TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2016 Volume 31(3): 343–358
doi:10.3846/16484142.2016.1212734

SHORT-TERM TRAFFIC FLOW PREDICTION USING A METHODOLOGY BASED ON AUTOREGRESSIVE INTEGRATED MOVING AVERAGE AND GENETIC PROGRAMMING

Chengcheng Xu[1,2], Zhibin Li[1,2], Wei Wang[1,2]

[1]*Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing, China*
[2]*Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Southeast University, Nanjing, China*

Submitted 9 May 2014; resubmitted 11 November 2014; accepted 3 March 2015

Abstract. The accurate short-term traffic flow forecasting is fundamental to both theoretical and empirical aspects of intelligent transportation systems deployment. This study aimed to develop a simple and effective hybrid model for forecasting traffic volume that combines the AutoRegressive Integrated Moving Average (ARIMA) and the Genetic Programming (GP) models. By combining different models, different aspects of the underlying patterns of traffic flow could be captured. The ARIMA model was used to model the linear component of the traffic flow time series. Then the GP model was applied to capture the nonlinear component by modelling the residuals from the ARIMA model. The hybrid models were fitted for four different time-aggregations: 5, 10, 15, and 20 min. The validations of the proposed hybrid methodology were performed by using traffic data under both typical and atypical conditions from multiple locations on the I-880N freeway in the United States. The results indicated that the hybrid models had better predictive performance than utilizing only ARIMA model for different aggregation time intervals under typical conditions. The Mean Relative Error (MRE) of the hybrid models was found to be from 4.1 to 6.9% for different aggregation time intervals under typical conditions. The predictive performance of the hybrid method was improved with an increase in the aggregation time interval. In addition, the validation results showed that the predictive performance of the hybrid model was also better than that of the ARIMA model under atypical conditions.

Keywords: short-term traffic-forecasting; hybrid model; ARIMA; genetic programming.

Introduction

The development of the dynamic freeway traffic management systems has prompted the research for proactive traffic management strategies to mitigate traffic congestion on freeways. Toward this goal, a large amount of studies have applied an extensive variety of time-series models to produce short-term traffic variables forecasting, such as traffic volume, traffic speed, travel time, etc. (Hamed *et al.* 1995; Vlahogianni *et al.* 2005; Ghosh *et al.* 2005, 2007; Chandra, Al-Deek 2009; Chen *et al.* 2012; Hamad *et al.* 2009; Wang, Shi 2013). The short-term traffic-forecasting models were developed to extrapolate traffic variables into the near-term future based on the past observations of the same traffic variables measured with traffic surveillance systems (Smith *et al.* 2002; Vlahogianni *et al.* 2005, 2007; Turochy 2006; Zhang, Xie 2008; Zhang, Ye 2008; Dimitriou *et al.* 2008; Huang, Sadek 2009; Hamad *et al.* 2009; Min, Wynter 2011; Chen *et al.* 2012; Dunne, Ghosh 2012; Wei, Chen 2012; Wang, Shi 2013). One of the practical applications of the short-term traffic-forecasting models is to help travellers select their travel routes or plan their trips in advance based on real-time traffic information. It can also help to develop proactive traffic management strategies for traffic congestion prevention and mitigation.

Over the past several decades, much effort has been devoted to the development and improvement of forecasting short-term traffic variables. Of the conventional statistical methods, the AutoRegressive Integrated Moving Average (ARIMA) family of models has been extensively utilized in constructing the forecasting models (Hamed *et al.* 1995; Williams 2001; Smith *et al.* 2002; Williams, Hoel 2003; Ghosh *et al.* 2005, 2007; Chandra, Al-Deek 2009). For example, Hamed *et al.* (1995) employed ARIMA to develop a model for short-term prediction of traffic volume in urban arterials. Smith *et al.* (2002) compared the predictive performance of the

Corresponding author: Chengcheng Xu
E-mails: *xuchengcheng@sen.edu.cn; iamxcc1@163.com*

Taylor & Francis
Taylor & Francis Group

ARIMA model and the nearest neighbour technique in forecasting traffic flow on highway. The results demonstrated that the ARIMA model produced better predictive performance than the nearest neighbour technique did. Ghosh *et al.* (2007) used the Bayesian ARIMA model in developing a short-term traffic flow-forecasting model. It was found that the Bayesian model could better match the traffic behaviour of extreme peaks and rapid fluctuation. However, the major limitation of the ARIMA model is the pre-assumed linear correlation structure among the time series values. The approximation of linear models to complex real-world problems is not always adequate (Zhang 2003; Aladag *et al.* 2009). Previous studies also suggested that the linear statistical algorithm was not adequate to capture the complicated process underlying traffic (Hamed *et al.* 1995; Williams 2001; Stathopoulos, Karlaftis 2003).

In response to the limitations associated with the conventional statistical methods, a number of studies have proposed non-parametric methods and artificial intelligence models for developing short-term traffic flow forecasting models. These models include Artificial Neural Network (ANN) model (Smith, Demetsky 1997; Zhang 2000), recurrent neural networks (Van Lint *et al.* 2002), genetically optimized neural networks (Vlahogianni *et al.* 2005, 2007), Support Vector Machine (SVM) prediction model (Vanajakshi, Rilett 2004; Zhang, Xie 2008), and wavelet network model (Xie, Zhang 2006). Although these models could capture the nonlinear pattern of traffic flow and produce better predictive performance than conventional statistical methods, the major limitation associated with these models is that these models work as black boxes, which cannot be directly used to identify the relationships between input variables and output variable by a mathematical equation.

This study aimed to propose a simple and effective hybrid model for forecasting traffic volume that combines the ARIMA model with Genetic Programming (GP). Combining these two models could enhance the possibility to capture the linear and nonlinear patterns within traffic flow data and to improve the predictive performance. Previous studies also suggested that combining different models could improve the prediction accuracy over the individual model (Zhang *et al.* 2011; Wang, Shi 2013). GP is a relatively new modelling technique, which was proposed to solve the classification and regression problems. The GP model is an evolutionary computation method introduced by Koza (1992). In recent years, GP model has gained considerable attention in transportation engineering for regression (Das *et al.* 2010) and classification analyses (Xu *et al.* 2013). The GP model has two major advantages over the traditional statistical regression and artificial intelligence models. First, with GP model, there is no need to specify any pre-specified functional forms. The solutions of the GP model can be any functional forms describable by mathematics. The GP model could select the best functional form for the solution to the problem based on the features presented from the data. Second, in contrast to the 'black box' solutions in artificial intelligence models, the solution of the GP model is an easily readable mathematical model, which defines the tangible relationship between input variables and output variable. This allows the results of GP models to be easily applied in practical engineering applications. In addition, previous studies also suggested that the GP model could produce better predictive performance over the traditional methods (Ong *et al.* 2005; Lensberg *et al.* 2006; Etemadi *et al.* 2009; Lee, Tong 2011). So far, no applications of the GP model for short-term traffic flow forecasting have been identified by the authors.

1. Methodology

The basic principles and modelling process of the ARIMA and GP models are summarized in the following as the foundation to describe the hybrid model.

1.1. The ARIMA Model

The ARIMA model was introduced by Box and Jenkins (1976). The Auto Regressive Moving Average (ARMA) has been widely used in forecasting time series. In an ARMA(p, q) model, the value of the time series in the next period is assumed to be a linear function of several past observations and random errors, as represented in the following:

$$y(t) = \theta_0 + \sum_{i=1}^{p} \phi_i y(t-i) + \varepsilon(t) - \sum_{j=1}^{q} \theta_j \varepsilon(t-j), \quad (1)$$

where: $y(t)$ and $\varepsilon(t)$ denote the actual value and random error at time period t, respectively; ϕ_i (i = 1, 2, …, p) and θ_j (j = 0, 1, 2, …, q) are the parameters of the model; p and q are integers and referred to as the orders of the autoregressive terms and moving average terms; ε_t are assumed to be white Gaussian noise.

After calibrating the model parameters ϕ_i and θ_j using specific sampled data, the one-step forecast of $y(t)$ can be estimated as:

$$\hat{y}(t) = E[y(t)] = E\left[\theta_0 + \sum_{i=1}^{p} \phi_i y(t-i) + \varepsilon(t) - \sum_{j=1}^{q} \theta_j \varepsilon(t-j)\right] =$$

$$E\left[\theta_0 + \sum_{i=1}^{p} \phi_i y(t-i) + \varepsilon(t) - \sum_{j=1}^{q} \theta_j \left(y(t-i) - \hat{y}(t-i)\right)\right] =$$

$$E\left[\theta_0 + \sum_{i=1}^{p} \phi_i y(t-i) - \sum_{j=1}^{q} \theta_i \left[y(t-i) - \hat{y}(t-i)\right]\right] + E[\varepsilon(t)] =$$

$$\theta_0 + \sum_{i=1}^{p} \phi_i y(t-i) - \sum_{j=1}^{q} \theta_i \left[y(t-i) - \hat{y}(t-i)\right], \quad (2)$$

where: $E[\varepsilon(t)]$ is the expected value of white Gaussian noise, i.e., $E[\varepsilon(t)] = 0$; θ_j (j = 0, 1, 2, …, q), ϕ_i(i = 1, 2; …, p) are the estimated parameters of the ARMA model; $y(t-i)$ are the known historical traffic volume data; and $\hat{y}(t-i)$ are the predicted volume of the ARMA model.

The ARIMA model is a generalization of the ARMA model. In an ARIMA(p, d, q) model, the parameter p and q are the same to those in the ARMA model. The parameter d represents the d-th order difference of the original data series, which aims to remove the trend

from the data series. By introducing the backshift operator B (that is, $By(t)=y(t-1)$), the Eq. (1) for ARMA(p, q) can be written as:

$$y(t)-\phi_1 y(t-1)-\phi_2 y(t-2)-...-\phi_p y(t-p)=$$
$$\theta_0+\varepsilon(t)-\left(\theta_1\varepsilon(t-1)+\theta_2\varepsilon(t-2)+\cdots+\theta_q\varepsilon(t-q)\right)\Rightarrow$$
$$\phi(B)y(t)=\theta_0+\theta(B)\varepsilon(t), \quad (3)$$

where: $\phi(B)$ is the autoregressive operator which is represented as a polynomial in the backshift operator: $\phi(B)=1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p$; $\theta(B)$ is the moving-average operator, represented as a polynomial in the backshift operator: $\theta(B)=1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q$. Similarly, the ARIMA model can be written as:

$$\phi'(B)(1-B)^d y(t)=\theta_0'+\theta'(B)\varepsilon'(t), \quad (4)$$

where: d represents the degree of non-seasonal differencing; $\phi'(B)$ and $\theta'(B)$ are the autoregressive and moving-average operators for the ARIMA model, respectively. In the ARIMA model, the d-th order difference of the data series, $(1-B)^d y(t)$, is used for forecasting, instead of the original data series $y(t)$.

