment by reducing carbon dioxide emissions and improve the level of traffic management and planning [2]. Traffic state is closely related to daily life and gets more and more concerns of many researchers. So traffic forecasting has always been a hotspot among researchers. In the past few years, a large amount of algorithms have been proposed to address traffic prediction problems. These studies mainly focus on the methodologies or models, such as parametric methods (e.g. historical average method [3], time series analysis method [4], state space models [5], etc.), non-parametric methods (e.g. the k-nearest neighbor method [6],

✉ Hong Zhang
zhanghong@lut.cn

1 College of Computer & Communication, Lanzhou University of Technology, Lanzhou, 730050, China

2 College of Electrical & Information Engineering, 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

support vector regression method [7, 8], and artificial neural network method [9, 10], etc.), and hybrid methods [11–13]. These methods or models have obtained better results, but every method has its applicable condition limitation and there is no one showing absolute superiority. The hybrid methods combining the merits of different methods produce higher prediction accuracy than single approaches and improve forecasting performance in traffic flow and travel time forecasting.

Traffic system is a typical complex system in terms of system engineering. The variations of traffic state are caused by complex interactions of several factors, such as temporal changes in traffic flow, road architecture, weather conditions, accidents, repair work and so on. Every variation will cause a variation of traffic state [14]. The problem of traffic flow forecasting cannot be solved only by a mathematical model. These factors must be comprehensively considered. Yet, many forecasting methods existing in the literature concern model studying with univariate, such as traffic flow or vehicles speed or travel time. These univariate forecasting methods use both historical and current traffic data to predict the future roadway conditions at the same location. When there is missing data, the forecasting accuracy of univariate methods will be greatly affected. Multivariate models can use multivariate to offset these shortages and yield unbiased estimates [15]. Multivariate models which use the related traffic parameters as exogenous variables perform better than univariate models. Thus, it is necessary to study multivariate hybrid methods and compensate for the shortages of univariate models in order to improve forecasting accuracy for dynamic traffic state.

2 Literature review

Predicting future traffic conditions is a difficult task because of the fact that many factors influence the variations of traffic state and that traffic flow not only exhibits characteristics of periodicity and tendency, but also reveals randomness evoked by exogenous factors. These challenges prompt the phenomenon that the forecasting of traffic state is always a hotspot. There are many hybrid methods on studying traffic flow forecasting.

Wang [16] proposed a traffic speed forecasting hybrid model combining a wavelet function, phase space reconstruction and support vector machine regression theory. Lopez-Garcia [17] combined genetic algorithm (GA) and cross entropy (CE) method to predict congestion of the I5 freeway in California. The results prove that the hybrid method is more accurate than GA or CE alone in forecasting short-term traffic congestion. Artificial neural networks (ANN) method combined with many other algorithms, such as simple statistical or data mining approaches, was developed to improve the performance of traffic forecasting [18, 19]. Corrêa [20] proposed a WARIMAX-GARCH method to improve forecasting accuracy of daily dam displacement in Brazil, which combined wavelet analysis with auto-regressive integrated moving average with exogenous variables (ARIMAX) and generalized auto-regressive conditional heteroscedasticity (GARCH). Paul [21] took maximum temperature as an exogenous variable and proposed an ARIMAX-GARCH-WAVELET model to predict wheat yield in Kanpur district of Uttar Pradesh in India. The results indicate that the ARIMAX-GARCH-WAVELET model outperforms other models as far as modelling and forecasting are concerned. Zhang [22] used traffic flow data from multi-detectors at nearby locations and built two multivariate forecasting approaches, the vector auto regression and the general regression neural network, to study the issue of short-term traffic flow forecasting. The results indicate that general regression neural network gives more robust and accurate forecasting for missing sample data. Yin [23] adopted time-delay reconstructions as well as embedding dimensions and built multivariate predicting method to forecast traffic time series. The results indicate that the multivariate forecasting method is more accurate than the univariate model.

Two conclusions are drawn from the results of these researches. One is that the hybrid methods, which take advantages of two or more models and consider both linear and nonlinear features of traffic flow, can improve the accuracy of traffic flow forecasting. The other is that multivariate methods incorporating exogenous variables are more accurate in forecasting traffic time series than univariate methods. Multivariate methods have a broad application prospect of predicting traffic state. Therefor hybrid methods with multivariate can improve the forecasting accuracy. However, most hybrid methods focus on the study of methodology and technology, and the multivariate mainly refers to the same kind data from multi-detectors at nearby locations, such as literature [22]. Only few literatures [23] study multivariate method with traffic parameters, such as speed and flow. But their methods are single methods like KNN, ANN, SVM or linear regression. Single methods can't fully mine traffic features underlying data and their forecasting performances are far less than those of hybrid methods. On the whole, studies on hybrid methods with multiple kinds of traffic parameters are few.

This paper mainly develops a hybrid and multivariate traffic flow forecasting method to improve short-term forecasting accuracy. The novelty of our work is that it combines statistical analysis method with computational intelligence techniques to further probe into the linear and non-linear characteristic of traffic flow. It can analyze multimodal features of traffic flow on a more microscopic

level and improve forecasting accuracy of short-term traffic flow. First, a wavelet analysis is employed to decompose traffic data to remove noise and pick up useful information. Then, time series analysis is used to analyze time-varying and periodic characteristic of traffic flow. Furthermore, a seasonal auto-regressive moving average with external input (SARIMAX) is established to forecast traffic flow, and wavelet forecast is adopted to forecast the values of X in SARIMAX model. Another novelty is that it studies the influence of the exogenous variable, occupancy, on the endogenous variable, traffic flow and develops a hybrid multivariate forecasting method, which conforms to the characteristics of traffic system. It compensates the shortages of univariate forecasting methods that have poor forecasting abilities when the data is missing or erroneous. It is also different from the existing methods in literatures where the multivariate mainly refers to the same kind data from multi-detectors at nearby locations. Multivariate forecasting methods are the research hotspots in the field of traffic flow forecasting.

