Physica A 466 (2017) 422–434

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method

Anyu Cheng [a], Xiao Jiang [a], Yongfu Li [a,*], Chao Zhang [b], Hao Zhu [a]

[a] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[b] Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, China

HIGHLIGHTS

- A multiple sources and multiple measures based traffic flow prediction algorithm using the chaos theory and support vector regression method is proposed.
- The chaotic characteristics of traffic flow associated with the speed, occupancy, and flow are identified using the maximum Lyapunov exponent.
- The phase space of multiple measures chaotic time series are reconstructed based on the phase space reconstruction theory.
- The support vector regression (SVR) model is designed to predict the traffic flow.
- Results show that the proposed method has better performance in terms of the accuracy and timeliness.

ARTICLE INFO

Article history:
Received 1 June 2016
Received in revised form 26 August 2016
Available online 27 September 2016

Keywords:
Traffic flow
Chaos theory
Bayesian estimation theory
Support vector regression
Phase space reconstruction
Multiple measures

ABSTRACT

This study proposes a multiple sources and multiple measures based traffic flow prediction algorithm using the chaos theory and support vector regression method. In particular, first, the chaotic characteristics of traffic flow associated with the speed, occupancy, and flow are identified using the maximum Lyapunov exponent. Then, the phase space of multiple measures chaotic time series are reconstructed based on the phase space reconstruction theory and fused into a same multi-dimensional phase space using the Bayesian estimation theory. In addition, the support vector regression (SVR) model is designed to predict the traffic flow. Numerical experiments are performed using the data from multiple sources. The results show that, compared with the single measure, the proposed method has better performance for the short-term traffic flow prediction in terms of the accuracy and timeliness.

1. Introduction

Recently, traffic problems, such as traffic congestion, traffic safety and emissions, have received much attention. To address these issues, various traffic flow models have been proposed to capture the characteristics of transportation system [1–8]. In addition, one effective method is to predict the traffic state accurately and timely. It is beneficial to develop

* Corresponding author.
E-mail address: laf1212@163.com (Y. Li).

http://dx.doi.org/10.1016/j.physa.2016.09.041
0378-4371/

new methodology to migrate the traffic congestion. However, on one hand, the traffic state involves multiple measures such as speed, occupancy, and flow; On the other hand, the inherent nonlinearity and complexity of traffic dynamics impede the real-time prediction accuracy of the traffic flow time series [9]. Hence, it is a challenge to deal with the characteristics of multiple measures and nonlinearity of traffic flow for the prediction.

This study is motivated by the need for a more accurate and timely algorithm so that the traffic state evolution can be predicted effectively. One feature of traffic flow associated with the nonlinearity is the chaotic phenomena. It means that small differences in the initial condition would yield large difference in the prediction results. To identify the chaotic characteristic of traffic flow, a common approach is to calculate the maximum Lyapunov exponent of time series [10]. Several methods have been proposed to calculate the maximum Lyapunov exponent. They include the Wolf method [11], the Jacobian method [12], the P norm algorithm [12], the singular value decomposition method [13], and the small-data method [14] as well as the wavelet transform [15]. Regarding the above mentioned methods, small-data based method has been widely used for the reliability and computational complexity. Consequently, this study will identify the chaotic characteristic of traffic flow using the small-data method.

In the literature, traffic flow prediction models can be roughly classified as linear model [16], nonlinear model [17], intelligent model [18], and other prediction model. From the linear model perspective, it can be further classified as historical average prediction model [19], time series prediction model [20], and Kalman filter model [21]. From the nonlinear model perspective, it can be further classified as the wavelet analysis based model [22] and the catastrophe theory based model [23]. From the other prediction model perspective, it can be further classified as the cellular automata model [24], the fuzzy regression model [25] and the nonparametric regression model [26]. The above mentioned traffic flow prediction models are relatively mature. However, these models restrict the application scope in terms of the accuracy as they do not consider the chaotic characteristics of traffic flow. Therefore, the artificial intelligent (AI) theory based intelligent prediction models has been proposed considering the chaotic characteristics of traffic flow. Since these intelligent models do not rely on the precise mathematical expressions between the dependent and independent variables, so they have strong learning ability and robustness. Hence, this study will adopt the intelligent model, i.e., support vector regression (SVR) model, to predict the traffic flow.

The SVR-based prediction model is a typical intelligent model as it has the potential to identify a highly nonlinear system for its ability to approximate complex nonlinear systems. Hence, SVR-based method plays an important role in short-term traffic flow prediction [27]. In this research line, Hu et al. [28] proposed an algorithm by combining the SVR and particle swarm optimization (PSO) for traffic flow prediction. Ahn et al. [29] proposed a real-time traffic flow prediction based on the Bayesian classifier and SVR. The results showed that the performance of the approach using the SVR-based estimation is better than that of the linear-based regression. Liu [30] further verified the effectiveness of the method by combining the phase reconstruction and SVR. Li et al. [31] proposed the method using the SVR model with Gauss loss function, i.e., Gauss-SVR, to forecast urban traffic flow. Wei et al. [32] proposed an adaptive SVR short-term traffic forecasting model. The results showed that the computational efficiency can be improved, compared with the standard SVR model. In summary, SVR model can avoid the local optima and slow convergence rate. Thereby, SVR model has been widely used for short-term traffic flow prediction.

The SVR models mentioned heretofore focus on the time series prediction of a single traffic measure. Considering the nonlinearity and complexity of traffic system, a single traffic measure is difficult to fully capture the characteristics of the system. Hence, it motivates us to characterize the traffic system using multiple measures such as average speed, average occupancy, and average traffic flow [33]. As a result, this study proposes a multiple sources and multiple measures based traffic flow prediction algorithm using chaos theory and support vector regression method. In particular, first, the chaotic characteristics of traffic flow associated with the speed, occupancy, and flow are identified using the small-data based the maximum Lyapunov exponent calculation. Then, the phase space of multiple measures chaotic time series are reconstructed through the selection of the embedding dimension and delay time based on the phase space reconstruction theory, and fused into a same multi-dimensional phase space using the Bayesian estimation theory. In addition, the SVR model is designed to predict the change trend of the time series of traffic states. Numerical experiments are performed using the data from multiple sources. The results show that, compared with the single measure, the proposed method has better performance for the short-term traffic flow prediction in terms of the accuracy and timeliness.

The rest of this paper is organized as follows. Section 2 presents the methodology used in this study, including the framework, chaotic characteristics identification, phase space reconstruction, Bayesian estimation theory based phase point fusion, and SVR based prediction model. Section 3 performs the numerical experiments to evaluate the performance of the proposed method. Section 4 concludes this study.

2. Methodology

In this paper, we study the object for the average speed, average occupancy and average traffic flow time series. These three measures can be integrated into a single time series of the fused traffic flow through the multiple sources integration method [33,34], so that the predicted traffic flow can be dynamically updated more accurately due to the correlation between these three measures. Hence, to improve the prediction accuracy, the multiple sources and multiple measures based traffic flow prediction model will be explained hereafter.