1.2. The GP Technique

The GP model is an evolutionary computation method introduced by Koza (1992). The GP model can be used to generate mathematical models, which represent approximate or exact solutions to a problem (Koza 1992). It can be considered as an extension of the genetic algorithms (GA). The main difference between GP and GA is the representation of individuals. The individuals in a GA model are numbers coded as fixed-length binary strings, while the individuals in a GP model are mathematical models coded as function trees (Koza 1992; Xu *et al.* 2013). An example of function tree in GP model is given in Fig. 1. The inner nodes represent the mathematical functions such as '+' and '÷', and the leaf nodes represent the predictors and constants. The left most tree in Fig. 1 represents the mathematical model $f(X,Y,Z)=(X+Y)\cdot\left[Z\div\log(0.2)\right]$. In a particular problem, the list of functions and predictors should be specified. The mathematical models in GP are generated from the pre-specified set of functions and predictors.

In general, GP works on a population of mathematical models (individuals) based on evolution theory. In each generation, multiple models are stochastically selected based on their fitness, and modified to form a new population of models by genetic operations. The new population of models is then used in the next iteration of the algorithm. A GP model will stop when the predetermined maximum number of generations has been produced or the predetermined fitness level has been reached for the population. The evolution process is expected to produce continuously a better model for a problem.

The new models in a GP model are usually created by three genetic operators, including crossover, mutation, and reproduction. The reproduction operator simply selects a proportion of models and includes them into the next generation without any alterations. The creation of new or offspring models from the crossover operation is accomplished by combining information extracted from the selected parents. Two parent models are randomly selected based on their fitness level and sub-trees are chosen from both parent models. Then the crossover operator swaps the sub-trees from the two parent models. Fig. 1 illustrates an example of crossover operation.

The purpose of mutation operator is to introduce new information into the population and avoid the premature convergence of a GP model. In mutation operation, a single parent is randomly selected based on its fitness level. A random sub-tree on the parent model is selected and replaced with a new random tree created from the pre-specified set of predictors and functions (Fig. 2). In the procedure of generating a random tree, the node at the initial tree depth level is first randomly chosen from the set of functions. Then its children node(s) are randomly chosen between functions set and predictors set. The random tree will stop growing when reaching the maximum tree level. Readers may consult Koza (1992) for full description of this procedure.

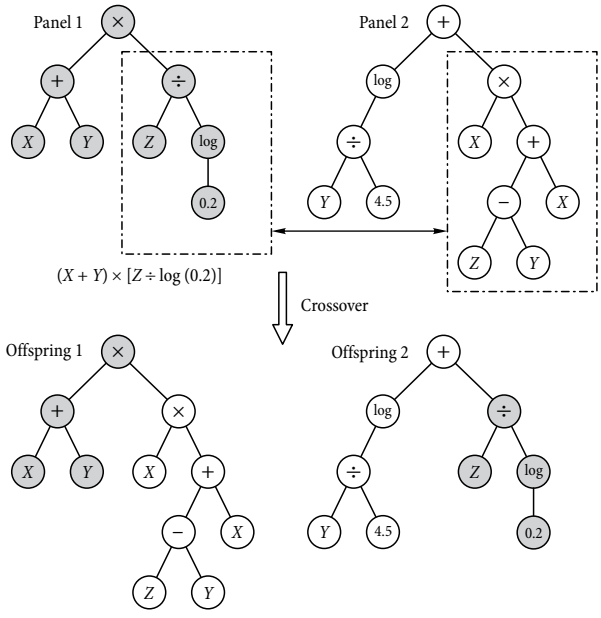

Fig. 1. Crossover operation in GP

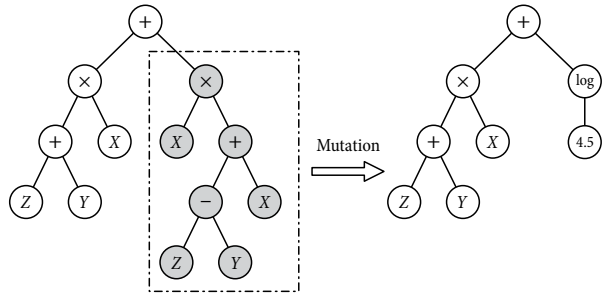

Fig. 2. Mutation operation in GP

The fitness function of a GP model determines how well a model in the population is able to solve the problem. The fitness function varies greatly across different types of problems. The fitness function is usually developed based on the error between the values predicted by the model and the actual data. In this study, a fitness function for short-term traffic flow forecasting was developed based on the Mean Absolute Error (MAE). Assuming a dataset $S=\{(y_1,x_1),(y_2,x_2),\ldots,(y_a,x_m)\}$ of input variables x_i for output variable y_i, the functional form of the fitness function is expressed as follows:

$$F(B_j)=\frac{1}{m}\sum_{i=1}^{m}\left|B_j(x_i)-y_i\right|, \tag{5}$$

where: $F(B_j)$ denotes the fitness of the j-th model B_j in the population; $B_j(\boldsymbol{x_i})$ is the value calculated by the j-th model B_j in the population.

The GP model uses the following steps to solve problems:

(a) initialization – create at random an initial population of M models;
(b) execute each model in the current population on training dataset and evaluate the fitness of each model in the current population;
(c) select the parent models, which will be used to produce offspring models;
(d) select the reproduction, crossover, and mutation operators probabilistically;
(e) generate a new model by performing one of the three genetic operators;
(f) repeat step (c) to step (e) until the predetermined population size M has been reached;
(g) replace the M old models by new generated M models;
(h) repeat step (b) to step (g) until the predetermined maximum generation N has been reached.

The model with the best fitness level in any generation is designated as the result of GPs.

1.3. The Hybrid Methodology Based on ARIMA and GP

Since it is difficult to completely know the characteristics of the traffic volume time series data, hybrid methodology that has both linear and nonlinear modelling capabilities can be a good strategy. By combining different models, different aspects of the underlying patterns of traffic flow may be captured. This study proposed a hybrid model that combines ARIMA for modelling the linear component L_t of traffic flow time series and the GP for modelling the nonlinear component N_t, as follows:

$$y(t)=L_t+N_t+\xi_t, \tag{6}$$

where: $y(t)$ represents the actual value at time period t; L_t and N_t denote the linear component and nonlinear component of the model respectively; ξ_t denotes the random error term. The residuals from the ARIMA model (r_t) were calculated as follows:

$$r_t=y(t)-\hat{L}_t, \tag{7}$$

where: $\hat{L}_t$ is the predicted value of L_t, which is estimated using the ARIMA model. By modelling the residuals from the ARIMA(r_t) using the GP model, nonlinear relationships can be discovered. With n input variables, the GP model for the residuals r_t can be written as:

$$r_t=f(r_{t-1},r_{t-2},\ldots,r_{t-n})+\xi_{rt}, \tag{8}$$

where: ξ_{rt} denotes the random error term; $f(r_{t-1},r_{t-2},\ldots,r_{t-n})$ represents the nonlinear function constructed using the GP model. Using the GP model to construct the nonlinear component of time series can generate a mathematical equation than ANN and SVM model. Thus, in practice, the predicted values using GP can be verified through the mathematical equation. The estimation of the residuals r_t can be determined by Eq. (8). Then the predicted values of the time series are estimated as follows:

$$\hat{y}(t)=\hat{L}_t+\hat{N}_t. \tag{9}$$

The proposed hybrid approach uses the following steps to forecast traffic flow:

1) Model the linear component of the time series using ARIMA model, and estimate $\hat{L}_t$ using ARIMA model.
2) Calculate the residuals from the ARIMA model using Eq. (7), and model them using the GP model in Eq. (8). The nonlinear component $\hat{N}_t$ are the predicted values of the developed GP model in Eq. (8).
3) Estimate the forecasts of the hybrid model by adding the predicted values of the ARIMA and GP model.

2. Data Sources and Evaluation Criteria

Data were obtained from the highway Performance Measurement System (PeMS) maintained by the California Department of Transportation (Caltrans), US. The PeMS database provided 30-sec raw loop detector data, including vehicle count, vehicle speed, and detector occupancy. The traffic data were collected from the Detector 401561 (Site A) and Detector 401517 (Site B) located on the northbound freeway I-880 (Fig. 3). The freeway has five lanes at the selected sites. The 30-sec raw traffic data were collected from all the five lanes. As shown in Fig. 3, the selected two detectors are far away from each other and have a number of ramps in between. Thus, the traffic data collected at the two sites are considered to have low correlations. The PeMS database also provides the detailed traffic incident data, including incident type, starting time, location and duration.

As discussed in Stathopoulos and Karlaftis (2003), Dunne and Ghosh (2012), and Chen *et al.* (2012), the traffic flow series recorded on weekdays were substantially different from those recorded on the weekends or holidays. The prediction models for weekday might produce unsatisfactory results for traffic data on weekends. Thus, for consistency purposes, this study only focuses on the weekday traffic flows.

Fig. 3. Study sites of freeway I-880N

The missing data problems are unavoidable in traffic flow data. Previous studies suggested that the missing data problem greatly affected traffic analysis (Zhong *et al.* 2004; Xin *et al.* 2006; Qu *et al.* 2009; Chen *et al.* 2003, 2012). The missing data should be imputed before developing the traffic-forecasting model. Different statistical methods and artificial intelligence models have been used for missing data imputation, such as, the Bayesian networks (Chen *et al.* 2003), the Bayesian principal component analysis (Qu *et al.* 2009), the ANN (Zhong *et al.* 2004), and the Probabilistic Principal Component Analysis (PPCA) (Qu *et al.* 2009). Since the PPCA can quickly produce accurate imputations (Qu *et al.* 2009), the PPCA was used in this study to impute the missing values in the dataset. The PPCA also has the advantage of appropriate combing both neighbouring historical flow data and current-day flow data (Qu *et al.* 2009). The reader may consult Oba *et al.* (2003) and Qu *et al.* (2009) for full description of the PPCA method.

The measurement noises and useless traffic fluctuations in the high-resolution traffic data (lower than 1 min) can decrease the predictive performance of the prediction models (Castro-Neto *et al.* 2009; Chen *et al.* 2012). Accordingly, the 30-sec raw detector data was first aggregated into 5-min traffic data by summing up the 10 observations of the 30-sec traffic volumes:

$$y = \sum_{i=1}^{n} q_i, \tag{10}$$

where: y denote the aggregated traffic volume; q_i represent the average traffic volume across different lanes; n represent the number of observations during the aggregation time interval. If there are any missing values of the 30-sec traffic volume during a 5-min interval, the traffic volume for this 5-min interval was labelled as a missing value. The PPCA method was conducted on the 5-min traffic data to impute all the missing values within it. The imputed 5-min traffic data were further aggregated into 10-min, 15-min and 20-min time interval using Eq. (10). The proposed hybrid models were fitted for these four different time-aggregations: 5, 10, 15, and 20 min.