The remaining sections of this paper are organized as follows. The third section presents the multivariate short-term traffic flow forecasting method and analyzes related theory. The fourth section presents the corroboration of prediction scheme and performance evaluation of the proposed method. The fifth section analyzes the results and draws conclusions.

3 The building of WSARIMAX

Only we understand the dynamic characteristics of traffic flow, we could construct an accurate and appropriate short-term traffic flow forecasting model. Many variables can be used to describe the traffic dynamical characteristics and reflect the traffic evolutionary process, such as traffic flow, vehicles speed, occupancy and travel time. But it is hard to collect and analyze all these variables. For many people traveling, the main problem they want to know is whether there is a traffic jam or what about the traffic condition directly. Traffic flow is the variable which reflects traffic congestion. Time occupancy is another variable showing vehicle density on roads. Taking time occupancy as exogenous variables, studying the hybrid forecasting method based on WSARIMAX model may improve the forecasting accuracy. Figure 1 shows the flow chart of the hybrid model. This flow chart can be divided into three parts, which are the wavelet de-noising and forecasting, WSARIMAX modeling, model checking and traffic flow forecasting based on the WSARIMAX model.

3.1 Wavelet de-noising and forecasting

Generally, there is noise in the original data due to the broken detectors or communication problem, etc. Noise will reduce forecasting accuracy and even give wrong

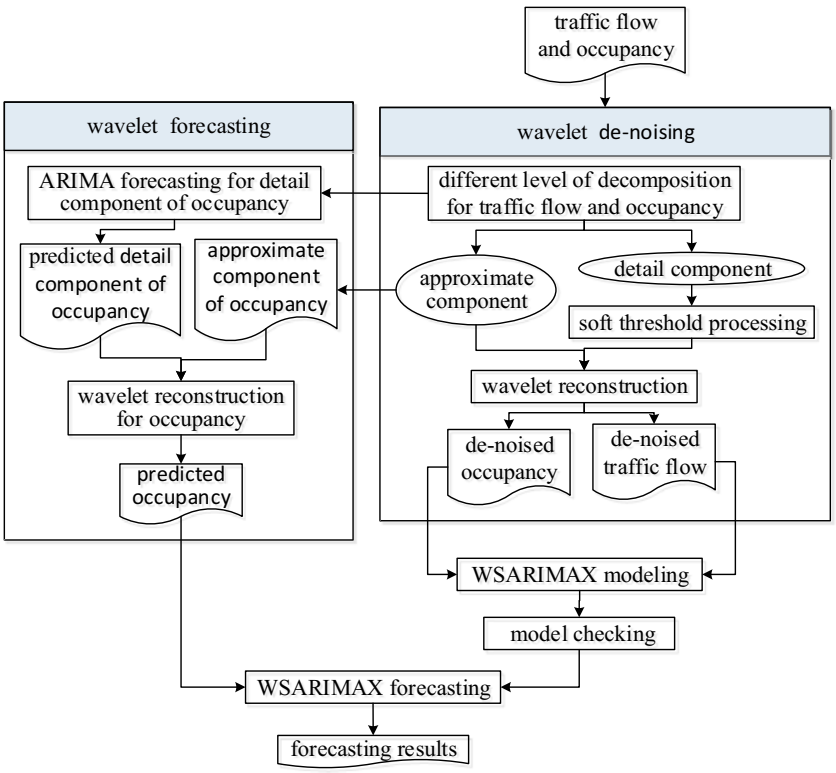


Fig. 1 Hybrid model with multivariate

forecasting results. Traffic flow is a non-stationary time series and de-noising process is critical and necessary before forecasting.

Wavelets analysis provides approximations of both stationary and non-stationary time series and expresses information both in time and frequency domain [24]. It allows us to examine the detail information of traffic data at different scales or different levels. It also has the abilities of probing short-term traffic flow characteristics, such as abrupt changes, spikes and periodic cycles and can be used to filter the noise information in traffic data and improve data quality. The maximal overlap discrete wavelet transform (MODWT), a classical wavelet transform, is well defined for all sample sizes, whereas for a complete decomposition of each levels [25]. So MODWT is well suitable for analyzing traffic data. For a time-series Y with N samples, a MODWT of level r can be described by (1):

$$\begin{aligned} Y &= \sum_{l=0}^{N-1} v_{r,t} S_{r,t+l \bmod N} + \sum_{j=1}^{r} \sum_{l=0}^{N-1} u_{j,l} W_{j,t+l \bmod N} \\ &\quad + \varepsilon_t \ (t = 1, 2, \ldots, T) \\ S_{j,t} &= \sum_{l=0}^{L_j - 1} g_{j,l} Y_{t-l \bmod N} W_{j,t} \\ &= \sum_{l=0}^{L_j - 1} h_{j,l} Y_{t-l \bmod N} \ (t = 1, 2, \ldots, T) \end{aligned} \tag{1}$$

where $S_{j,t}$ and $W_{j,t}$ are scaling and wavelet coefficients at j^{th} level respectively, $g_{j,l}$ and $h_{j,l}$ are j^{th} level scaling filter and wavelet filter respectively, $v_{r,t}$ and $u_{j,l}$ are scaling filters and wavelet filters obtained by periodizing $g_{j,l}$ and $h_{j,l}$ ε_t is an error term.