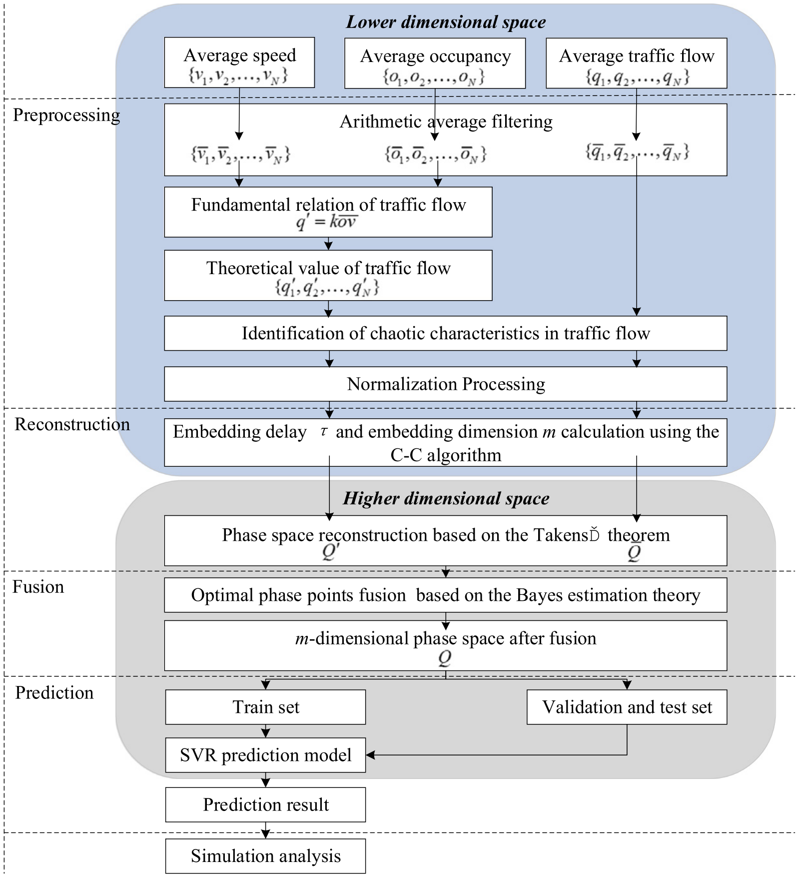


Fig. 1. The proposed framework.

2.1. Framework

We propose a two-level framework to predict the traffic flow as shown in Fig. 1. The proposed framework includes two components. The first component processes and normalizes the data, identifies of chaotic characteristics, calculates the embedding dimension m and delay τ in a lower dimensional space, and the second component reconstructs the data, fuses the multiple measures, builds the SVR prediction model and predicts traffic flow in a higher dimensional space.

2.2. Analyzing and judging the chaotic characteristics of traffic flow

Considering the chaotic characteristics of traffic system, it is possible to study the change trend of traffic flow with the method of chaotic dynamics theory. Therefore, in this paper, the maximum Lyapunov exponent of multiple measures, such as average traffic flow and theoretical traffic flow, are first calculated based on small-data method [14] so as to verify the traffic system's chaotic characteristics.

The average period P of traffic flow time series $\{x(t)|t = 1, 2, \ldots, n\}$ is first calculated based on the fast Fourier transform [9]. The embedding dimension m and delay time τ of phase space reconstruction are calculated based on the C–C algorithm [35]. As a result, we obtain the phase space $\{X_i \mid i = 1, 2, \ldots, K\}$ of the traffic time series, where $K = n-(m-1)\tau$ is the number of phase points. Then, the nearest neighbor phase point $X_{\hat{i}}$ of each phase point X_i in the phase space can be distinguished and meanwhile they are limited to brief separation, which implies that:

$$d_i(0) = \min_{X_{\hat{i}}} \|X_i - X_{\hat{i}}\| \quad (|i - \hat{i}| > p). \tag{1}$$

The neighboring phase point of the distance $d_i(j)$ after the j discrete time step for each phase point X_i in phase space is calculated as follows:

$$d_i(j) = |X_{i+j} - X_{\hat{i}+j}| \quad (j = 1, 2, \ldots, \min(K - i, K - \hat{i})). \tag{2}$$

Finally, the average $y(i)$ of all $\ln(d_i(j))$ for each j is calculated by:

$$y(i) = \frac{1}{q\Delta t}\sum_{i=1}^{q}\ln(d_i(j)). \tag{3}$$

In Eq. (3), q is the number of non-zero $d_i(j)$. Using the least squares regression method, the maximum Lyapunov exponent λ is the slope of the straight line. If $\lambda > 0$, the corresponding time series of traffic flow is chaotic. The method makes full use of the spatial information of time series, and consequently it is more reliable compared with the small data set based method.

2.3. Multi-measure phase space reconstruction

Phase space reconstruction is regarded as the basis of chaotic time series analysis and widely used in nonlinear system analysis [34]. Takens [36] proposed the delay coordinates method of phase space reconstruction for chaotic time series analysis. According to the literature [33], suppose a time series represented by $[Y_1, Y_2, \ldots, Y_N]^{\mathrm{T}}$ contains N measures, each element Y_i can be viewed as an individual time series consisting of n data points, i.e., $Y_i = \{y_{i,t} | t = 1, 2, \ldots, n\}$. Based on the Takens' embedding theorem [36], a chaotic nonlinear time series can be reconstructed as a recovery phase space of the original dynamical system by selecting the embedding dimension m and delay τ. More specifically, for each time series Y_i of measure $i = 1, 2, \ldots, N$, its optimal embedding dimension m_i and delay time τ_i can be determined using the C–C algorithm [35].

For each measure $i = 1, 2, \ldots, N$, the corresponding time series $Y_i = [y_{i,1} \quad y_{i,2} \quad \cdots \quad y_{i,n}]$can be reconstructed into a phase space represented by:

$$Z_i = \begin{bmatrix} y_{i,1} & y_{i,1+\tau} & \cdots & y_{i,1+(m-1)\tau} \\ \vdots & \vdots & \vdots & \vdots \\ y_{i,k} & y_{i,k+\tau} & \cdots & y_{i,k+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i,K} & y_{i,K+\tau} & \cdots & y_{i,K+(m-1)\tau} \end{bmatrix}. \tag{4}$$

The kth row of Z_i is:

$$Z_{i,k} = (y_{i,k}, y_{i,k+\tau}, \ldots, y_{i,k+(m-1)\tau}) \quad \forall k = 1, 2, \ldots, K, \tag{5}$$

where $Z_{i,k}$ represents the kth phase point of the ith time series data, the number of phase points K is determined by $n - (m-1)\tau$, and all rows together represent the m-dimensional phase space.

2.4. Phase point fusion in phase space based on Bayesian estimation

This section introduces a multiple sources data fusion method to fuse the information on multiple dimensions into a single time series to improve the short-term prediction accuracy. In general, the reconstructed phase space of a single measure can approximately prescribe the characteristic of the attractor of the original dynamical system [33]. However, due to the data incompleteness, a single-measure time series is unable to comprehensively reveal the characteristics of the system in other dimensions. Multiple sources data fusion method fuses the multiple-measure time series into a single one, which integrates the system characteristics associated with the different dimensions.