Previous study suggested that the traffic flow prediction model developed by normal traffic data may produce poor predictive performance when incidents or atypical situations are present (Castro-Neto *et al.* 2009; Guo *et al.* 2013). Hence, the predictive performance of the proposed hybrid model was evaluated with traffic data under both normal conditions (Scenario 1) and incident conditions (Scenario 2). In Scenario 1, the used traffic flow data were not significantly affected by incidents, such as crashes. The traffic flow data at Sites A and B were collected from 1 May 2012 to 1 June 2012. To achieve more reliable and accurate estimations, a long period of traffic flows were selected as training dataset (Zhang *et al.* 2011). The traffic flow data from the weekdays in May 2012 were used as the training dataset and the traffic flow data on 1 June 2012 were used as the validation dataset for Scenario 1. Table 1 summarizes the descriptive statistics of the training and validation dataset for Scenario 1 based on the 30-sec traffic data.

In Scenario 2, the traffic data under incident conditions were collected to test the predictive performance of the proposed hybrid model under incident conditions. The only difference between Scenarios 1 and 2 was that the validation dataset for Scenario 2 contained the traffic flow data under incident conditions. The predictive performance of the models developed based on the training dataset in Scenario 1 was tested on the validation dataset for Scenario 2. Table 2 summarizes the descriptive statistics and characteristics of the traffic data under incident conditions in Scenario 2.

To compare the predictive performance of the ARIMA and the proposed hybrid model, the following four performance indexes were used:

1) the Mean Absolute Error (MAE):

$$MAE = \frac{1}{m}\sum_{t=1}^{m}\left|y(t) - \hat{y}(t)\right|; \tag{11}$$

2) the Mean Relative Error (MRE):

$$MRE = \frac{1}{m}\left(\sum_{t=1}^{m}\frac{\left|y(t) - \hat{y}(t)\right|}{y(t)}\right)\cdot 100\%; \tag{12}$$

Table 1. Descriptive Statistics of traffic data used in Scenario 1

Selected sites	Training dataset [vehicles/hour]			Validation dataset [vehicles/hour]		
	Mean	S.D.	Missing [%]	Mean	S.D.	Missing [%]
Site A	818	465	3.60	840	468	0
Site B	924	545	3.54	904	505	0

Table 2. Descriptive statistics of traffic data used in Scenario 2

Selected sites	Date	Time	Mean	S.D.	Missing [%]
Site A	8 June 2012	12:30÷13:25	793	174	0%
Site B	4 April 2012	11:40÷12:45	701	194	0%

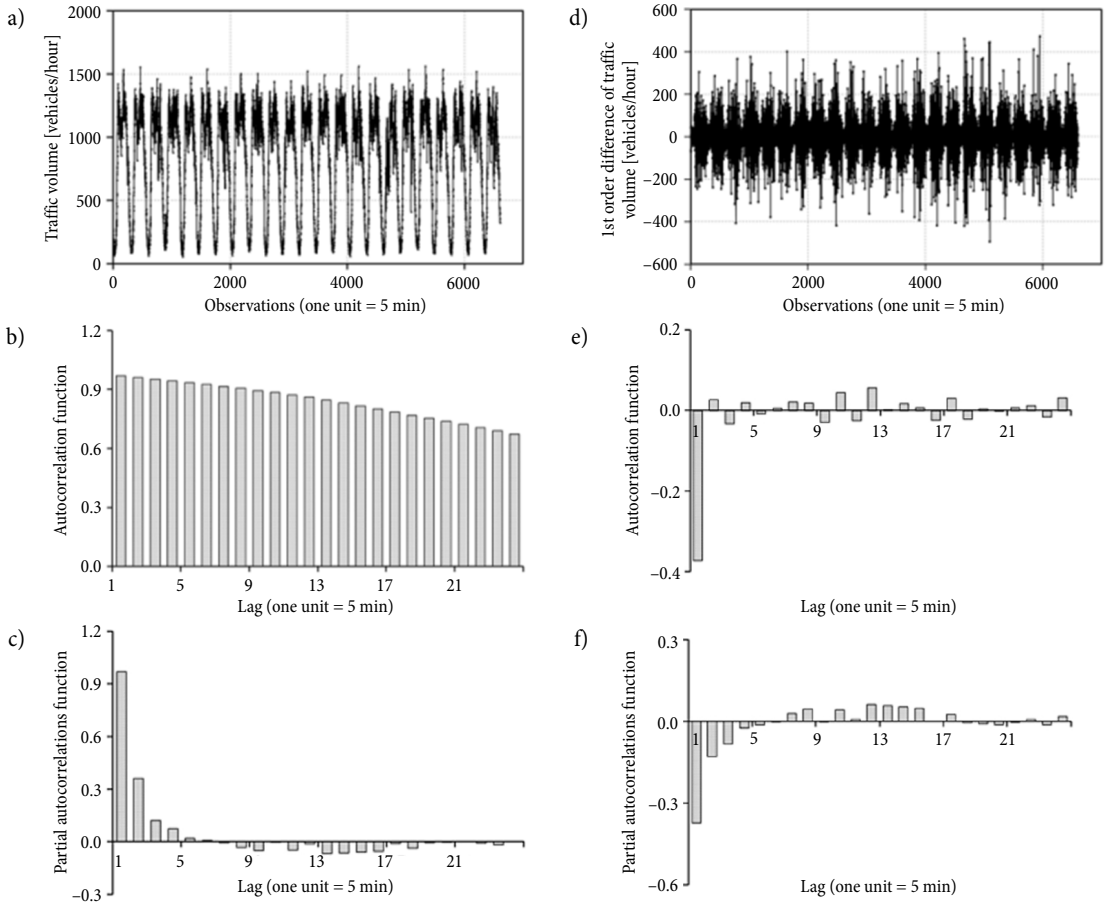

Fig. 4. 5-min average traffic flow, ACF and PACF at Site A

3) the Mean Square Error (MSE):

$$MSE = \frac{1}{m}\left(\sum_{t=1}^{m}\left(y(t)-\hat{y}(t)\right)^2\right); \tag{13}$$

4) the Mean Square Relative Error (MSRE):

$$MSRE = \frac{1}{m}\left(\sum_{t=1}^{m}\left(\frac{y(t)-\hat{y}(t)}{y(t)}\right)^2\right)\cdot 100\%. \tag{14}$$

3. Results of Data Analysis

3.1. Model Development

A statistical analysis of a time series requires that the time series are stationary. In other words, this time series should have the same statistical behaviour at each point in time. Forecast of statistical models, including the ARIMA model, based on non-stationary series usually exhibit large errors (Washington *et al.* 2003). Readers may consult Washington *et al.* (2003) for full explanation of the requirement of stationarity in the time series analysis. Thus, before modelling a time series, the data must be stationary. Fig. 4a illustrates the 5-min traffic data of the whole training dataset at Site A. Fig. 4b and 4c illustrate the AutoCorrelation Function (ACF) and the Partial AutoCorrelation Function (PACF) of the 5-min traffic data, respectively. The ACF plot indicates that the traffic volume series is non-stationary, since the ACF decays very slowly.

The 5-min traffic volume series become stationary after the first-order differencing. The first-order difference of 5-min traffic volume does not have a visible trend and its ACF and PACF decay quickly (Fig. 4d–f). The Augmented Dickey Fuller (ADF) test was further conducted to test the stationarity. The ADF test result indicates that the null hypothesis of non-stationarity can be rejected at the 0.01 significance level after the first differencing was performed. Thus, the first-order difference of 5-min traffic volume is stationary and can be used for the ARIMA model development.

To identify the best ARIMA model for the 5-min traffic data at Site A, the ARIMA models were developed for different combinations of parameter p and q. The parameter p and q were set from 0 to 10. The Akaike's Information Criterion (AIC) was used to find the best ARIMA model. It was found that the AIC reached a minimum when p and q were set to be 3 and 2, respectively. Besides, it was ensured that all the variables in the ARIMA model were statistically significant (Table 3). The residuals analysis was further conducted for the developed ARIMA model to make sure there is no pattern remaining. Fig. 5 illustrates the graphical check of the residuals from the developed ARIMA model for the 5-min traffic data at Site A. As shown in Fig. 5a, 5b, the autocorrelations of the residuals from the ARIMA model are very small and insignificant. The partial autocorrelations (Fig. 5c) and inverse autocorrelations (Fig. 5d) of the residuals are also negligible. The white noise test

was also conducted on the residuals. The results of the white noise test in Table 4 indicate that the residuals from ARIMA model have no pattern remaining, and that the best ARIMA model for the 5-min traffic data at Site A has been identified. The other 7 ARIMA models for different time-aggregations were developed using the same procedure. Tables 3 and 5 summarize the estimation results of the ARIMA models at Sites A and B for different aggregation time intervals, including 5, 10, 15 and 20 min.

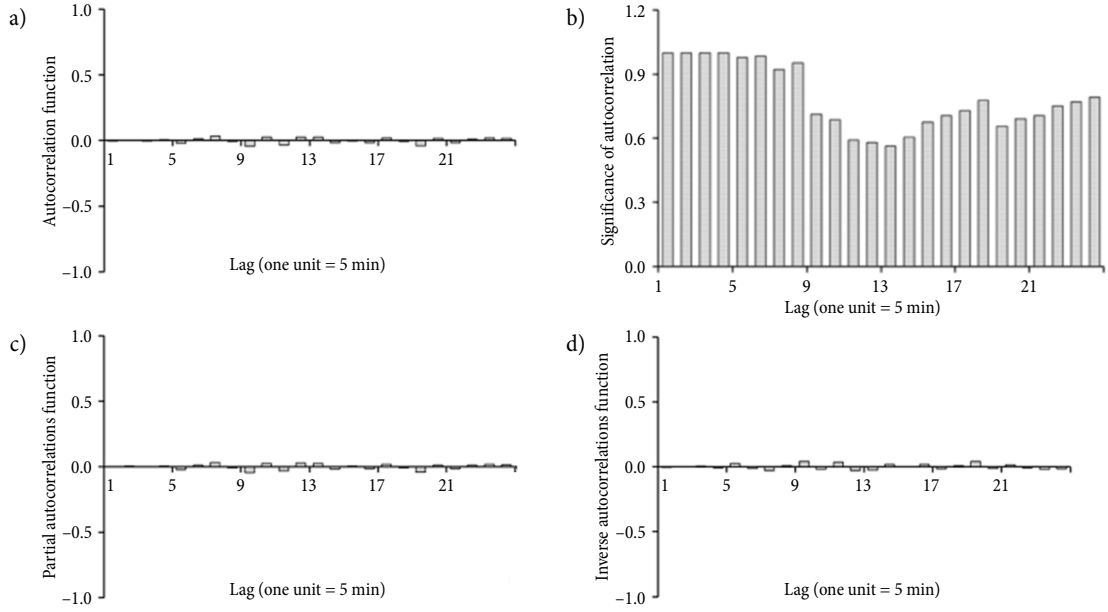

Fig. 5. The graphical check of residuals from the developed ARIMA model for 5-min traffic data at Site A