From (1), we can see that wavelet analysis makes it possible to observe the sequence Y from different time intervals and different frequencies, and allow the using of a longer time interval when low frequency information is needed and of a shorter time interval when high frequency information is needed. The noise is contained in the higher frequency series. By processing the wavelet coefficients, the larger ones are kept or shrunk and the others are set to zero to achieve the de-noising purpose [26]. A wavelet de-noising can be concluded into the following steps.

(1) Decompose a sequence Y.
(2) Choose a threshold value and apply thresholding to the detail coefficients for each level.
(3) Perform wavelet reconstruction to obtain the de-noised data.

It can be seen from (1) that if want to predict Y+1 it suffices to predict the approximation at level r, denoted by

$$\mathrm{Y}_{A_{r,t}} = \sum_{l=0}^{N-1} v_{r,t} S_{r,t+l \bmod N} (t = 1, \ldots, T), \tag{2}$$

and detail coefficients at all levels, denoted by

$$Y_{D_m,t} = \sum_{j=1}^{r} \sum_{l=0}^{N-1} u_{j,l} W_{j,t+l \bmod N} \begin{pmatrix} m = 1, \ldots, r, \\ t = 1, \ldots, T \end{pmatrix} \tag{3}$$

That is to say, it is sufficient to predict scaling and wavelet coefficients at each levels respectively, which are relatively stationary or smooth and can be forecasted with ARIMA model. Wavelet methodology for prediction of time-series data based on multiscale decomposition was developed by [27]. It can be used to forecast the non-stationary time series data. What's more, it needn't to make the difference transformation, which may lose some useful information in the data. So wavelet methodology is applied to forecast the values of occupancy, which is the values of X in WSARIMAX model. The following steps are chosen for mapping and forecasting the values of occupancy.

- Wavelet transform methodology is used to decompose occupancy time series data and obtain scaling as well as wavelet coefficients at each levels;
- ARIMA model is chosen for forecasting these scaling and wavelet coefficients;
- Inverse wavelet transform is applied to reconstruct traffic occupancy series and obtain the predicted values of occupancy
- The predicted values of occupancy are used for the input data of X in model WSARIMAX to forecast the values of traffic flow.

3.2 WSARIMAX modeling

After wavelet de-noising and wavelet forecasting, SARIMAX model is used to fit traffic flow with occupancy data acting as exogenous variables. And WSARIMAX model is constructed to forecast traffic flow with de-noised traffic flow and occupancy as training or testing data and predicted occupancy as predicted values for X.

The SARIMAX model is a generalization of seasonal ARIMAX model, which is capable of incorporating an external input variable (X) and constructing a multivariate time series model. It is in-fact a combination of AR (auto regression), MR (moving average) and integration exogenous variables X [28]. If $y_t (t = 1, 2, \cdots T)$ is a stationary time series or a nonstationary time series that can be transformed into a stationary one and $x_t (t = 1, 2, \cdots T)$

is an exogenous variable for y_t, a SARIMAX$_{(p,d,q)(P,D,Q)_s}$ model can be represented by (4):

$$\left(1-\sum_{i=1}^{p}\emptyset_i B^i\right)\left(1-\sum_{i=1}^{P}\varphi_i(B^s)^i\right)(1-B)^d(1-B^s)^D \times\left(y_t-\sum_{i=0}^{l}\vartheta_i(B)^i\right)=\left(1+\sum_{i=1}^{q}\theta_i B^i\right)\left(1+\sum_{i=1}^{Q}\psi_i(B^s)^i\right)a_t \tag{4}$$

where p and P denote the order of the non-seasonal and seasonal autoregressive (SAR) part respectively, q and Q denote the order of the non-seasonal and seasonal moving average (SMA) part respectively, d and D denote the number of non-seasonal and seasonal difference respectively, s denotes the number of seasons, y_t is a non-stationary seasonal time series, B is a delay operator. And $\sum_{i=1}^{p}\emptyset_i B^i$ is the AR part of order p, $\sum_{i=1}^{q}\theta_i(B)^i$ is the MA part of order q, $\sum_{i=0}^{l}\vartheta_i(B)^i$ is the X part of order l, $\sum_{i=1}^{P}\varphi_i(B^s)^i$ is the SAR part of order P, $\sum_{i=1}^{Q}\psi_i(B^s)^i$ is the SMA part of order Q. a_t is an innovation consisting of random variables from an uncorrelated stochastic process with zero mean and constant conditional variance.

A SARIMAX model can be used to fit a seasonal time series data with exogenous variables, which are closely related and interact with each other, and accurately forecast the future values of a time series. So it is more suitable for forecasting traffic flow correlated by vehicle speed, occupancy and so on.

3.3 Model checking and WSARIMAX forecasting

In order to obtain an adequate model and improve forecasting performance, the following steps must be carried out to estimate the model parameters.