This study adopts a multivariate data fusion method based on the Bayesian estimation theory [33]. Bayesian estimation based data fusion methods make full use of the prior information of the measures and sample information. Thereby, Bayesian estimates in parameter estimation usually have smaller mean square error [33,34]. By applying the Bayesian estimation based multiple sources data fusion method, the phase points of multiple measures can be fused into a single one such that the characteristics provided by the phase points can complement each other and mitigate the measurement errors in a single measure. By doing so, the estimation results can better approximate the true state of the system.

Following the method proposed in Ref. [33], we can obtain a new fusion phase space with m-dimensions that can be achieved with the following form

$$F = [F_1, \ldots, F_k, \ldots, F_K]^{\mathrm{T}}. \tag{6}$$

Each F_k in Eq. (6) can be expressed as:

$$F_k = (f_k, f_{k+\tau}, \ldots, f_{k+(m-1)\tau}) \quad \forall k = 1, 2, \ldots, K \tag{7}$$

where K represents the number of phase space, and k represents the arbitrary coordinate point in the time series. Each phase point F_k includes the main features of every single variable in the new phase space F, and is able to approach the true state of traffic flow. Therefore, the new fusion phase space can accurately characterize the traffic flow.

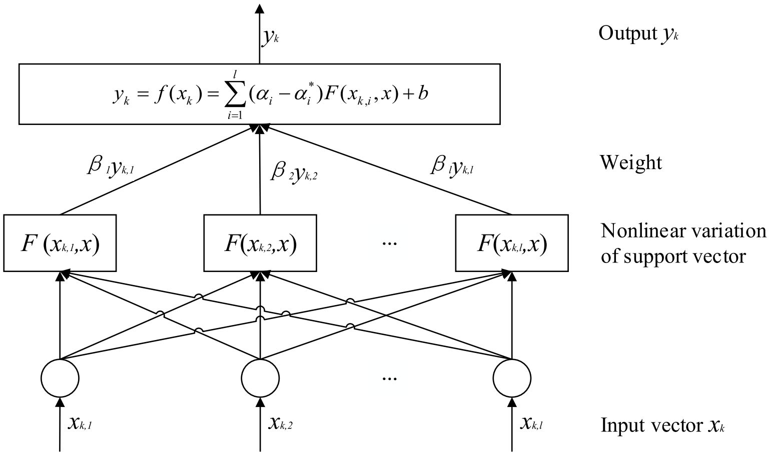


Fig. 2. SVR prediction structure based on phase space reconstruction.

Table 1
Input sets and output sets.

Input sets	Output sets
$x_1 = (f_1, f_{1+\tau}, \ldots, f_{1+(m-1)\tau})$	$y_1 = f_{1+(m-1)\tau+1}$
$\vdots$	$\vdots$
$x_k = (f_k, f_{k+\tau}, \ldots, f_{k+(m-1)\tau})$	$y_k = f_{k+(m-1)\tau+1}$
$\vdots$	$\vdots$
$x_K = (f_K, f_{K+\tau}, \ldots, f_{K+(m-1)\tau})$	$y_K = f_{K+(m-1)\tau+1}$

2.5. SVR based prediction model

Based on the support vector machine (SVM) theory [37], SVR is to approximate the given observations in an m-dimensional space by a linear function in another feature space. As shown in Fig. 2, the entire SVR prediction structure includes three-part: input vector, nonlinear variation of support vector and output.

In this study, the proposed SVR prediction model regards the phase point of the fusion phase space as the input vector. By Eq. (7), the input sets and output sets of the SVR prediction model based on the phase space reconstruction are show in Table 1.

The vector $x_k = (f_k, f_{k+\tau}, \ldots, f_{k+(m-1)\tau})$ is the input vector of the model, $\beta = (\beta_1, \beta_2, \ldots, \beta_l)$ is the vector of output weight, and $y_k = f_{k+(m-1)\tau+1}$ is the output of the model. In this paper, we use the ε-insensitive loss function in support vector machine and namely get the ε-SVR model, and adopt Gaussian radial basis function $F(x_k, x) = \exp(-\gamma \|x_k - x\|^2)$ as the ε-SVR kernels. The generic ε-SVR estimation function takes the form:

$$y_k = f(x_k) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)F(x_{k,i}, x) + b \tag{8}$$

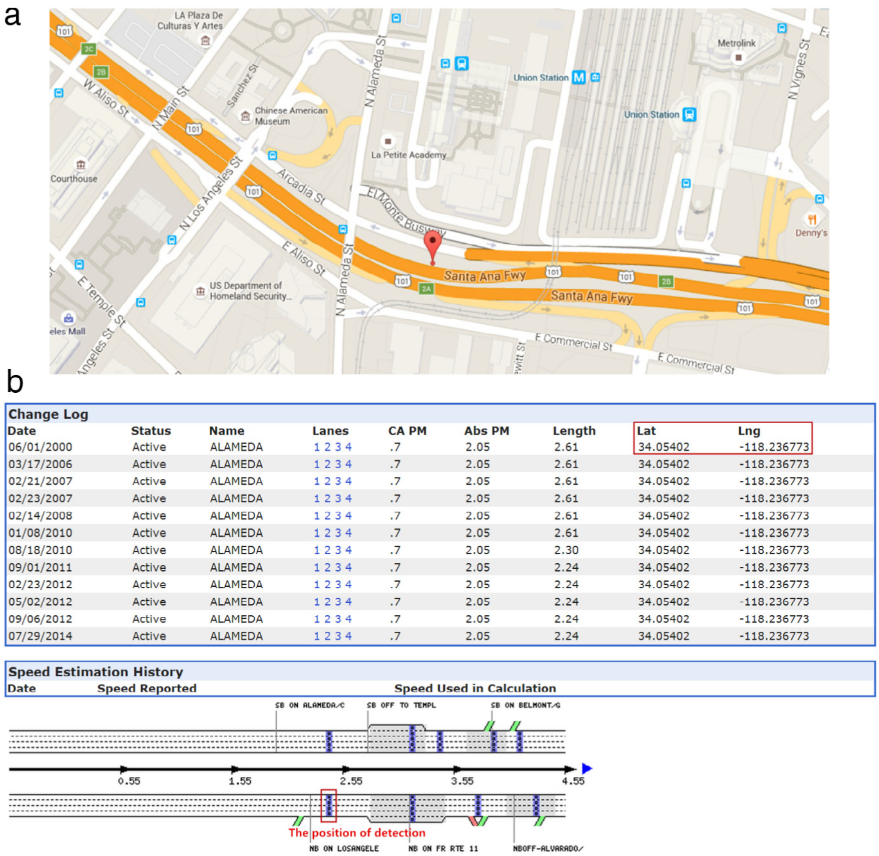
where α_i and α_i^* are the Lagrange multipliers, $F(x_k, x)$ is the kernel function to transpose into high-dimensional feature space using low-dimensional space data without knowing the transformation.

For ε-SVR, an important issue is the selection of model parameters, i.e., penalty C, radius ε and the kernel function parameter λ. Therefore, we adopt genetic algorithm (GA) [38] to optimize the parameters of ε-SVR.