Table 3. Estimation results of the ARIMA models for different aggregation intervals at Site A

Aggregation level	Parameter		Estimate	Standard error	*t*-value	Significance
5-min ARIMA (3, 1, 2)	Constant		$-1.061 \cdot 10^{-4}$	0.008	–0.010	0.9897
	Auto regressive	Lag 1	1.232	0.036	34.500	<0.0001
		Lag 2	–0.158	0.026	–6.110	<0.0001
		Lag 3	–0.128	0.022	–5.940	<0.0001
	Moving average	Lag 1	1.675	0.032	51.620	<0.0001
		Lag 2	–0.720	0.033	–22.030	<0.0001
10-min ARIMA (4, 1, 4)	Constant		$-3.269 \cdot 10^{-4}$	0.017	–0.020	0.9845
	Auto regressive	Lag 1	2.065	0.019	111.190	<0.0001
		Lag 2	–2.237	0.035	–64.470	<0.0001
		Lag 3	1.237	0.035	35.520	<0.0001
		Lag 4	–0.172	0.018	–9.480	<0.0001
	Moving average	Lag 1	2.256	0.006	348.580	<0.0001
		Lag 2	–2.647	0.007	–399.970	<0.0001
		Lag 3	1.648	0.006	296.760	<0.0001
		Lag 4	–0.394	0.005	–83.420	<0.0001
15-min ARIMA (3, 1, 4)	Constant		$-3.112 \cdot 10^{-4}$	0.025	–0.010	0.990
	Auto regressive	Lag 1	0.745	0.066	11.210	<0.0001
		Lag 2	–0.857	0.055	–15.530	<0.0001
		Lag 3	0.698	0.060	11.570	<0.0001
	Moving average	Lag 1	0.796	0.068	11.760	<0.0001
		Lag 2	–0.964	0.061	–15.860	<0.0001
		Lag 3	0.714	0.065	10.980	<0.0001
		Lag 4	–0.163	0.024	–6.820	<0.0001

End of Table 3

Aggregation level	Parameter		Estimate	Standard error	*t*-value	Significance
20-min ARIMA (4, 1, 3)	Constant		–0.001	0.035	–0.020	0.984
	Auto regressive	Lag 1	–0.338	0.088	–3.830	0.0001
		Lag 2	–0.250	0.065	–3.870	0.0001
		Lag 3	0.749	0.065	11.550	<0.0001
		Lag 4	0.087	0.034	2.590	0.010
	Moving average	Lag 1	–0.428	0.086	–4.950	<0.0001
		Lag 2	–0.422	0.086	–4.920	<0.0001
		Lag 3	0.562	0.086	6.560	<0.0001

Table 4. Autocorrelation check for white noise for the 5-min ARIMA model at Site A

To lag	*Pr* > *Chisq*	Autocorrelations					
6	1.000	0.000	0.000	–0.001	0.004	–0.021	0.011
12	0.108	0.030	–0.009	–0.045	0.026	–0.033	0.025
18	0.350	0.026	–0.017	–0.002	–0.017	0.018	–0.010
24	0.440	–0.040	0.015	–0.019	0.009	0.018	0.016
30	0.584	0.001	–0.006	–0.007	–0.018	–0.008	0.041

Table 5. Estimation results of the ARIMA models for different aggregation intervals at Site B

Aggregation level	Parameter		Estimate	Standard error	*t*-value	Significance
5-min ARIMA (1, 1, 2)	Constant		$1.124 \cdot 10^{-4}$	0.007	0.020	0.9869
	Auto regressive	Lag 1	–0.869	0.095	–9.160	<0.0001
	Moving average	Lag 1	–0.415	0.098	–4.240	<0.0001
		Lag 2	0.372	0.049	7.570	<0.0001
10-min ARIMA (3, 1, 3)	Constant		$-8.820 \cdot 10^{-5}$	0.019	0.000	0.996
	Auto regressive	Lag 1	0.394	0.199	1.980	0.048
		Lag 2	0.769	0.244	3.150	0.002
		Lag 3	–0.444	0.090	–4.950	<0.0001
	Moving average	Lag 1	0.592	0.199	2.980	0.003
		Lag 2	0.724	0.289	2.500	0.012
		Lag 3	–0.646	0.131	–4.920	<0.0001
15-min ARIMA (3, 1, 2)	Constant		$-3.103\text{E} \cdot 10^{-4}$	0.030	–0.010	0.992
	Auto regressive	Lag 1	0.838	0.109	7.690	<0.0001
		Lag 2	–0.291	0.088	–3.290	0.001
		Lag 3	0.139	0.030	4.670	<0.0001
	Moving average	Lag 1	0.911	0.109	8.340	<0.0001
		Lag 2	–0.365	0.095	–3.830	0.000
20-min ARIMA (3, 1, 2)	Constant		0.002	0.003	0.620	0.5331
	Auto regressive	Lag 1	1.775	0.041	43.510	<0.0001
		Lag 2	–0.678	0.066	–10.310	<0.0001
		Lag 3	–0.113	0.028	–3.960	<0.0001
	Moving average	Lag 1	1.780	0.034	51.880	<0.0001
		Lag 2	–0.782	0.034	–22.770	<0.0001

The GP models were developed to predict the nonlinear component of the traffic flow time series. The parameters used in the GP models are given in Table 6. The function set contained 8 standard arithmetic operators, including +, –, ×, ÷, *protected square root*, sin, cos, and *pow*(2, *x*). If A ≤ 0, the protected square root of *A* equals to 0. When *A* > 0, the protected square root of *A* equals to the square root of *A*. The function *pow*(2, *x*) represents two raised to the power, *x*. The population size was set to 1000, and the maximum number of generations was 100. The reproduction probability was 0. The purpose of doing so was to let the crossover and

mutation operation govern the evolutionary process (Xu *et al.* 2013). The probabilities of the crossover and mutation were set to be 0.4 and 0.6, respectively. Implementing a lower crossover probability and a higher mutation probability can avoid genetic drift (Das *et al.* 2010), which is the accumulation to a sub-optimal solution in the search space. The terminal set included the constant terminals (randomly generated floating point numbers between –10 and 10) and the residual lagged variables (i.e., r_{t-1}, r_{t-2}, …, r_{t-n}).

To select an optimal number of residual lagged variables, the GP model was conducted in a successive phase in which the number of residual lagged variables n was set from 1 to 10. The number of 10 is expected to cover the possible n that ensures the best prediction accuracy. The optimal number of residual lagged variables in previous studies that use the similar hybrid model is usually lower than 10 (Zhang 2003; Aladag *et al.* 2009; Lee, Tong 2011; Zhang *et al.* 2011). The value would be selected when the prediction accuracy of the GP model reached a maximum. After the development of the GP model for the 5-min traffic data at Site A, the residuals from the hybrid model for 5-min interval was also analysed to ensure that there is no pattern left. The white noise test of the residuals from the hybrid model in Table 7 indicates that there is no pattern remaining in the residuals from the hybrid model for 5-min interval. Thus, the best GP model for 5-min interval at Site A has been identified. The other 7 GP models for different time-aggregations were developed using the same procedure. The white noise tests also indicate that there are no patterns left for these 7 hybrid models. Figs 6 and 7 illustrate the GP models for different aggregation time intervals at Sites A and B.

Table 6. The configuration parameters of the GP model

Configuration parameters	Selected values
Number of individuals	1000
Number of generations	100
Genetic operations	crossover and mutation
Crossover probability	0.4
Mutation probability	0.6
Functions set	+, –, ×, ÷, protected square root, *pow*(2, *x*), sin, cos
Terminal set	random constant (between –10 and 10), residual lagged variables
Fitness function	MAE

a) Aggregation time length = 5 min

$$
\begin{cases}
r_t = 0.176\left\{\sin\left[(f_{19}-0.625)/r_{t-1} - 2.290 - r_{t-3} - 3r_{t-6}\right] + f_{20} + r_{t-6}\right\}\\
f_{20} = f_{18}/f_{17} - \sin(f_{17}/f_{18}) - 2f_{19}\\
f_{19} = 0.288\sin\left[\sin(f_{17}/f_{18})\right]/\left[f_{18}/f_{17} - \sin(f_{17}/f_{18})\right]\\
f_{18} = 0.841\left(f_{12}/|f_{11}r_{t-1}| - f_{13}\right)\\
f_{17} = \cos\left\{\sin\left[pow(2, f_{16}) - 0.159\right]\right\}^4\\
f_{16} = \sin\left(-0.828f_{15} - 0.276 - r_{t-6} - r_{t-7}\right) + f_8 - |f_{11}r_{t-1}|\\
f_{15} = \left[1.083\left(f_{14}^2 - 1.987\right) - 0.494\right]/r_{t-1}r_{t-4} - r_{t-1} - r_{t-3}\\
f_{14} = 0.033f_{13}/\left\{\left(f_8 - |f_{11}r_{t-1}|\right)\left[1 - pow(2, f_5) + f_9 - |f_9|\right]\right\}\\
f_{13} = \left[|f_{11}r_{t-1}| + |f_8| - |f_9| + pow(2, f_5) - 1 - r_{t-2}\right]\\
f_{12} = \left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)/\left[\left|f_4\left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)\right| - 1.641\right]\\
f_{11} = \left\{pow\left[2, pow(2, f_{16}) - 1\right] - 1\right\}^2 - r_{t-6} - r_{t-7}\\
f_{10} = |f_9|\left[1 - pow(2, f_5) + f_9 - |f_9|\right]\\
f_9 = \sin\left[\left|f_4\left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)\right| - 1.641 + r_{t-2} + 2f_8\right]\\
f_8 = \sin\left[pow\left(2, f_7/r_{t-4}\right) - 4r_{t-6} - 2.329 + r_{t-2}\right] + \left|f_4\left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)\right|\\
f_7 = \left[-1.364\left(f_6 - 0.106\right)r_{t-2} - 0.494\right]/r_{t-4}\\
f_6 = pow\left\{2, 0.106\left[pow(2, f_5) - 1\right] - r_{t-1}\right\} - 1 + \left|f_4\left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)\right|\\
f_5 = \sin\left[\left|r_{t-7} + f_1^{0.5}\right|r_{t-3}^2 r_{t-6}/\left(r_{t-7} + f_1^{0.5} + r_{t-1}^2 f_1\right)\right]\\
f_4 = 2pow\left\{2, \cos\left[pow(2, f_3) - 1\right]\right\} - 2\\
f_3 = 0.794\left[0.139\left(-0.467f_2 - r_{t-7} - 1.642\right) + r_{t-1} + 1.364\right]\\
f_2 = \left[\left|\left(r_{t-7} + f_1^{0.5}\right)r_{t-3}^2 r_{t-6}\right| - r_{t-7}\right]/r_{t-2} + r_{t-6} + 1.549\\
f_1 = \cos\left[\sin\left(0.996 - r_{t-3}\right)\right]
\end{cases}
$$