I Identify a suitable ARIMA and SARIMAX model. The first step is to plot the traffic time series data and examine features such as trend and periodicity. It is necessary to make sure whether there is a linear trend, a curved trend, a periodicity or not. If a linear trend or a curved trend or a periodicity exists, a first order difference or a logarithmic transformation before differencing or a seasonal difference is needed to make the series stationary. Then for the stationary series, auto correlation function (ACF) and partial auto correlation function (PACF) are plotted to determine the plausible values for the parameters p, d, q, P, D, Q and l in (2). Based on maximum R^2 and minimum Akaike's information criterion (AIC) showed in (5) and (6) respectively, the best fitted models are selected.

$$R^2=1-\sum_{t=1}^{T}e_t^2\Big/\sum_{t=1}^{T}(y_t-\overline{y})^2(t=1,2,\ldots,T) \tag{5}$$

$$\mathrm{AIC}=\ln\sigma_t^2+\frac{n+2t}{n}\quad(t=1,2,\ldots,T) \tag{6}$$

where e_t is the residuals of fitted SARIMAX model, y_t is the traffic flow series, $\overline{y}=\sum_{t=1}^{T}y_t/T$, σ_t is the estimate of variance, n is the number of samples and k is the number of parameters.

II Define the method to be used to estimate the SARIMAX parameters. The most common methods are the maximum likelihood estimation method (MLEM) and non-linear least squares method (NLSM). The parameter estimation of the SARIMAX model can be done by using a software package, like SAS, MATLAB, EViews and R. In this paper, MLEM is adopted to estimation the parameters using R software.

III Make a diagnostic check to choose the best fitted as well as an adequate model through Ljung-Box-test, ACF and PACF graphs of the residuals. The model, of which the residuals show white noise, is retained. The errors of traffic flow forecasting both in-sample and out-of-sample can be checked also. The smaller the errors, the better the forecasting performance. Here, traffic features and exogenous factor are considered to improve the forecasting accuracy.

When an adequate model is obtained, the WSARIMAX model is used to forecast traffic flow with the predicted occupancy as the predicted values for X.

4 Corroboration of forecasting scheme

4.1 Data sources and description

Traffic data studied in this paper comes from the Minnesota Department of Transportation which has been responsible for collecting and publishing daily flow and occupancy data about Twin Cities' freeways. We select data from six detectors located on I-694 EB to investigate performance of above proposed model. 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 there are 18 detectors and 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. One month of flow

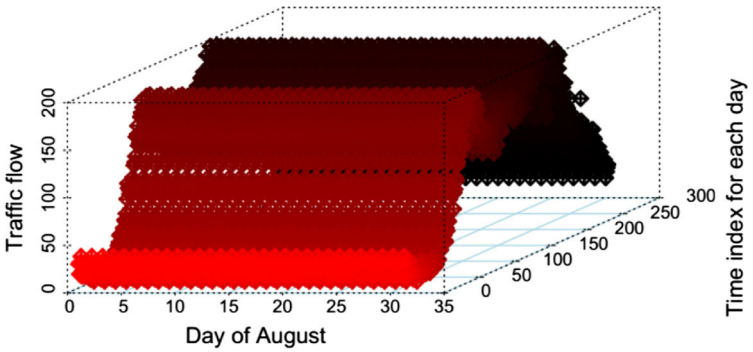
Fig. 2 Distribution of all stations and selected ones

and occupancy data, from August 1, 2016 to August 31, 2016, is selected as sample data, which has been aggregated into five-minute intervals. Thus, for each day, 288 flow and occupancy values are available (24 h× 60m/5m) as input to the prediction scheme. Partial missing data on day 20 and 21 is replaced by data calculated with moving average method.

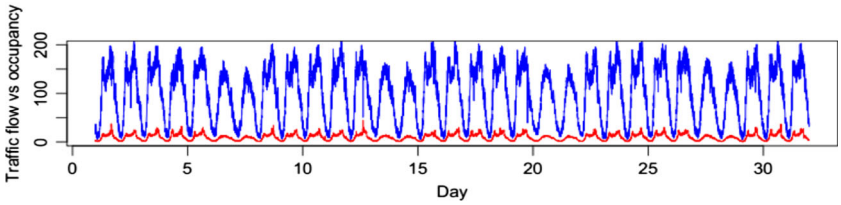
Figure 3 presents the characteristics of sample traffic data on I-694 EB in the Twin Cities from August 1, 2016 to August 31, 2016. From chart a), it can be seen that traffic flow has similar patterns over different days and different times of a day and reaches peaks at the same time index. Chart b) shows that traffic flow and occupancy in weekdays and weekend days take 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. It also can be seen that the changing trend of traffic flow and occupancy is very similar and occupancy data can be used as a supplement to traffic flow data. All these give rise to the idea that a hybrid multivariable model, a model combining time series seasonal model with wavelet analysis, may improve forecasting accuracy. So we mainly focus on developing a multivariable forecasting model based on time series seasonal model and wavelet analysis with traffic flow and occupancy data on weekdays. Because there are 8 weekend days in August 2016, 23 weekdays of traffic flow and occupancy data will be used to investigate and forecast.

4.2 Data preprocessing

As described in the above section, MODWT is used to analyze and de-noise the traffic data. First of all both variables are preprocessed to remove outliers, which are any values less than $Q_1 - 1.5 * (Q_3 - Q_1)$ or greater than $Q_3 + 1.5 * (Q_3 - Q_1)$ where Q_1 and Q_3 are the first and third quartile respectively [29]. The normal thresholds range for flow and occupancy is from 0 to 243.26 and 0 to 31.17 respectively. After many times of training and checking, 4 levels and 2 levels MODWT are respectively used to decompose occupancy and flow data. For each level of decomposed wavelet coefficients, soft threshold function method is applied to select the specified threshold