To use the SVR model to predict the chaotic time series, the kth input vector is the kth phase point of the fusion phase space, and the dimension of input vector is equal to the chaotic time series in the reconstructed phase space of embedding dimension m, and the output data is the last dimension of the $(k+1)$th input vector. Similar to Ref. [33], three indices are used to evaluate the prediction accuracy: mean absolute error (*MAE*), mean absolute relative error (*MARE*), and equal coefficient (*EC*).

3. Numerical example

To investigate the effectiveness of the proposed short-term traffic state prediction method, a numerical experiment is implemented based on the traffic data collected by the loop detector at US101-N @ CA PM.7 (Abs PM 2.05) District 7, Los Angeles County, City of Los Angeles through the California's Freeway Performance Measurement System (PeMS) [39].



Date	Status	Name	Lanes	CA PM	Abs PM	Length	Lat	Lng
06/01/2000	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
03/17/2006	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
02/21/2007	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
02/23/2007	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
02/14/2008	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
01/08/2010	Active	ALAMEDA	1 2 3 4	.7	2.05	2.61	34.05402	-118.236773
08/18/2010	Active	ALAMEDA	1 2 3 4	.7	2.05	2.30	34.05402	-118.236773
09/01/2011	Active	ALAMEDA	1 2 3 4	.7	2.05	2.24	34.05402	-118.236773
02/23/2012	Active	ALAMEDA	1 2 3 4	.7	2.05	2.24	34.05402	-118.236773
05/02/2012	Active	ALAMEDA	1 2 3 4	.7	2.05	2.24	34.05402	-118.236773
09/06/2012	Active	ALAMEDA	1 2 3 4	.7	2.05	2.24	34.05402	-118.236773
07/29/2014	Active	ALAMEDA	1 2 3 4	.7	2.05	2.24	34.05402	-118.236773

Fig. 3. Data collection location: (a) coordinates of data collection location; (b) loop detector location at US101-N in PeMS.

The data set contains information on average speed ($\{v_1, v_2, \ldots, v_n\}$, measured by radar detectors), average occupancy ($\{o_1, o_2, \ldots, o_n\}$, measured by loop detectors), and average traffic flow ($\{q_1, q_2, \ldots, q_n\}$, vehicle count). Each traffic measure is regarded as one measure in the multi-source time series for describing the evolution of traffic state. Data groups are generated every 5 min containing the measurements of each traffic measure. Thereby, 288 data groups are generated every day. The data collected from December 28 to December 31, 2015 consisting of 1152 data groups is used in the numerical experiment. Fig. 3 illustrates the data collection location.

To reduce the measurement noises, an arithmetic average filtering [40] is used to process the raw data. Fig. 4(a)–(c) illustrate the original time series and the post-processing time series. Fig. 4 shows that the evolution patterns of different traffic measures differ from each other.

Based on the method in [33], the fundamental relation of traffic flow theory is:

$$q' = \frac{\bar{o}}{l+d}\bar{v}. \tag{9}$$

In this experiment, $l + d$ is assumed to be constant. According to Eq. (9), the theoretical traffic flow q' can be computed in terms of the measured average speed $\bar{v}$ and average occupancy $\bar{o}$. In addition, considering the inconsistent dimensions among average traffic flow and theoretical traffic flow, Fig. 5 shows the measured average traffic flow q and the theoretical traffic flow q' which are normalized to the range $[0, 1]$. From Fig. 5, the overall shape of the measured traffic flow time series is similar to that of the theoretical traffic flow time series, however, the individual features of the measured and theoretical traffic flow also exist.

3.1. Selection of parameters in reconstructing phase space

First, the multiple measures are normalized to a range of [0–1]. Second, the embedding dimension m and delay τ of the measured average traffic flow time series and the theoretical traffic flow time series are calculated using the C–C algorithm [35]. The C–C algorithm defines the statistic variable based on the time series correlation integral. Following

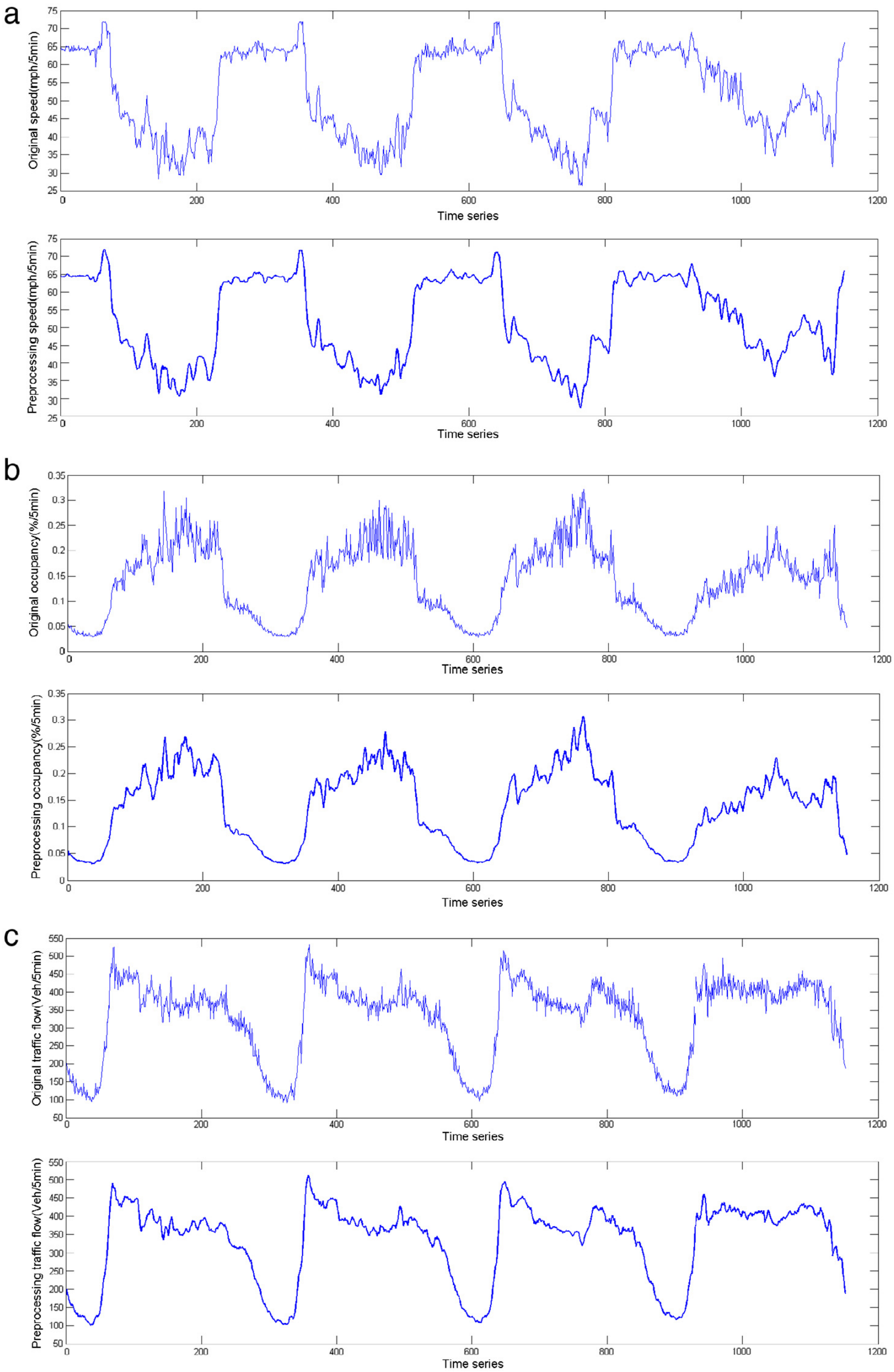


Fig. 4. Time series data: (a) Average speed; (b) Average occupancy; (c) Average traffic flow.

the C–C algorithm, the time series correlation integral $C(m, N, r, t)$ is defined as:

$$C(m, N, r, t) = \frac{2}{M(M-1)} \sum_{1 \le i < j \le M} \theta(r - d_{ij}), \quad r > 0 \tag{10}$$

where $M = N - (m-1)t$, $d_{ij} = \|X_i - X_j\|$, $\theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \ge 0 \end{cases}$.