b) Aggregation time length = 10 min

$$
\begin{cases}
r_t = -0.192\sin f_6 \times \left[2.336f_2 + \cos\left(2.336f_2\right) + pow(2, f_3)\right]\\
f_6 = \left[4.587pow(2, f_5) - 0.634\right]/r_{t-6} - 0.176 + r_{t-1} - r_{t-6}\\
f_5 = \left|0.006pow(2, f_3)/f_4 - 0.003f_4\right|\\
f_4 = -(\sin r_{t-2} + |\sin r_{t-2}|/\sin r_{t-2}) - pow(2, f_3)\\
f_3 = \left|\left[2\cos(2.336\cos f_2) + r_{t-2} + r_{t-5}\right] \times (\sin r_{t-2} + |\sin r_{t-2}|/\sin r_{t-2})\right|\\
f_2 = 10.989\left[pow(2, f_1) + r_{t-2} - 0.113\right] \times \left[r_{t-6}\left(2\sin|r_{t-2}|/\sin r_{t-2} + 1\right) - r_{t-2}\right]\\
f_1 = r_{t-3}\left[r_{t-6}\left(2\sin|r_{t-2}|/\sin r_{t-2} + 1\right) - r_{t-2}\right]
\end{cases}
$$

c) Aggregation time length = 15 min

$$
\begin{cases}
r_t = -0.288\sin f_6 - 0.14r_{t-1} - 0.14\left[\sin\left(1.012r_{t-6}^2\right) + r_{t-6}\right]/f_3 - 0.14f_4\\
f_6 = 4\sin f_5 - \left[\sin\left(1.012r_{t-6}^2\right) + r_{t-6}\right]/f_3 - f_4\\
f_5 = f_4 - \left[\sin\left(1.012r_{t-6}^2\right) + r_{t-6}\right]/f_3\\
f_4 = 1.048r_{t-5}\left(f_3 - 0.559\right) - r_{t-8} - 2r_{t-3} + \sin\left(1.012r_{t-6}^2\right) \times \left[\sin\left(1.012r_{t-6}^2\right) + r_{t-6}\right]\\
f_3 = 2pow(2, f_2) - 2.289 - 2r_{t-2} + r_{t-3} + r_{t-6}f_1 + pow(2, f_2)\\
f_2 = r_{t-3}r_{t-6}f_1 + \sin\left(1.012r_{t-6}^2\right) + r_{t-6}\\
f_1 = \left\{2r_{t-1}\left[\sin\left(1.012r_{t-6}^2\right) + r_{t-6}\right] - 1.258\right\}r_{t-8} + r_{t-6}
\end{cases}
$$

d) Aggregation time length = 20 min

$$
\begin{cases}
r_t = -0.56\sin f_8 - 0.28r_{t-5} + 0.28f_2 + 0.14r_{t-3} + f_2\\
f_8 = pow\left[2, pow(2, f_7) - 1 - 2r_{t-2}\right] - 1\\
f_7 = (0.056f_5^2/r_{t-4} \times f_6 r_{t-1} + r_{t-4}) \times r_{t-2}\\
f_6 = f_4 - f_5^2 + 0.056f_5^2/r_{t-4}\\
f_5 = -4\sin\left[-pow(2, f_3) + 1 - r_{t-3}\right] + 2f_4 - 1.967 + 1.5f_2 - r_{t-3} - r_{t-1}\\
f_4 = f_3 - f_2\left(|f_2| - 0.5f_2\right)\\
f_3 = f_2\left(|f_2| - 0.5f_2\right)/r_{t-4}\\
f_2 = -0.278f_1 + 0.038f_1 r_{t-3}\\
f_1 = -0.963\left[2\sin(-60.606/r_{t-5} - 30.303r_{t-2}) + r_{t-5}\right]r_{t-1} - r_{t-1}
\end{cases}
$$

Fig. 6. The GP model for different time-aggregations at Site A

Table 7. Autocorrelation check for white noise for the 5-min hybrid model at Site A

To lag	*Pr*>*Chisq*	Autocorrelations					
6	0.941	0.001	0.005	–0.004	–0.024	–0.016	0.013
12	0.151	0.059	0.015	–0.005	0.055	0.003	0.046
18	0.150	0.050	–0.010	0.009	–0.025	0.014	–0.025
24	0.182	–0.048	0.001	–0.026	0.000	0.018	0.012
30	0.288	0.005	–0.008	–0.002	–0.016	0.000	0.042

a) Aggregation time length = 5 min

$$\begin{cases} r_t = -0.111 r_{t-1}(f_{14} f_{15} + f_{15} - r_{t-1}) \\ f_{15} = f_1 + 0.106 r_{t-1} r_{t-3} - 0.098 r_{t-1} - 2\sin[pow(2, f_{12}) - 1] \\ f_{14} = \sin\left|\left[(f_{13} - 0.033)/1.048 - r_{t-1}\right] r_{t-5} - 1.745\right| \\ f_{13} = \left\{0.106\left|2\sin[pow(2, f_{12}) - 1] - r_{t-1}\right|/r_{t-3} - 0.178\right\}^2 \\ f_{12} = \left|\left[(0.106 r_{t-1} r_{t-3} - 0.098 r_{t-1})^2 - r_{t-1}\right]/r_{t-5} - 1.531\right| \\ f_{11} = f_7 f_9 + f_{10} - 0.106 f_{10}/(f_7 f_9 + f_{10}) + r_{t-5} \\ f_{10} = 0.64[0.61(f_9 - r_{t-5}) + r_{t-1} - 0.818]/r_{t-6} + 0.61 r_{t-5} \\ f_9 = \left|\sin\{\cos[(f_8 + f_6)/r_{t-1} + r_{t-6}]\} + r_{t-2}\right| \\ f_8 = \left\{\left[(f_6 + r_{t-5}) r_{t-5} - 1.531\right]^2 - 0.106/(0.468 - f_4)\right\}/f_7 + r_{t-6} \\ f_7 = -(0.106 r_{t-5} + 0.612) f_2 + f_6 \\ f_6 = 2\sin[f_5 - 0.106/(0.468 - f_4)]/r_{t-1} \\ f_5 = 0.211\cos(0.468 - f_4) - 0.106 r_{t-6} + 0.106 r_{t-4} - r_{t-5} - 0.706 \\ f_4 = \sin(0.106 r_{t-2} r_{t-3} - 0.106 r_{t-2} f_3 - r_{t-2} r_{t-5})/r_{t-1} \\ f_3 = [pow(2, -2 f_2/r_{t-2} - r_{t-3}) - r_{t-1} - 0.894](0.106 r_{t-5} + 2.612) f_2 \\ f_2 = \sin\left\{\left[(|\sin f_1| - r_{t-1})/r_{t-3} - 1.259\right] \times r_{t-1} - 0.106\right\} \\ f_1 = (0.806 - 0.712 r_{t-1} - 0.106 r_{t-1} r_{t-5})/r_{t-2} \end{cases}$$

b) Aggregation time length = 10 min

$$\begin{cases} r_t = 0.106(f_8 + f_3^4 f_4 - 2 f_3^5 + f_3^8 + 2 r_{t-5} + 2 r_{t-3}) \\ f_8 = [pow(2, 2 r_{t-5} + 0.212 f_7) - 1 - r_{t-2}] r_{t-4} - (f_1 + 2 f_3)/(1.779 + r_{t-5}) \\ f_7 = \{0.909 - pow[2, 1 - pow(2, f_6)]\}/r_{t-1} \\ f_6 = |pow(2, -1.024 f_5) - 1| + r_{t-5} \\ f_5 = 0.132(1.779 + r_{t-5})^2/(f_1 + 2 f_3) + r_{t-5} - r_{t-2} \\ f_4 = r_{t-1}/f_1 + \cos(f_1 - r_{t-3}/f_1)/(0.391 - r_{t-4}) \\ f_3 = -\cos\left\{\left[\cos(1.075 f_2^2) + r_{t-3}\right]/r_{t-2}\right\} \\ f_2 = \cos(f_1 - r_{t-3}/f_1)/(0.391 - r_{t-4}) - r_{t-1} \\ f_1 = \left[\sin^2(1.908 r_{t-1}) + 0.391 - r_{t-4}\right]/r_{t-4} + 0.391 - r_{t-4} \end{cases}$$

c) Aggregation time length = 15 min

$$\begin{cases} r_t = 0.341\sin f_{10} - 0.176 f_4 + 0.176 f_5 + 0.176 pow(2, f_5 + r_{t-2}) - 0.086 \\ f_{10} = 2\sin(-0.494\sin f_9/f_2 + 0.506 r_{t-7} + r_{t-8}) \\ f_9 = f_2 \sin f_7 - 2 f_3 + f_4 - |2\sin f_3 \times f_2 f_4| \\ f_8 = f_4 - f_5 - pow(2, f_5 + r_{t-2}) + 1 - 2\sin f_7 \\ f_7 = \{\sin[1 + f_2 - pow(2, f_6)] + r_{t-8}\}/r_{t-3} \\ f_6 = -r_{t-3} pow(2, f_5 + r_{t-2}) + r_{t-3} - r_{t-8} + 0.779 \\ f_5 = \sin\left[\left(|2\sin f_3 \times f_2 f_4| - f_4 - r_{t-5}\right) r_{t-3} + r_{t-7} + r_{t-8}\right] \\ f_4 = (r_{t-3} - 0.003) f_3 - \sin[(\sin f_2 + r_{t-8})/r_{t-5}] \\ f_3 = \sin\{\sin[(\sin f_2 + r_{t-8})/r_{t-5}]\} \\ f_2 = pow\{2, \sin[\cos(f_1 + r_{t-7})]\} - 1 \\ f_1 = \left|-0.96 r_{t-3} + 0.003\right|^{1/2} \times r_{t-5} r_{t-6} r_{t-7} - r_{t-8} \end{cases}$$

d) Aggregation time length = 20 min

$$\begin{cases} r_t = 0.779 pow[2, -1.51\sin(0.795 f_8 + f_1)] - 1.4 \\ f_8 = -1.087\sin(f_7 + f_6) - f_3/(\sin f_7 + f_6 + f_1) - f_6 - r_{t-3} \\ f_7 = pow(2, |f_5 - f_6| + 3 f_3 - f_1 + r_{t-5}) - 1 - r_{t-3} - f_6 \\ f_6 = 0.801(f_4 - f_3 - r_{t-3})(0.00006/r_{t-5}^2 - f_1^2)/f_2 \\ f_5 = -0.003 f_2/(0.00006/r_{t-5}^2 - f_1^2) - f_3 + 0.288 + r_{t-1} + f_1 \\ f_4 = \sin\left[-f_2\left(0.00006/r_{t-5}^2 - f_1^2\right) - r_{t-4}\right] - f_1 + r_{t-5} \\ f_3 = -0.00006/r_{t-5}^2 + f_1^2 + f_2(0.00006/r_{t-5}^2 - f_1^2) \\ f_2 = 1.866 r_{t-5}^2 f_1^4 - r_{t-1} - f_1 \\ f_1 = -1.361\sin(0.00006/r_{t-5}^2) + 0.236 \end{cases}$$

Fig. 7. The GP model for different time-aggregations at Site B

3.2. Predictive Performance under Normal Conditions

Tables 8 and 9 compare the predictive performance of the ARIMA models against that of the proposed hybrid models for Sites A and B under normal conditions. These two tables report four performance indexes on the validation dataset for Scenario 1 for different aggregation time intervals, including MAE, MRE, MSE and MSRE. As shown in Tables 8 and 9, the hybrid model produces better predictive performance than that of the ARIMA models for different aggregation intervals. By comparing the performance indexes for different aggregation time intervals, it can be found that the predictive performance of the hybrid method increases with an increase in the aggregation time interval. This may imply that data aggregation could suppress the effects of the measurement noises and useless traffic fluctuation information.