a) 3D display of traffic flow over one month



b) Traffic flow and occupancy series over one month

Fig. 3 Traffic data series from August 1 2016 to August 31, 2016

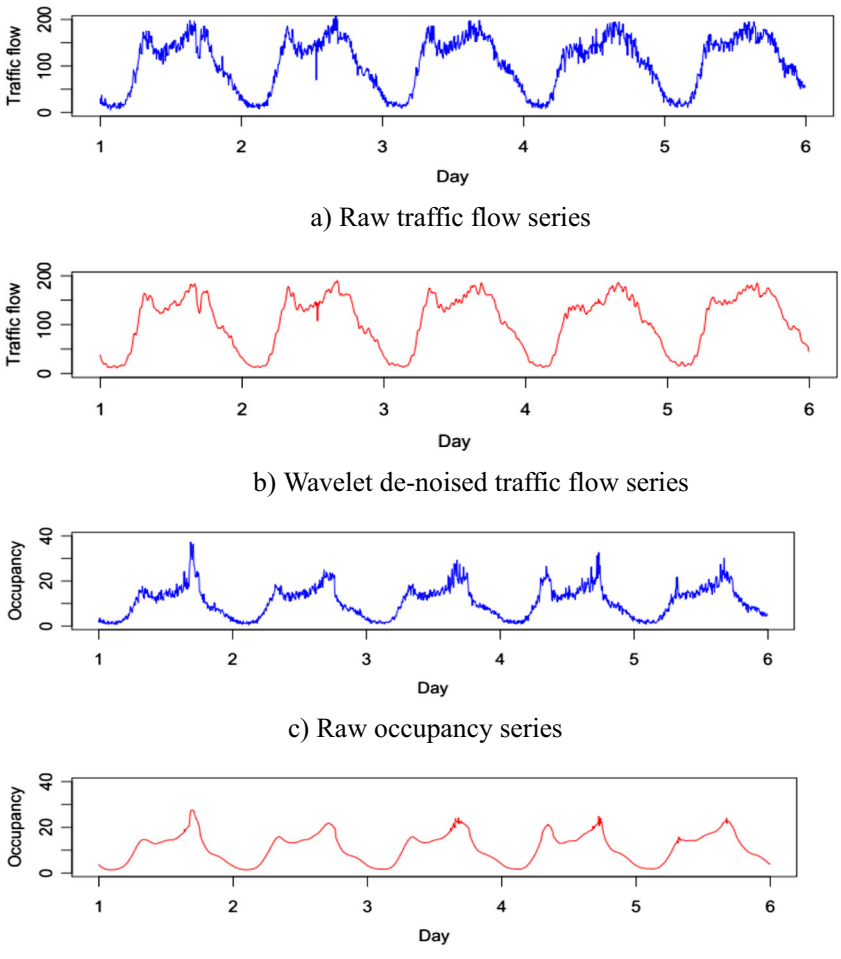
value of the decomposition coefficients to zero. Figure 4 presents 5 consecutive days of raw and wavelet de-noised data respectively.

4.3 Fitting of WSARIMAX model

After having been removed outliers and made wavelet de-noising, traffic flow data is tested for stationarity by using augmented dickey Fuller (ADF) test and phillips Perron (PP) test. On taking the first order difference at lag 288($y_t - y_{t-288}$), traffic flow series becomes stationary. Variation of flow series versus occupancy series is shown in Fig. 5. From chart a), it can be seen when values of occupancy are less than about 19, the ratio of flow and occupancy shows a direct proportional linear correlation. Chart b) further shows the changing trend of scaled traffic flow and occupancy series is very similar and their distribution takes on the state of almost overlapping. All these prove that occupancy data can be used as a supplement to traffic flow data and adopting SARIMAX model with occupancy as exogenous variables to forecast traffic flow is advisable and reasonable. On the basis of maximum R^2 and minimum AIC values, the best SARIMAX model built for the data set under consideration is SARIMAX$(2, 0, 0)(0, 1, 2)_{288}$ with the AIC of 10803.8, which is less than that of other models. The chosen model is described as following:

$$(1 + 0.87B)(1 - 0.105B^2)(1 - B^{288})(\mathrm{y}_t - 0.306x_t)$$
$$= (1 + 0.464B^{288})(1 - 0.527(B^{288})^2)a_t \quad \sigma_a^2 = 105.8$$

The estimated standard errors are 0.058, 0.058, 0.017, 0.048, and 0.048. Figure 6 shows the result of model diagnostic checking. As chart a) shows, the standardized residuals are small and evenly distributed near the zero, which are seemingly random series. And the maximum value is not more than 3. The ACF plot of the residuals indicates that a significant correlations doesn't exist in the residual series. The normal Q-Q plot of residuals indicates that the distribution of the error series of SARIMAX model satisfies normality except for only a few points at the end of