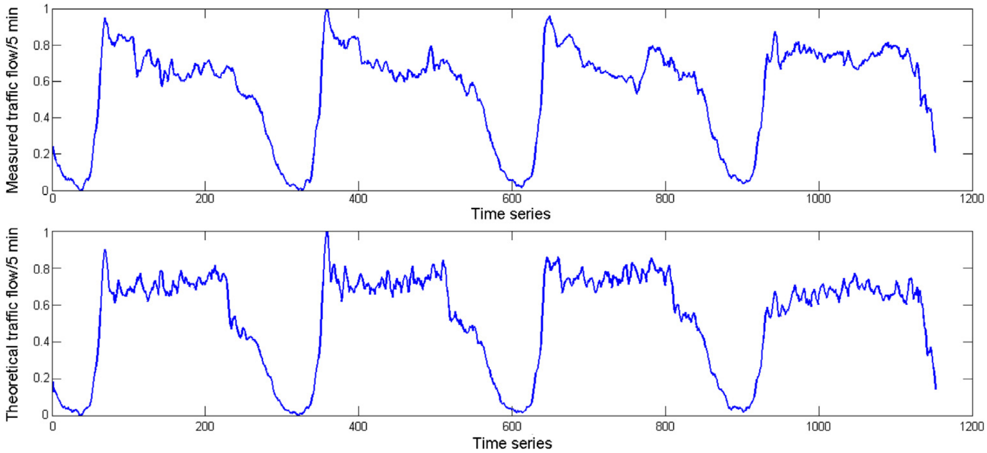


Fig. 5. The measured and theoretical traffic flow time series.

Table 2
Parametric values based on time series data.

Measures	τ	τ_w	m
$q(t)$	21	79	5
$q'(t)$	22	81	5

The statistic variable $S_2(m, N, r, t)$ is defined as, i.e., if $N \to \infty$

$$S_2(m, r, t) = \frac{1}{t}\sum_{s=1}^{t}[C_s(m, r, t) - C_s^m(1, r, t)]. \tag{11}$$

Selecting the maximum and minimum radius r, the deviation of the statistic variable is defined as:

$$\Delta S_2(m, t) = \max\{S_2(m, r, t)\} - \min\{S_2(m, r, t)\}. \tag{12}$$

$\Delta S_2(m, t)$ measures the maximum deviation of $S_2(m, r, t) \sim t$ for all the radii r. Based on BDS statistical conclusion [35], N, m, and r can be reasonably estimated, and make an assumption that $N = 3000$, $m = 2, 3, 4, 5$ and $r_i = i \times 0.5\sigma$, where $\sigma = std(x)$ is the standard deviation of the time series and $i = 1, 2, 3, 4$, calculation:

$$\bar{S}_2(t) = \frac{1}{16}\sum_{m=2}^{5}\sum_{i=1}^{4} S_2(m, r_i, t) \tag{13}$$

$$\Delta\bar{S}_2(t) = \frac{1}{4}\sum_{m=2}^{5}\Delta S_2(m, t). \tag{14}$$

According to the correlation of $\Delta\bar{S}_2(t)$ and time t, the optimal delay τ can be calculated as it is the time that corresponds to the first minimum value calculated in terms of $\Delta\bar{S}_2(t)$. Considering the $\bar{S}_2(t)$ and $\Delta\bar{S}_2(t)$, a new index is defined as:

$$S_{2cor}(t) = \Delta\bar{S}_2(t) + |\bar{S}_2(t)|. \tag{15}$$

Finally, the optimal embedding window τ_w can be calculated as it is the time that corresponds to the global minimum value calculated in terms of $S_{2cor}(t)$. Fig. 6 shows the delay time and embedding window associated with the measured time series.

According to the modified C–C algorithm, the values of delay time τ and embedding window τ_w can be estimated simultaneously through the correlation integral method. Then, the embedding dimension m is dependent on delay time τ and embedding window τ_w, and can be calculated by $\tau_w = (m - 1)\tau$. As shown in Fig. 6, the optimal delay time τ is that which corresponds to the first minimum value calculated in terms of $\Delta\bar{S}_2(t)$ and the embedding window τ_w is that which corresponds to the minimum value calculated in terms of $S_{2cor}(t)$. Table 2 shows the delay time τ, the embedding window τ_w, and the embedding dimension m of measured average traffic flow and theoretical traffic flow.

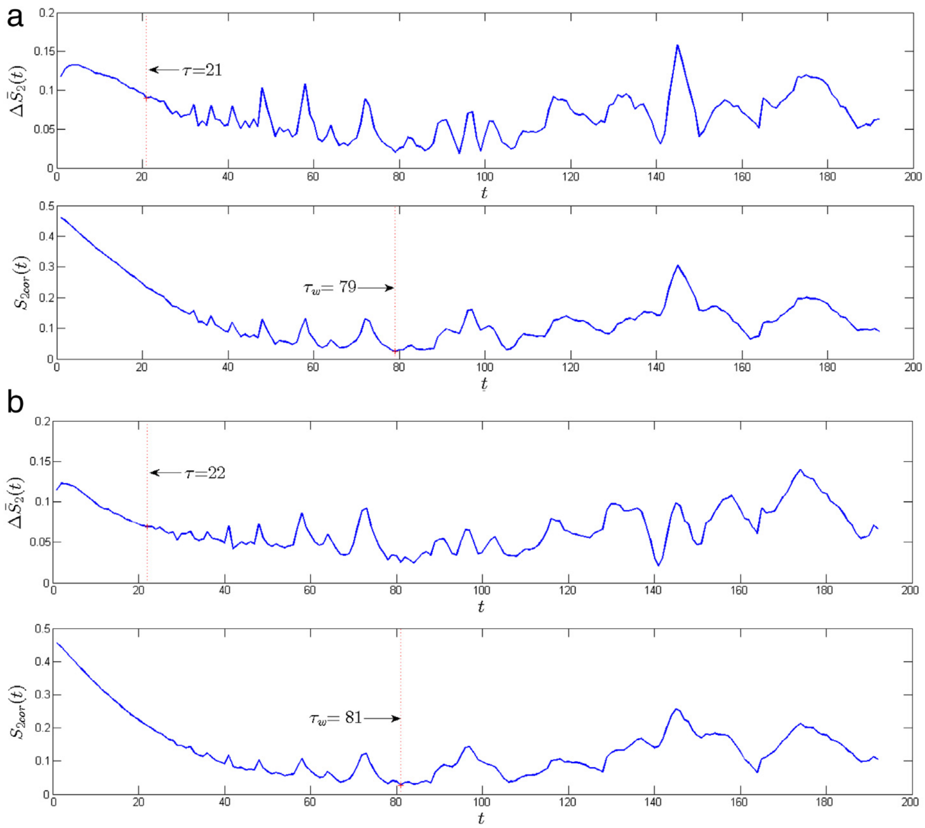


Fig. 6. Delay time and embedding window of time series using C–C algorithm: (a) Measured average traffic flow; (b) Theoretical traffic flow.