For further comparison of the predictive performance of the ARIMA and hybrid model, Figs 8 and 9 illustrate the predicted volumes of the models against the actual values for different time-aggregations at Sites A and B. In addition, Figs 8 and 9 also summarize the regression coefficients for the fitted linear relationship between the actual and predicted values. For different time-aggregations at both sites, the *R*-square values of the hybrid models are all greater than those of the ARIMA model, indicating that the predicted values of the hybrid method have higher correlation with the actual values.

The above results reveal that the hybrid models have better forecasting accuracy than the ARIMA model. This indicates the advance nature and effectiveness of combining the GP model with the ARIMA model. The hybrid strategy can better capture the characteristics of the traffic flow time series data. Moreover, the hybrid model can display a mathematical equation which can be easily used to forecast traffic volume in practice. For example, the hybrid model for 20-min interval at Site A is composed of a linear component and a nonlinear component. The linear component is estimated by the ARIMA model for the 20-min interval shown in Table 3, and the nonlinear component is obtained by the equations shown in Fig. 6.

For illustrative purposes, the prediction results of the hybrid model and the original observations for different aggregate time intervals at Sites A and B are shown in Figs 10 and 11. The hybrid model provides reasonably accurate forecasts of traffic volume. In general, the hybrid model has lower prediction errors for larger aggregation time intervals, and has higher prediction errors for greater traffic volumes.

Table 8. Comparison of the predictive performance of the models for Site A in Scenario 1

Aggregation level	ARIMA				ARIMA + GP			
	MAE[a]	MRE	MSE	MSRE	MAE[a]	MRE	MSE	MSRE
5 min	69.48	9.62%	4827.47	0.93%	54.02	6.93%	2918.59	0.48%
10 min	57.36	7.45%	3290.17	0.56%	36.12	5.32%	1304.65	0.28%
15 min	58.68	7.15%	3443.34	0.51%	36.36	5.29%	1322.05	0.28%
20 min	62.52	7.70%	3908.75	0.59%	30.48	4.10%	929.03	0.17%

Note: [a]The unit of the MAE is [vehicles/hour].

Table 9. Comparison of the predictive performance of the models for Site B in Scenario 1

Aggregation level	ARIMA				ARIMA + GP			
	MAE[a]	MRE	MSE	MSRE	MAE[a]	MRE	MSE	MSRE
5 min	74.76	8.24%	5589.06	0.68%	54.84	6.72%	3007.43	0.45%
10 min	68.69	7.32%	4718.04	0.54%	44.76	5.27%	2003.46	0.28%
15 min	70.90	7.09%	5026.24	0.50%	40.45	4.53%	1636.36	0.21%
20 min	79.81	7.84%	6369.96	0.61%	39.92	4.15%	1593.93	0.17%

Note: [a]The unit of the MAE is [vehicles/hour].

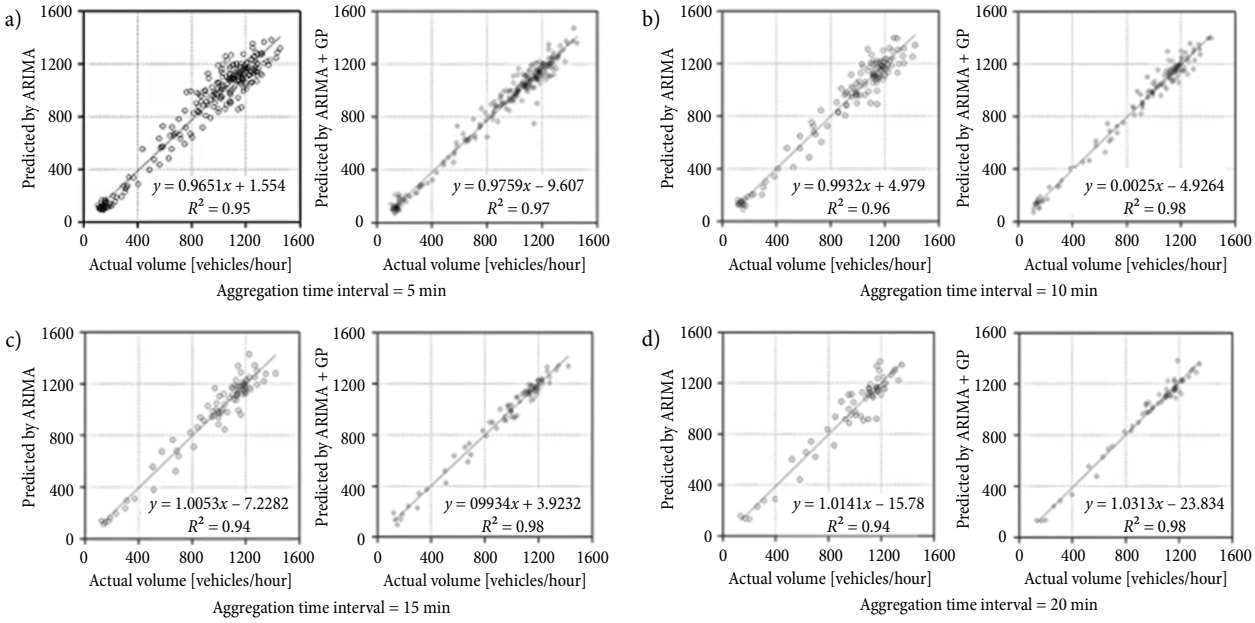

Fig. 8. Predicted versus actual volumes for different aggregation time intervals at Site A

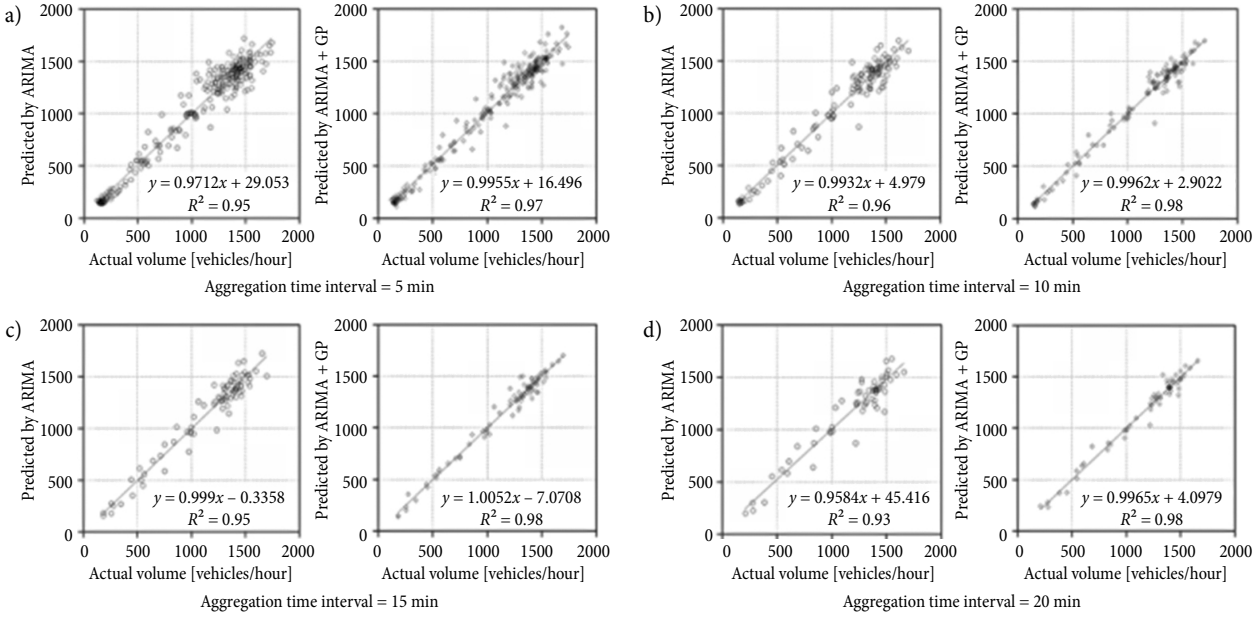

Fig. 9. Predicted versus actual volumes for different aggregation time intervals at Site B

Comparisons of prediction accuracy have also been made with several previous studies shown in Table 10. The prediction accuracy of the proposed is relatively good compared with the models in previous studies (Table 10). Table 10 also gives the improvements of the proposed models in previous studies over traditional models. It can be concluded that the improvements of the proposed model in this study are relatively high.