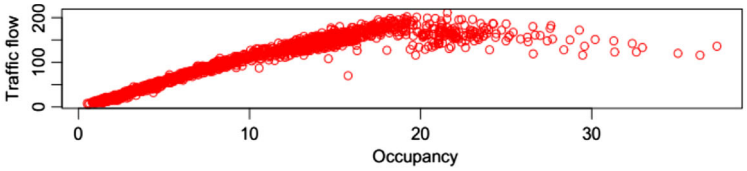
a) Raw traffic flow series

b) Wavelet de-noised traffic flow series

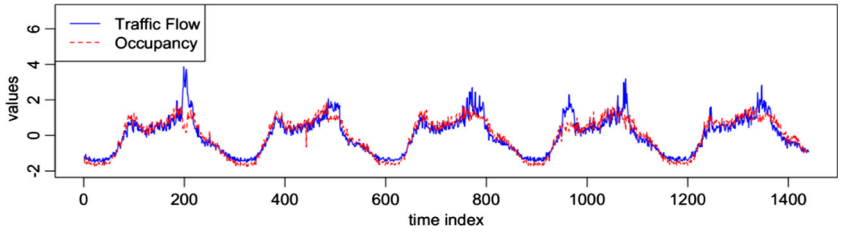
c) Raw occupancy series

d) Wavelet de-noised occupancy series

Fig. 4 5 consecutive days of raw and de-noised dada



a) Scatter plot of flow and occupancy



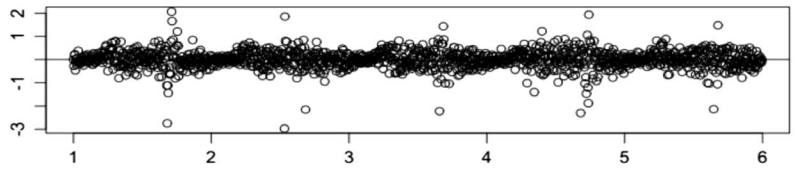
b) Scaled traffic flow and occupancy

Fig. 5 Plot of traffic flow and occupancy

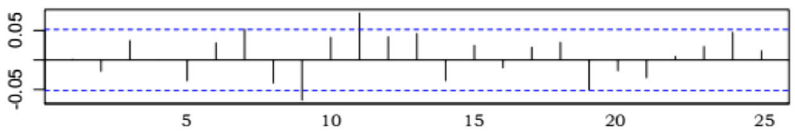
the line. All these features show that the SARIMAX model is correct and fits the data set well.

4.4 Forecasting with WSARIMAX model

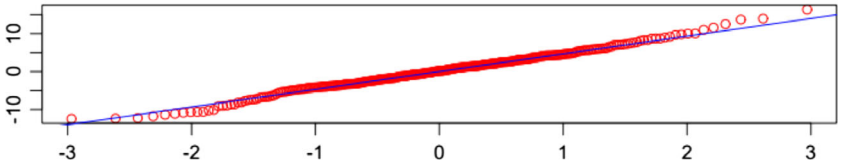
Two datasets are used to test the performance of the developed model. Figure 7 describes the one-step ahead rollback forecasting results with the proposed model. The blue solid line stands for the actual values of traffic flow, and the red points represent the predicted values. One dataset (dataset1) is the dataset described in Section 4.1, of which the first 5 days of data is used to estimate model order, and the remaining 12 days of data is used to check model, and the other 6 days of data is used to test model. Chart a) is the forecasting results using 5 consecutive days of data as input, and chart b) is the forecasting results using 3 consecutive days. The other dataset (dataset2), 5 consecutive weekdays of traffic flow and occupancy data from another highways in the Twin Cities, is applied to further check the performance of the developed model. Chart c) is the forecasting results for the fifth day using the first 4 days of data. From Fig. 7, it can be seen the predicted values closely match with the actual values in most situations, especially in non-peak period. And different days of input data yields different forecasting results, which can be seen from chart a) and chart b). Chart a) is a little better than chart b), especially on the abrupt changing points. In fact, when 2 and 3 weeks of weekday's data is used as the input, the forecasting results



a) Standardized residuals



b) ACF of residuals



c) Normal Q-Q plot of residuals

Fig. 6 Model diagnostic checking

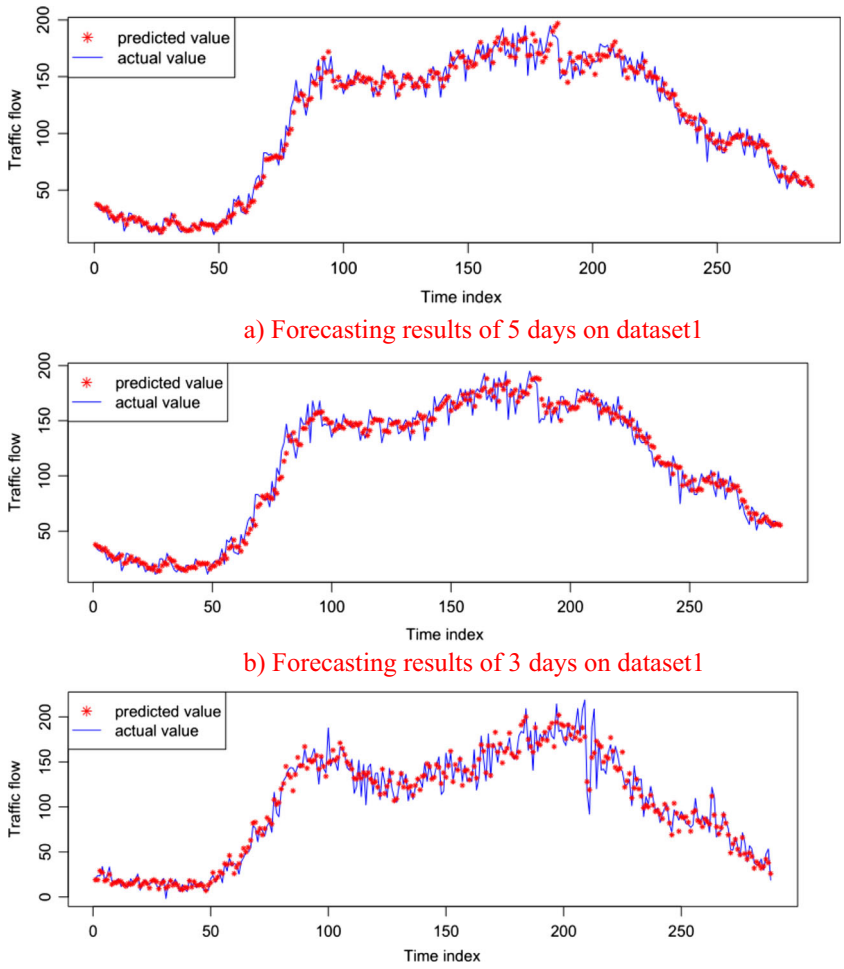


Fig. 7 Results of one-step ahead rollback forecasting

are almost same with the chart a). Thus forecasting results of a complete integral weeks of data used as the input dataset are better. From chart c), the portability of the model is better except only a few predicted values don't match with the actual values when traffic flow fluctuates rapidly. But the deviation is not far. So the proposed model can better capture the fluctuation of traffic flow and predict the changes of short-term traffic flow.