Table 3
The maximum Lyapunov exponent of time series.

Measures	P	λ_1
$q(t)$	288	0.5234
$q'(t)$	288	0.4135

3.2. Identification of chaotic characteristics in traffic flow

In this paper, from the perspective of nonlinear time series, the maximum Lyapunov exponent for judging chaos characteristics of traffic flow is calculated by the small-data method [14] based on chaotic dynamics theory, which provides the foundation for the prediction of short-term traffic flow. First, the average period P of time series, including the measured average traffic flow and the theoretical traffic flow, is calculated by the fast Fourier transform (FFT) [9]. Then, the maximum Lyapunov exponent of multiple measures is calculated based on small-data method and C–C algorithm is adopted to calculate the embedding delay τ and dimension m. The estimated value of maximum Lyapunov exponent of multiple measures is shown respectively in Fig. 7.

The slope of the least squares regression line is the estimated value of the maximum Lyapunov exponent λ_1 in Fig. 7. In Fig. 7(a), the expression of the least squares regression line is $y = 0.5234x + 79.6075$, so the maximum Lyapunov exponent of measured average traffic flow is $\lambda_1 = 0.5234$. In Fig. 7(b), the expression of the least squares regression line is $y = 0.4135x + 101.8$, so the maximum Lyapunov exponent of theoretical traffic flow is $\lambda_1 = 0.4135$. The estimated value of the maximum Lyapunov exponent of the multiple measures time series is shown in Table 3, where all $\lambda_1 > 0$ indicate that the multiple measures are chaotic in this paper.

3.3. Fusion and prediction of time series data of multiple measures

To capture the traffic system characteristics from different data sources, this study focuses on multiple measures rather than any single measure. According to Table 2, to capture all attributes of multi-measure, we select the maximum embedding dimension $m = 5$ and the minimum delay time $\tau = 21$ as the values of phase space reconstruction parameters. Based on Eq. (9) and Fig. 5, the theoretical traffic flow can be calculated in terms of the average speed and average occupancy. Then, the

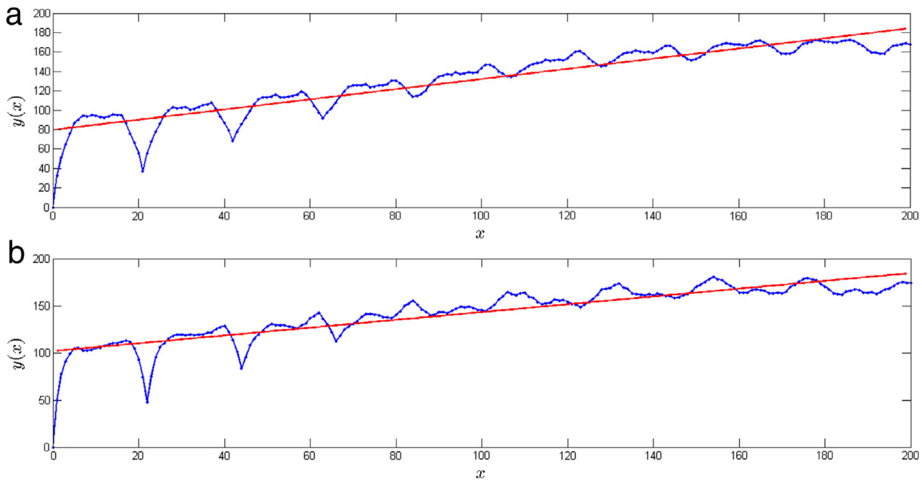


Fig. 7. The maximum Lyapunov exponent of time series using small-data method: (a) Measured average traffic flow; (b) Theoretical traffic flow.

Table 4
Prediction performance comparisons.

Predicted data	*MAE*	*MARE*	*EC*
Measured traffic flow	0.0224	0.0383	0.9795
Fusion traffic flow	0.0148	0.0227	0.9862

groups of the measured and theoretical traffic flow time series are embedded into a higher dimensional space. Consequently, we can obtain an optimal fusion phase space based on the fusion of the phase points of the two groups of measures, according to Eq. (6), in the higher dimensional space.

As shown in Fig. 8, Fig. 8(a) is the phase space reconstruction diagram of the measured average traffic flow time series, and local characteristics 1, 2 and 3 are marked on it. Fig. 8(b) is the phase space reconstruction diagram of the theoretical traffic flow time series, and local characteristics 4, 5, and 6 are marked on it. Fig. 8(c) is the fusion phase space reconstruction diagram of the multi-source time series, and local characteristics 7, 8, and 9 are marked on it. The local characteristics 1 and 3 of the measured average traffic flow time series and the local characteristics 4 and 6 of the theoretical traffic flow time series are very similar to the characteristics 7 and 9 of the multi-source fusion time series, respectively. However, local characteristic 2 of the measured average traffic flow time series is different from local characteristic 5 of the theoretical traffic flow time series, but the local characteristic 8 of the fusion phase space reconstruction diagram of the multi-source time series reflects them at the same time. Therefore, the multi-source fusion time series phase space contains all the main features of the measured and theoretical traffic flow time series. Based on the above analyses, the proposed method can reflect the characteristics of traffic system effectively. In addition, compared to an individual measure, the proposed method can elicit richer information. Consequently, it can reflect the main characteristics of traffic system accurately.

According to Table 1, we can get the input sets and output sets of the SVR prediction model based on the multi-source fusion time series phase space, and get the number of phase points $K = 1068$, which means that the total number of input sets and output sets is 1068 sets of data. We select 900 groups of data as the train sets and 168 groups of data as the prediction validation sets. Based on the train sets, the phase space reconstruction based SVR prediction model can be trained. As shown in Fig. 9 are the prediction result of the measured average traffic flow and the multi-source fusion time series based on the trained SVR prediction model. To evaluate the prediction performance, errors associated with the measured average traffic flow and the multi-source fusion time series using the proposed framework in terms of *MAE*, *MARE*, and *EC* are shown in Table 4.

As shown in Fig. 9, it can be seen that both prediction change trend based on the measured average traffic flow and the theoretical traffic flow are in accordance with the practical trend. As shown in Table 4, the prediction error evaluation index $MAE < 0.03$, $MARE < 0.04$ and $EC > 0.97$ show that the fitting effect of the actual and estimated values based on SVR prediction model are well. All of the analysis above has demonstrated that the proposed framework can predict traffic flow with high reliability and relatively high accuracy in terms of *MAE*, *MARE*, and *EC*.