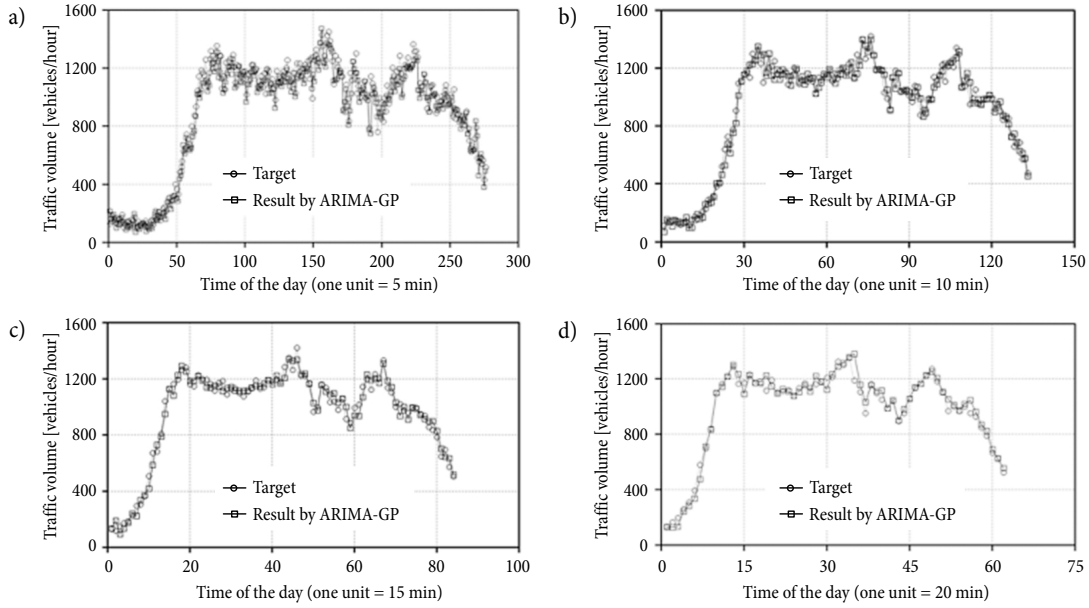

Fig. 10. Prediction results for different aggregation time intervals by using the hybrid method at Site A

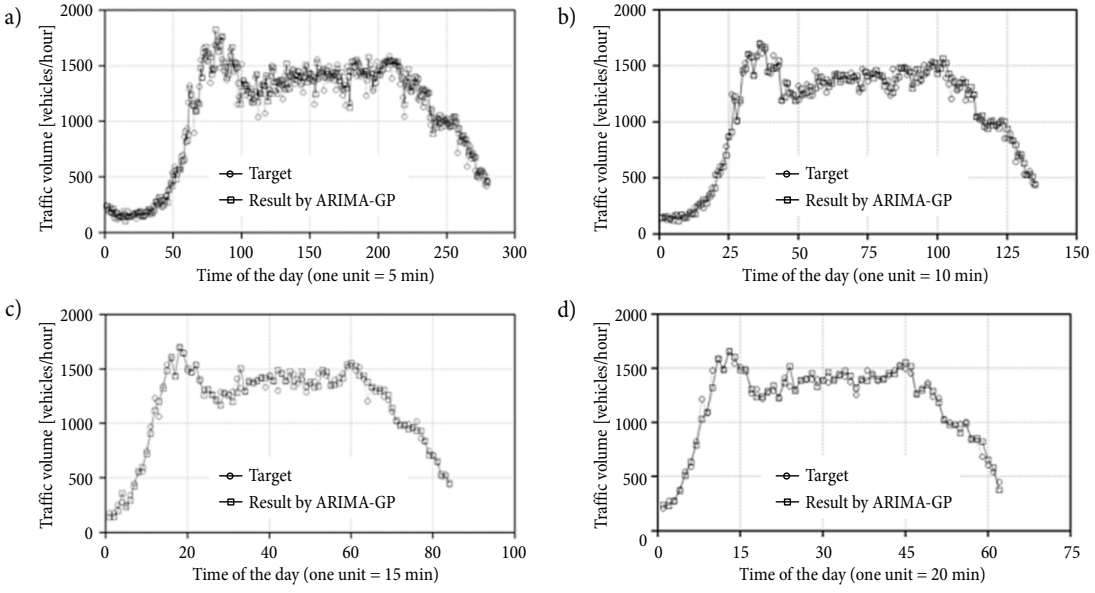

Fig. 11. Prediction results for different aggregation time intervals by using the hybrid method at Site B

Table 10. Prediction accuracy comparison

Authors	MRE of the proposed model in the study [%]	Improvement in MRE over traditional model [%]
Smith *et al.* (2002)	8.8	1.1
Zhang, Xie (2008)	8.8	2.3
Chandra, Al-Deek (2009)	9.4	1.4
Castro-Neto *et al.* (2009)	9.0	0.8
Lee, Tong (2011)	7.6	0.5
Dimitriou *et al.* (2008)	10.2	1.6

Table 11 gives the Central Processing Unit (CPU) times needed for the estimation of the hybrid model parameters, and the CPU times needed for the application of estimated hybrid models for one prediction using a desktop computer (3.4 GHz CPU and 8GB RAM).

Table 11. The CPU times for model calibrations and predictions

Aggregation level	Model parameter estimation [min]		One prediction using estimated model [sec]	
	Site A	Site B	Site A	Site B
5 min	33.5	38.2	<0.1	<0.1
10 min	21.7	25.3	<0.1	<0.1
15 min	26.1	18.7	<0.1	<0.1
20 min	19.7	24.6	<0.1	<0.1

Although calibrating a hybrid model needs a relatively long time, the estimated model needs very short time to make a prediction. The CPU running times required by one prediction of the estimated models are less than 0.1 second. Thus, the developed models have the potential to be used for online traffic control and management.

3.3. Predictive Performance under Atypical Conditions

The predictive performance of the hybrid model and the ARIMA model on the validation dataset for Scenario 2 (incident conditions) was tested. Since the durations of the most incidents on the I-880N freeway are lower than 60 minutes, we only tested the predictive performance of the hybrid model for the 5-min interval. The prediction model for the long time interval, such as the 20-min interval, can only make 3 predictions for a 60-min period. This may lead to unstable estimates of the predictive performance of the hybrid model. Fig. 12a and 12b illustrate the traffic flow data under incident conditions and the traffic data under normal conditions (average volumes across the 23 weekdays in May 2012). Traffic volumes under incident conditions were significantly lower than those observed on the normal weekdays.

Fig. 12c and 12d illustrate the actual values and predicted values from the ARIMA and hybrid models for two sites. During the period of incident, the predicted values of the hybrid models are more closed to actual values than those predicted by the ARIMA models for both sites, indicating that the predictive performance of the hybrid model is better than that of the ARIMA model even under incident conditions.

The predictive performance indexes of the hybrid model and the ARIMA model under incident conditions are given in Table 12. It should be noted that these performance indexes were calculated for the period that began about 20 minutes before the occurrence of the incident and ended about 20 minutes after the traffic flow back to normal conditions. Previous study suggested that this could help evaluate the models' capability of responding to unexpected changes in traffic flow, as well as the ability of these models to recover the prediction performance when traffic flow returns to the normal patterns (Castro-Neto *et al.* 2009). As shown in Table 12, compared with the ARIMA model, the hybrid model can increase the MRE by about 9% on the validation dataset for Scenario 2. Thus, combining the GP model with the ARIMA model can better capture the characteristics of the short-term traffic flow time series data under incident conditions.

Table 12. Comparison of the predictive performance of the ARIMA and hybrid models in Scenario 2

Selected sites	ARIMA		ARIMA + GP	
	MAE[a]	*MRE* [%]	MAE	*MRE* [%]
Site A	171.68	19.13	80.92	9.70
Site B	150.26	17.09	75.49	9.21

Note: [a]The unit of the MAE is [vehicles/hour].

Conclusions

This study proposed a hybrid methodology, which combines the ARIMA and GP models for short-term traffic flow forecasting. Compared with the models in previous studies, the proposed method has the following advantages. First, the hybrid model can better capture the linear and nonlinear patterns within traffic flow data and improve the predictive performance. Second, the GP technique in the hybrid model does not need pre-specified functional forms and can select the best functional form based on the training data. Finally, un-

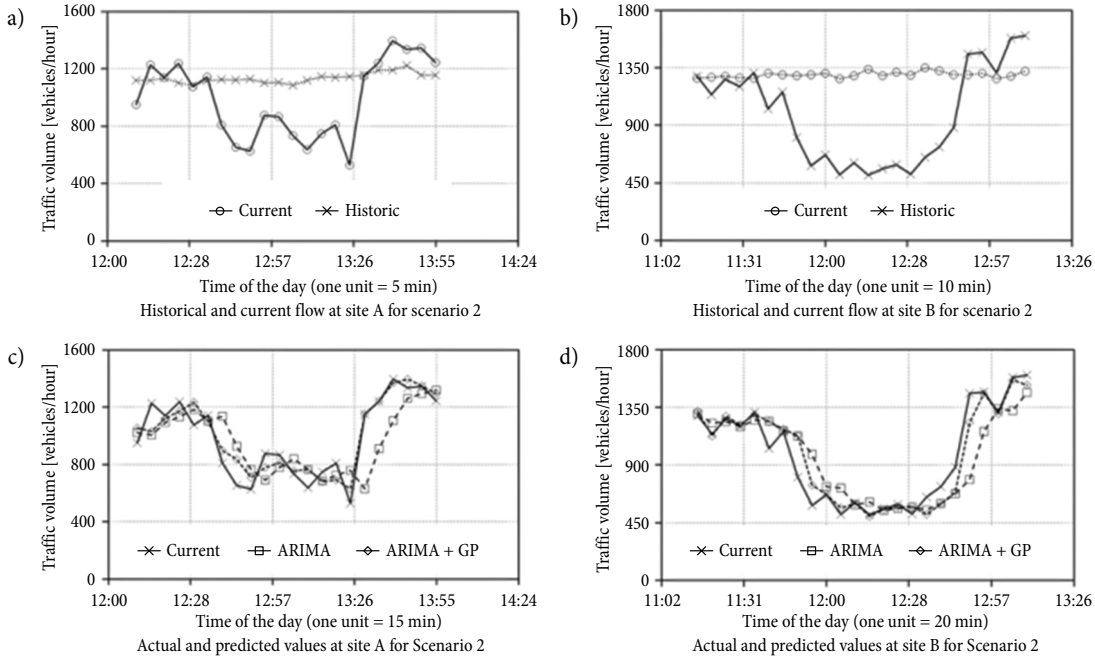

Fig. 12. Actual and predicted values from ARIMA models and hybrid models for two sites under incident conditions

like the ANN and SVM, which work as black boxes, the GP can generate easily readable mathematical equations. Thus, the proposed model can be easily applied in practical engineering applications. The major shortcoming of the proposed model is that the GP model is a computationally intense algorithm that requires a great amount of machine running time. It usually takes relatively long time for training a GP model when the number of observations in the training dataset is quite large. However, the calibrated model only needs extremely short time to make predictions.

The hybrid models were fitted for four different time-aggregations: 5, 10, 15, and 20 min. The validations were performed by using traffic data under both normal and incident conditions obtained from multiple locations on the I-880N freeway in the United States. The results showed that the hybrid models have better predictive performance than utilizing only ARIMA model for different aggregation time intervals under normal conditions. The MRE of the hybrid models was found to be from 4.1 to 6.9% for different aggregation time intervals under normal conditions. The predictive performance of the hybrid method increases with an increase in the aggregation time interval. In addition, the validation results also showed that the hybrid model can still produce satisfactory predictive performance under incident conditions. The predictive performance of the hybrid model is better than that of the ARIMA model under incident conditions.

With regard to the aggregation level, the hybrid model for 5-min interval is more appropriate for practical application. The reasons are as follows. First of all, for incident traffic conditions, the hybrid model is expected to forecast traffic flow in high resolution, as the dynamic traffic management system needs to mitigate and minimize the adverse effects of incidents in a timely fashion. In addition, for the normal traffic conditions, the hybrid model for 5-min interval can also achieve relatively good prediction accuracy of 93%. The hybrid model for 5-min interval can provide good prediction accuracy for both normal and incident traffic conditions. Second, the 5-min traffic data are commonly used in practical engineering. The hybrid model for 5-min interval can be easily applied in practical applications by using the 5-min traffic data. Finally, previous studies about short-term traffic-forecasting also recommended to developed prediction model for 5-min interval.