A comparison of the proposed model with SARIMA [28], WSARIMA (data preprocessed by wavelet translation acts as the input data of SARIMA), and SARIMAX is attempted to further explore the forecasting performance in terms of quantity. Forecast for traffic flow computed by SARIMA model is as following:

$$(1-0.074B)(1-0.95B^2)(1+0.098B^3)(1-B^{288})y_t = (1-0.606B^{288})(1-0.961(B^{288})^2) \times(1+0.645(B^{288})^3)a_t\sigma_a^2 = 128.2$$

Forecast for traffic flow computed by WSARIMA model is as following:

$$(1-0.196B)(1-B^{288})y_t = (1+0.167B^{288}) \times(1+0.634(B^{288})^2)(1+0.299(B^{288})^3)a_t\sigma_a^2 = 5.374$$

Table 1 RMSE values from one-step to ten-steps ahead

	Numbers of forecasting steps ahead									
	1	2	3	4	5	6	7	8	9	10
SARIMA	10.58	11.26	12.65	13.97	14.99	15.94	16.79	18.09	19.14	20.09
WSARIMA	10.54	11.05	12.43	13.57	14.57	15.36	16.24	17.45	18.66	19.65
SARIMAX	10.28	10.79	11.83	12.79	13.47	13.96	14.41	15.25	15.96	16.73
WSARIMAX	**9.21**	**10.15**	**11.23**	**12.32**	**13.04**	**13.65**	**14.11**	**14.92**	**15.83**	**16.3**

Table 2 MAE values from one-step to ten-steps ahead

	Numbers of forecasting steps ahead									
	1	2	3	4	5	6	7	8	9	10
SARIMA	7.87	8.54	9.71	10.52	11.35	11.97	12.54	13.68	14.49	15.09
WSARIMA	7.85	8.34	9.53	10.4	11.1	11.59	12.15	13.44	14.3	14.98
SARIMAX	7.67	8.21	8.94	9.54	10.05	10.48	10.79	11.4	12.01	12.25
WSARIMAX	**6.59**	**7.25**	**7.99**	**8.71**	**9.31**	**9.76**	**10.21**	**11.01**	**11.64**	**11.82**

Forecast for traffic flow computed by SARIMAX model is as following:

$$(1+0.475B)(1-0.396B^2)(1-B^{288})(y_t-0.369x_t)$$
$$=(1-0.059B^{288})(1-0.699(B^{288})^2)$$
$$\times(1+0.234(B^{288})^3)a_t\sigma_a^2=128.5$$

Two measurements of effectiveness are applied to evaluate the forecasting performance, which are root mean squared error (RMSE) and mean absolute error (MAE) showed in (7) and (8) respectively.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i)-\hat{x}(i))^2} \tag{7}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|x(i)-\hat{x}(i)| \tag{8}$$

where $x(i)$ and $\hat{x}(i)$ is actual values and predicted values respectively, n is the total number of predicted values.

Values of RMSE and MAE from one-step ahead to ten-step ahead forecasting are listed in Tables 1 and 2 respectively. RMSE and MAE values of the proposed model are marked in bold font. For one-step ahead forecasting, RMSE values of other three models are not far from each other, but RMSE value of WSARIMAX is much smaller than those of other three models. Compared with other three models, the proposed method gets improvements of 12.95%, 12.62% and 10.41% in forecasting accuracy respectively. For ten-steps ahead forecasting, the improvement in forecasting accuracy is more obvious. It gets improvements of 18.87%, 17.05% and 2.57% in forecasting accuracy. What's more, it can be seen that while forecasting steps increase. RMSE values of SARIMA and WSARIMA increase rapidly with 20.09 and 19.65 for ten-step ahead forecasting, but RMSE values of SARIMAX and WSARIMAX increase gradually with 16.73 and 16.3 for ten-step ahead forecasting. Same results that the proposed model performs well in improving forecasting accuracy and that they can capture the long-term traffic flow trend are obtained from Table 2. Since RMSE and MAE values of WSARIMAX are lower than those of other three models regardless of one-step or ten-steps ahead, the proposed model performs better in improvement of forecasting accuracy. The lower values of all the two statistics reflect the superiority of the proposed method in forecasting traffic flow.

5 Conclusions

In order to investigate the issue of how to use multiple traffic data for improving forecasting accuracy, this paper proposed a new hybrid model called WSARIMAX to forecast traffic flow by considering occupancy data as exogenous variables. Two datasets from freeways in the Twin Cities are adopted to model and evaluate forecasting performance of the proposed model. RMSE and MAE from one-step to ten-steps ahead forecasting is applied to compare the forecast performance of different models. It is found that WSARIMAX model outperforms SARIMA, WSARIMA and SARIMAX models as far as modeling and forecasting of traffic flow are concerned. For multi-steps ahead forecasting, the improvement in forecasting accuracy is more obvious. So following the conclusions are drawn.