4. Conclusions

To evaluate the effects of chaotic characteristics on traffic flow prediction, this study proposes a multiple sources and multiple measures based algorithm using the chaos theory and support vector regression method. To this end, the chaotic characteristics of the multi-measure, including the average speed, average occupancy and average traffic flow, are identified

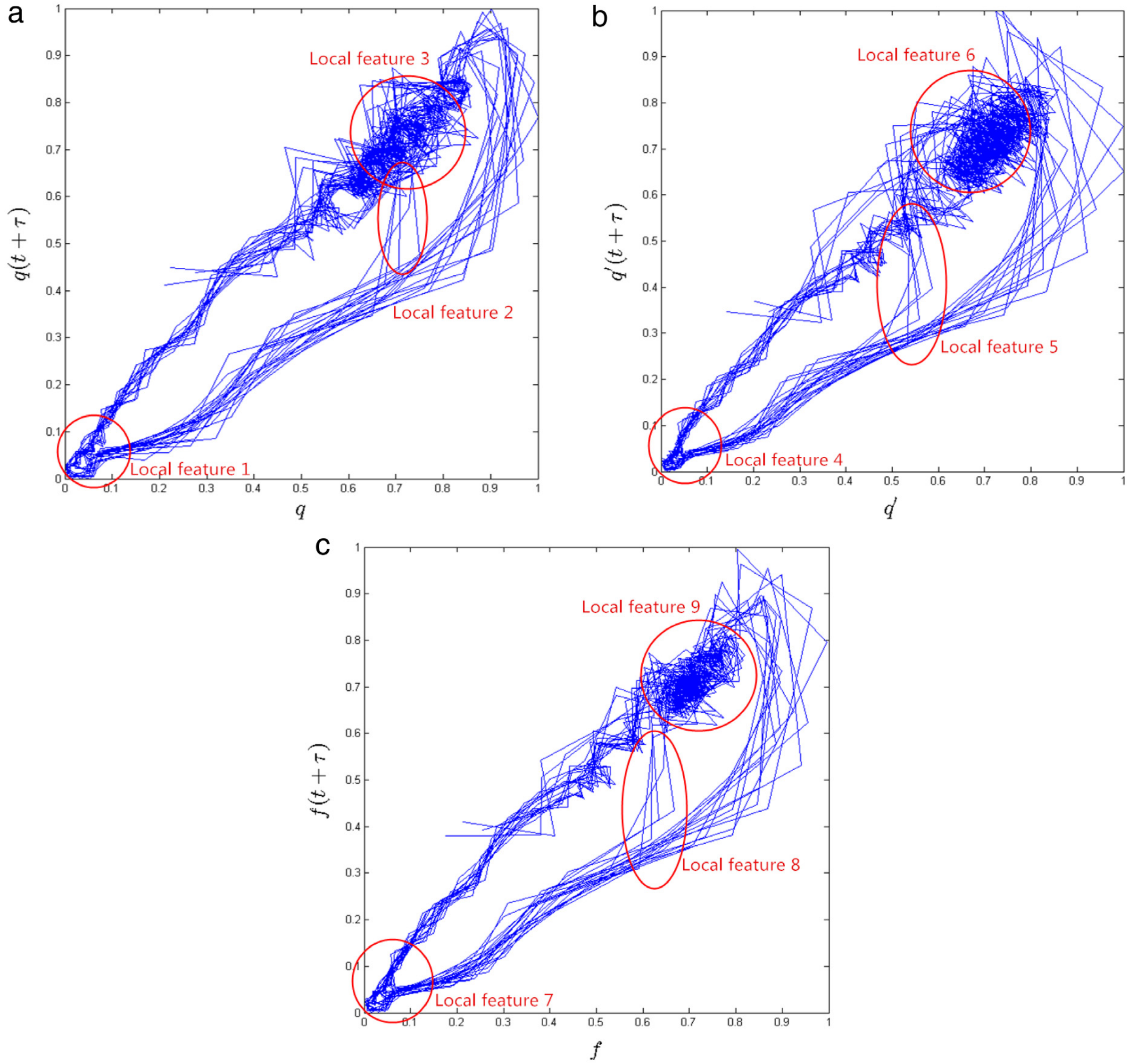


Fig. 8. Time series phase space reconstruction diagram: (a) the measured average traffic flow; (b) the theoretical traffic flow; (c) the fusion of the multi-source time series.

based on the maximum Lyapunov exponent. Then, as the multiple measures can provide more rich information than a single measure, the multiple traffic measures in low dimensional space are transformed into a high dimensional phase space using the phase space reconstruction theory and consequently the optimal fusion phase point of traffic measures is obtained based on the Bayesian estimation theory. Finally, the phase space based SVR prediction model is proposed to predict the traffic flow.

Compared with the single traffic measure based prediction methods, this proposed multiple sources and multiple measures based prediction framework can capture rich information and the main characteristics of traffic flow accurately in a high dimension phase space. Simulation results demonstrate that the phase space based SVR prediction model has relatively strong fitting capability and high predictive accuracy, and thereby can be used in the state evolution prediction of traffic flow.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China under Grant 61304197 and 71301095, by the Scientific and Technological Talents of Chongqing under Grant cstc2014kjrc-qnrc30002, by the National Key Research and Development Program under Grant 2016YFB0100906, Wenfeng Talents of CQUPT under Grant 2014-277, the Natural

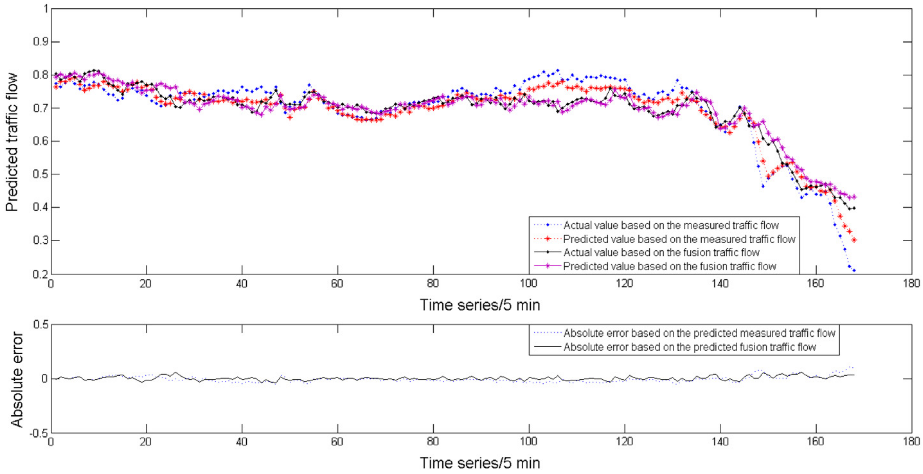


Fig. 9. Prediction of measured traffic flow and fusion traffic flow using SVR model.

Science Funds of Chongqing under Grant No. cstc2014jcyjA60003, the Doctoral Start-up Funds of CQUPT under Grant No. A2012-26. The authors make equal contribution to this paper and are listed in alphabetical order.