The proposed hybrid model has the potential to be used for short-term traffic flow forecasting in practice. However, before the hybrid method is used in practical applications, additional research is still needed to further improve the model predictive performance. First, the effects of the other factors such as time of the day and weather conditions could be considered. Incorporating these factors as input variables may further improve the model fitness. Second, this study only modelled the traffic data from a single isolated detector. By combing the traffic information from adjacent loop detectors, the predictive performance of the hybrid model may be further improved. Finally, additional traffic data from other freeways are needed to test the transferability of the proposed model. The authors recommend that future studies may focus on these issues.

Acknowledgements

This research was jointly sponsored by the China's National Key Basic Research Program (No. 2012CB725400) and National Natural Science Foundation of China (Project No. 51322810).

References

Aladag, C. H.; Egrioglu, E.; Kadilar, C. 2009. Forecasting nonlinear time series with a hybrid methodology, *Applied Mathematics Letters* 22(9): 1467–1470. http://doi.org/10.1016/j.aml.2009.02.006

Box, G. E. P.; Jenkins, G. M. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day. 575 p.

Castro-Neto, M.; Jeong, Y.-S.; Jeong, M.-K.; Han, L. D. 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions, *Expert Systems with Applications* 36(3): 6164–6173. http://doi.org/10.1016/j.eswa.2008.07.069

Chandra, S. R.; Al-Deek, H. 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 13(2): 53–72. http://doi.org/10.1080/15472450902858368

Chen, C.; Kwon, J.; Rice, J.; Skabardonis, A.; Varaiya, P. 2003. Detecting errors and imputing missing data for single-loop surveillance systems, *Transportation Research Record: Journal of the Transportation Research Board* 1855: 160–167. http://doi.org/10.3141/1855-20

Chen, C.; Wang, Y.; Li, L.; Hu, J.; Zhang, Z. 2012. The retrieval of intra-day trend and its influence on traffic prediction, *Transportation Research Part C: Emerging Technologies* 22: 103–118. http://doi.org/10.1016/j.trc.2011.12.006

Das, A.; Abdel-Aty, M.; Pande, A. 2010. Genetic programming to investigate design parameters contributing to crash occurrence on urban arterials, *Transportation Research Record: Journal of the Transportation Research Board* 2147: 25–32. http://doi.org/10.3141/2147-04

Dimitriou, L.; Tsekeris, T.; Stathopoulos, A. 2008. Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow, *Transportation Research Part C: Emerging Technologies* 16(5): 554–573. http://doi.org/10.1016/j.trc.2007.11.003

Dunne, S.; Ghosh, B. 2012. Regime-based short-term multivariate traffic condition forecasting algorithm, *Journal of Transportation Engineering* 138(4): 455–466. http://doi.org/10.1061/(ASCE)TE.1943-5436.0000337

Etemadi, H.; Rostamy, A. A. A.; Dehkordi, H. F. 2009. A genetic programming model for bankruptcy prediction: empirical evidence from Iran, *Expert Systems with Applications* 36(2): 3199–3207. http://doi.org/10.1016/j.eswa.2008.01.012

Ghosh, B.; Basu, B.; O'Mahony, M. 2007. Bayesian time-series model for short-term traffic flow forecasting, *Journal of Transportation Engineering* 133(3): 180–189. http://doi.org/10.1061/(ASCE)0733-947X(2007)133:3(180)

Ghosh, B.; Basu, B.; O'Mahony, M. 2005. Time-series modelling for forecasting vehicular traffic flow in Dublin, in

Transportation Research Board 84th Annual Meeting Compendium of Papers CD-ROM, 9–13 January 2005, Washington, DC, 1–22.

Guo, F.; Krishnan, R.; Polak, J. 2013. A computationally efficient two-stage method for short-term traffic prediction on urban roads, *Transportation Planning and Technology* 36(1): 62–75. http://doi.org/10.1080/03081060.2012.745721

Hamad, K.; Shourijeh, M. T.; Lee, E.; Faghri, A. 2009. Near-term travel speed prediction utilizing Hilbert–Huang transform, *Computer-Aided Civil and Infrastructure Engineering* 24(8): 551–576. http://doi.org/10.1111/j.1467-8667.2009.00620.x

Hamed, M. M.; Al-Masaeid, H. R.; Said, Z. M. B. 1995. Short-term prediction of traffic volume in urban arterials, *Journal of Transportation Engineering* 121(3): 249–254. http://doi.org/10.1061/(ASCE)0733-947X(1995)121:3(249)

Huang, S.; Sadek, A. W. 2009. A novel forecasting approach inspired by human memory: The example of short-term traffic volume forecasting, *Transportation Research Part C: Emerging Technologies* 17(5): 510–525. http://doi.org/10.1016/j.trc.2009.04.006

Koza, J. R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. A Bradford Book. 840 p.

Lee, Y.-S.; Tong, L.-I. 2011. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming, *Knowledge-Based Systems* 24(1): 66–72. http://doi.org/10.1016/j.knosys.2010.07.006

Lensberg, T.; Eilifsen, A.; McKee, T. E. 2006. Bankruptcy theory development and classification via genetic programming, *European Journal of Operational Research* 169(2): 677–697. http://doi.org/10.1016/j.ejor.2004.06.013

Min, W.; Wynter, L. 2011. Real-time road traffic prediction with spatio-temporal correlations, *Transportation Research Part C: Emerging Technologies* 19(4): 606–616. http://doi.org/10.1016/j.trc.2010.10.002

Oba, S.; Sato, M.-A.; Takemasa, I.; Monden, M.; Matsubara, K.-I.; Ishii, S. 2003. A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics* 19(6): 2088–2096. http://doi.org/10.1093/bioinformatics/btg287

Ong, C.-S.; Huang, J.-J.; Tzeng, G.-H. 2005. Building credit scoring models using genetic programming, *Expert Systems with Applications* 29(1): 41–47. http://doi.org/10.1016/j.eswa.2005.01.003

Qu, L.; Li, L.; Zhang, Y.; Hu, J. 2009. PPCA-based missing data imputation for traffic flow volume: a systematical approach, *IEEE Transactions on Intelligent Transportation Systems* 10(3): 512–522. http://doi.org/10.1109/TITS.2009.2026312

Smith, B. L.; Demetsky, M. J. 1997. Traffic flow forecasting: comparison of modeling approaches, *Journal of Transportation Engineering* 123(4): 261–266. http://doi.org/10.1061/(ASCE)0733-947X(1997)123:4(261)

Smith, B. L.; Williams, B. M.; Oswald, R. K. 2002. Comparison of parametric and nonparametric models for traffic flow forecasting, *Transportation Research Part C: Emerging Technologies* 10(4): 303–321. http://doi.org/10.1016/S0968-090X(02)00009-8

Stathopoulos, A.; Karlaftis, M. G. 2003. A multivariate state space approach for urban traffic flow modeling and prediction, *Transportation Research Part C: Emerging Technologies* 11(2): 121–135. http://doi.org/10.1016/S0968-090X(03)00004-4

Turochy, R. E. 2006. Enhancing short-term traffic forecasting with traffic condition information, *Journal of Transportation Engineering* 132(6): 469–474. http://doi.org/10.1061/(ASCE)0733-947X(2006)132:6(469)

Vanajakshi, L.; Rilett, L. R. 2004. A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed, in *2004 IEEE Intelligent Vehicles Symposium (IV2004)*, 14–17 June 2004, Parma, Italy, 194–199. http://doi.org/10.1109/ivs.2004.1336380

Van Lint, J.; Hoogendoorn, S.; Van Zuylen, H. 2002. Freeway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural networks, *Transportation Research Record: Journal of the Transportation Research Board* 1811: 30–39. http://doi.org/10.3141/1811-04

Vlahogianni, E. I.; Karlaftis, M. G.; Golias, J. C. 2007. Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks, *Computer-Aided Civil and Infrastructure Engineering* 22(5): 317–325. http://doi.org/10.1111/j.1467-8667.2007.00488.x

Vlahogianni, E. I.; Karlaftis, M. G.; Golias, J. C. 2005. Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach, *Transportation Research Part C: Emerging Technologies* 13(3): 211–234. http://doi.org/10.1016/j.trc.2005.04.007

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. http://doi.org/10.1016/j.trc.2012.08.004

Washington, S. P.; Karlaftis, M. G.; Mannering, F. 2003. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC. 440 p.

Wei, Y.; Chen, M.-C. 2012. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks, *Transportation Research Part C: Emerging Technologies* 21(1): 148–162. http://doi.org/10.1016/j.trc.2011.06.009

Williams, B. 2001. Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling, *Transportation Research Record: Journal of the Transportation Research Board* 1776: 194–200. http://doi.org/10.3141/1776-25

Williams, B. M.; Hoel, L. A. 2003. Modeling and forecasting vehicular traffic flow as a seasonal arima process: theoretical basis and empirical results, *Journal of Transportation Engineering* 129(6): 664–672. http://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664)

Xie, Y.; Zhang, Y. 2006. A wavelet network model for short-term traffic volume forecasting, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 10(3): 141–150. http://doi.org/10.1080/15472450600798551

Xin, W.; Hourdos, J.; Michalopoulos, P. 2006. Preprocessing volume input data for improved traffic simulation, *Transportation Research Record: Journal of the Transportation Research Board* 1965: 192–200. http://doi.org/10.3141/1965-20

Xu, C.; Wang, W.; Liu, P. 2013. A genetic programming model for real-time crash prediction on freeways, *IEEE Transactions on Intelligent Transportation Systems* 14(2): 574–586. http://doi.org/10.1109/TITS.2012.2226240

Zhang, G. P. 2003. Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing* 50: 159–175. http://doi.org/10.1016/S0925-2312(01)00702-0

Zhang, H. M. 2000. Recursive prediction of traffic conditions with neural network models, *Journal of Transportation Engineering* 126(6): 472–481. http://doi.org/10.1061/(ASCE)0733-947X(2000)126:6(472)

Zhang, N; Zhang, Y.; Lu, H. 2011. Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways, *Transportation Research Record: Journal of the Transportation Research Board* 2215: 85–92. http://doi.org/10.3141/2215-09

Zhang, Y.; Xie, Y. 2008. Forecasting of short-term freeway volume with v-support vector machines, *Transportation Research Record: Journal of the Transportation Research Board* 2024: 92–99. http://doi.org/10.3141/2024-11

Zhang, Y.; Ye, Z. 2008. Short-term traffic flow forecasting using fuzzy logic system methods, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 12(3): 102–112. http://doi.org/10.1080/15472450802262281

Zhong, M.; Lingras, P.; Sharma, S. 2004. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques, *Transportation Research Part C: Emerging Technologies* 12(2): 139–166. http://doi.org/10.1016/j.trc.2004.07.006