Data de-noising before forecasting is an important step. Through wavelet de-noising, the noise information, such as short-term irregular variations, is removed and the preserved data reflects real characteristics of traffic variations. So wavelet de-noising not only avoids over-fitting or under fitting and improves forecasting performance but also decreases the parameters of modeling, which can be seen from modeling and checking of SARIMA and WSARIMA.

Multivariate predicting methods considering exogenous traffic variables are better and more accurate than univariate models. A multivariable model yields unbiased forecasting results and offsets the shortages of a univariate model, which can be seen from the results of model diagnostic checking for SARIMA and SARIMAX.

The data quality has a great influence on forecasting results. However, with the help of current information detecting techniques, the impact of data quality on forecasting results is far less than that of external factors on forecasting results, which can be seen from the error comparison of the four models. That is to say, error variations between SARIMA and WSARIMA as well as SARIMAX

and WSARIMAX are small, yet error variations between SARIMA and SARIMAX are large.

The study results are encouraging. The multivariate forecasting method has a broad application prospect and can be used for prediction of complex multivariable systems. In the future, traffic data from different roadways requires further exploration to test robustness of the proposed model.

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

Funding Information 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. Transp Res C: Emerg Technol 68:507–508
2. Rota BCR, Simic M (2016) Traffic flow optimization on freeways. Procedia Comput Sci 96:1637–1646
3. Zhang Y, Zhang Y (2016) A comparative study of three multivariate Short-Term freeway traffic flow forecasting methods with missing data. J Intell Transp Syst 20(3):205–218
4. Ghosh B, Basu B (2009) Multivariate Short-Term Traffic Flow Forecasting Using Time-Series Analysis. IEEE Trans Intell Transp Syst 10(2):246–254
5. Dong C, Shao Z, Xiong C, Zhang H (2015) A spatial-temporal-based state space approach for freeway network traffic flow modelling and prediction. Transportmetrica: A Transport Science 11(6):1–14
6. 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 Transp Technol 6:200–206
7. 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. Physica A Statistical Mechanics &, Its Applications 466:422–434
8. Cong Y, Wang J, Li X (2016) Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm. Procedia Eng 137:59–68
9. Tang J, Liu F, Zou Y, Zhang W, Wang Y (2017) An improved fuzzy neural network for traffic speed prediction considering periodic characteristic. IEEE Transactions on Intelligent Transportation Systems, 99:1:11
10. Moretti F, Pizzuti S, Panzieri S, Annunziato M (2015) Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. Neurocomputing 167(C):3–7
11. Hu W, Yan L, Liu K, Wand H (2016) A short-term traffic flow forecasting method based on the hybrid PSO-SVR. Neural Process Lett 43:155–172
12. Wang C, Ye Z (2015) Traffic flow forecasting based on a hybrid model. J Intell Transp Syst 20(5):428–437
13. Liu S, Hellendoorn H, Schutter BD (2017) Model predictive control for freeway networks based on Multi-Class traffic flow and emission models. IEEE Trans Intell Transp Syst 18(2):306–320
14. Liu S, Chen W, Chi Q, Yan H (2017) Day-to-day dynamical evolution of network traffic flow with elastic demand. Acta Phys Sin 66(6):8–22
15. Ni D (2016) Traffic Flow Theory. In: Ni D (ed) Chapter 24 multiscale traffic flow modeling. Butterworth-Heinemann, Oxford, pp 361–377
16. Wang J, Shi Q (2013) Short-term traffic speed forecasting hybrid model based on chaos–wavelet analysis-support vector machine theory. Transportation Research Part C: Emerging Technologies 27:219–232
17. Lopez-Garcia P, Onieva E, Osaba E, Masegosa A (2016) A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy. IEEE Trans Intell Transp Syst 17(2):557–569
18. Moretti F, Pizzuti S, Annunziato M, Annunziato M (2015) Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. Neurocomputing 167(C):3–7
19. Chen H, Grant-Muller S, Mussone L, F Montgomery F (2001) A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. Neural Computing &, Applications 10(3):277–286
20. Corrêa J, Neto A, Júnior L et al (2016) Time series forecasting with the WARIMAX-GARCH method. Neurocomputing 216:805–815
21. Paul R (2015) ARIMAX-GARCH-WAVELET model for forecasting volatile data. Model Assist Stat Appl & Appl 10(3):243–252
22. Zhang Y, Zhang Y (2016) A comparative study of three multivariate short-term freeway traffic flow forecasting methods with missing data. J Intell Transp Syst 20(3):205–218
23. Yin Y, Shang P (2016) Forecasting traffic time series with multivariate predicting method. Appl Math Comput 291:266–278
24. Ghosh H, Paul R, Prajneshu (2010) Wavelet frequency domain approach for statistical modeling of rainfall time-series data. Journal of Statistical Theory and Practice 4(4):813–825
25. Wenigera M, Kappa F, Friederichsa P (2017) Spatial verification using wavelet transforms: a review. Q J R Meteorol Soc 143(702):120–136
26. Lu J, Lin H, Ye D, Zhang Y (2016) A new wavelet threshold function and denoising application. Math Probl Eng 2016(3):1–8
27. Aminghafari M, Poggi J (2012) Nonstationary time series forecasting using wavelets and kernel smoothing. Communication in Statistics - Theory Methods 41(3):485–499
28. Kumar S, Vanajakshi L (2015) Short-term traffic flow prediction using seasonal ARIMA model with limited input data. European Transport Research Review 7(3):21
29. Brett l (2015) Machine learning with R - Second Edition. PACKT Publishing, Birmingham

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.

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.

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 Xi'an 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.

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.