References

[1] T. Nagatani, Modified KdV equation for jamming transition in the continuum models of traffic, Physica A 261 (3–4) (1998) 599–607.
[2] T. Tang, W. Shi, S. Yang, H. Shang, A car-following model accounting for the driver's attribution, Physica A 413 (2014) 583–591.
[3] T. Tang, W. Shi, H. Shang, Y. Wang, An extended car-following model with consideration of the reliability of inter-vehicle communication, Measurement 58 (2014) 286–293.
[4] T. Tang, W. Shi, H. Shang, Y. Wang, A new car-following model with consideration of inter-vehicle communication, Nonlinear Dynam. 76 (2014) 2017–2023.
[5] Y. Li, Y. Kang, B. Yang, S. Peeta, L. Zhang, T. Zheng, Y. Li, A sliding mode controller for vehicular traffic flow, Physica A 462 (2016) 38–47.
[6] Y. Li, L. Zhang, S. Peeta, X. He, T. Zheng, Y. Li, A car-following model considering the effect of electronic throttle opening angle under connected environment, Nonlinear Dynam. 85 (4) (2016) 2115–2125.
[7] Y. Li, L. Zhang, B. Zhang, T. Zheng, H. Feng, Y. Li, Non-lane-discipline-based car-following model considering the effect of visual angle, Nonlinear Dynam. 85 (3) (2016) 1901–1912.
[8] Y. Li, L. Zhang, S. Peeta, H. Pan, T. Zheng, Y. Li, X. He, Non-lane-discipline-based car-following model considering the effects of two-sided lateral gaps, Nonlinear Dynam. 80 (1–2) (2015) 227–238.
[9] H. Zhang, X. Sun, Y. He, Analysis and prediction of complex dynamical characteristics of short-term traffic flow, Acta Phys. Sin. 63 (4) (2014) 040505.
[10] H. Kantz, A robust method to estimate the maximal Lyapunov exponent of a time series, Phys. Lett. A 185 (1) (1994) 77–87.
[11] A. Wolf, J.B. Swift, H.L. Swinney, J.A. Vastano, Determining Lyapunov exponents from a time series, Physica D 16 (3) (1985) 285–317.
[12] G. Barna, I. Tsuda, A new method for computing Lyapunov exponents, Phys. Lett. A 175 (6) (1993) 421–427.
[13] K. Briggs, An improved method for estimating Lyapunov exponents of chaotic time series, Phys. Lett. A 151 (1–2) (1990) 27–32.
[14] M.T. Rosenstein, J.J. Collins, C.J. De Luca, A practical method for calculating largest Lyapunov exponents from small data sets, Physica D 65 (1) (1993) 117–134.
[15] H. Liu, Y. Zhao, Z. Dai, X. Gong, Z. Yu, Calculation of the largest Lyapunov exponent in the discrete dynamical system with wavelet analysis, Acta Phys. Sin. 50 (12) (2001) 2311–2317.
[16] Q. Meng, Y. Peng, A new local linear prediction model for chaotic time series, Phys. Lett. A 370 (5) (2007) 465–470.
[17] J.B. Elsner, A.A. Tsonis, Nonlinear prediction, chaos, and noise, Bull. Ame. Meteorol. Soc. 73 (1) (1992) 49–60.
[18] J. Zhu, J. Cao, Y. Zhu, Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections, Transp. Res. C 47 (2014) 139–154.
[19] B.L. Smith, B.M. Williams, R.K. Oswald, Comparison of parametric and nonparametric models for traffic flow forecasting, Transp. Res. C 10 (4) (2002) 303–321.
[20] K.S. Lorek, G.L. Willinger, A multivariate time-series prediction model for cash-flow data, Account. Rev. (1996) 81–102.
[21] L.L. Ojeda, A.Y. Kibangou, W.C.C. De, Adaptive Kalman filtering for multi-step ahead traffic flow prediction, in: American Control Conference, 2013, pp. 4724–4729.
[22] J. Gao, Z. Leng, Y. Qin, Z. Ma, X. Liu, Short-term traffic flow forecasting model based on wavelet neural network, in: 25th Chinese Control and Decision Conference, 2013, pp. 5081–5084.
[23] J. Gu, S. Chen, Nonlinear analysis on traffic flow based on catastrophe and chaos theory, Discrete Dyn. Nat. Soc. (2014) 1–11.
[24] J. Bao, W. Chen, Z. Xiang, Prediction of traffic flow based on cellular automaton, in: International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration, 2015, pp. 88–92.
[25] W. Zheng, D. Lee, Q. Shi, Short-term freeway traffic flow prediction: Bayesian combined neural network approach, J. Transp. Eng. 132 (2) (2006) 114–121.
[26] S. Clark, Traffic prediction using multivariate nonparametric regression, J. Transp. Eng. 129 (2) (2003) 161–168.
[27] Q. Xu, R. Yang, Traffic flow prediction using support vector machine based method, J. Highway Transp. Res. Dev. 22 (12) (2005) 131–134.
[28] J. Hu, P. Gao, Y. Yao, X. Xie, Traffic flow forecasting with particle swarm optimization and support vector regression, in: 2014 IEEE 17th International Conference on Intelligent Transportation Systems, 2014, pp. 2267–2268.
[29] J.Y. Ahn, E. Ko, E.Y. Kim, Predicting spatiotemporal traffic flow based on support vector regression and Bayesian classifier, in: 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, 2015, pp. 125–130.
[30] J. Liu, Short-term traffic flow prediction model of phase space reconstruction and support vector regression with combination optimization, Comput. Eng. Appl. 50 (3) (2014) 13–17.
[31] M. Li, W. Hong, H. Kang, Urban traffic flow forecasting using Gauss–SVR with cat mapping, cloud model and PSO hybrid algorithm, Neurocomputing 99 (2013) 230–240.

[32] D. Wei, H. Liu, An adaptive-margin support vector regression for short-term traffic flow forecast, J. Intell. Transp. Syst. 17 (4) (2013) 317–327.
[33] Y. Li, X. Jiang, H. Zhu, X. He, S. Peeta, T. Zheng, Y. Li, Multiple measures-based chaotic time series for traffic flow prediction based on Bayesian theory, Nonlinear Dynam. 85 (1) (2016) 179–194.
[34] R. Cong, S.L. Liu, R. Ma, An approach to phase space reconstruction from multivariate data based on data fusion, Acta Phys. Sin. 57 (12) (2008) 7487–7493.
[35] Z. Lu, Z. Cai, K. Jiang, Determination of embedding parameters for phase space reconstruction based on improved C–C method, J. Syst. Simul. 19 (11) (2007) 2527–2529.
[36] F. Takens, Detecting Strange Attractors in Turbulence, Springer, Berlin Heidelberg, 1981, pp. 366–381.
[37] C. Wu, J. Ho, D. Lee, Travel-time prediction with support vector regression, IEEE Trans. Intell. Transp. Syst. 5 (4) (2004) 276–281.
[38] K. Chen, Forecasting systems reliability based on support vector regression with genetic algorithms, Reliab. Eng. Syst. Safety 92 (4) (2007) 423–432.
[39] http://pems.dot.ca.gov/.
[40] A. Stathopoulos, M.G. Karlaftis, A multivariate state space approach for urban traffic flow modeling and prediction, Transp. Res. C 11 (2) (2003) 121–135.