KSCE Journal of Civil Engineering (2018) 22(10):4107-4115

DOI 10.1007/s12205-018-0429-4

Transportation Engineering

pISSN 1226-7988, eISSN 1976-3808
www.springer.com/12205

An Algorithm for Traffic Flow Prediction Based on Improved SARIMA and GA

Xianglong Luo*, Liyao Niu, and Shengrui Zhang*****

Received April 5, 2017/Revised 1st: September 8, 2017, 2nd: November 30, 2017, 3rd: February 11, 2018/Accepted February 12, 2018/Published Online May 31, 2018

Abstract

The traffic flow prediction plays a key role in modern Intelligent Transportation Systems (ITS). Although great achievements have been made in traffic flow prediction, it is still a challenge to improve the prediction accuracy and reduce the operation time simultaneously. In this paper, we proposed a hybrid prediction methodology combined with improved seasonal autoregressive integrated moving average (ISARIMA) model and multi-input autoregressive (AR) model by genetic algorithm (GA) optimization. Since traffic flow data has strong spatio-temporal correlation with neighboring stations, GA is used to select those stations which are highly correlated with the prediction station. The ISARIMA model is used to predict the traffic flow in test station at first. A multi-input AR model with traffic flow data in optimal selected stations is built to predict the traffic flow in test station as well. The final prediction result can be gained by combining with the results of ISARIMA and multi-input AR model. The test results from traffic data provided by TDRL at UMD Data Center demonstrate that proposed algorithm has almost the same prediction accuracy with artificial neural networks (ANNS). However, its operation time is almost the same with SARIMA model. It is proved to be an effective method to perform traffic flow prediction.

Keywords: *traffic flow prediction, SARIMA, spatio-temporal correlation, GA*

1. Introduction

With the vigorous growth of economy and the improvement of living standards, numbers of vehicles are increasing continuously and sharply. Vehicles have made people's lives more comfortable and convenience, but meanwhile traffic congestion has also become a very serious problem. To alleviate traffic congestion and improve transportation operation efficiency, accurate and timely traffic flow information is strongly needed for individual drivers and Traffic Management Departments.

The importance of Intelligent Transportation Systems (ITS) has led to a rapid development of various methods in order to predict traffic condition accurately, including traffic flow distribution, average travel velocity and time. The top task of ITS is traffic flow prediction, because the accurate traffic flow prediction is the precondition of traffic guidance, planning and control. Traffic flow prediction is a complex issue, and a large number of algorithms have been proposed in the past few years. These methods can be classified roughly into three categories: parametric approach, non-parametric approach and hybrid approach.

In earlier times, a number of scholars proposed many developed parametric methods in traffic flow prediction, such as Historical Average (HA) method, Autoregressive Integrated Moving Average method (ARIMA) (Voort *et al*., 1996), Seasonal Autoregressive Integrated Moving Average method (SARIMA) (Williams and Hoel, 2003), and Kalman filter (Okutani and Stephanedes, 1984; Xie *et al*., 2007; Wang and Papageorgiou, 2005; Ji *et al*., 2010; Gong and Zhang, 2013). The parametric methods were widely used in traffic flow prediction, which were the basis of the presented algorithms. Due to the strong stochastic characteristics of the traffic flow, the prediction accuracy of simple parametric algorithms could not meet the requirement of current engineering application. Many non-parametric algorithms were proposed as well.

Zhang and Ye (2008) proposed a fuzzy logic system to improve traffic flow prediction accuracy. Sun *et al.* (2012) proposed a model for traffic flow prediction with graphic lasso and neural networks. Huang and Sun (2013) applied kernel regression with sparse metric learning to predict short-term traffic flow. Habtemichael and Cetin (2016) proposed a non-parametric and data driven methodology for short-term traffic prediction based on identifying similar traffic patterns with an enhanced K-nearest neighbor (KNN) algorithm. Hou and Sun (2015) developed a multilayer feedback neural network for traffic flow prediction. Wei *et al.* (2013) proposed a KNN based neuro-fuzzy system for time series prediction, and Cheng *et al.* (2016) provided a

*Associate Professor, School of Highway and School of Information Engineering, Chang'an University, Xi'an 710064, P.R. China (Corresponding Author, E-mail: xllluo@chd.edu.cn)
**Assistant Engineer, Sichuan Operations Center of Iflytek CO., LTD. Chengdu 610041, P.R. Chian (E-mail: niuliyao1993@163.com)
***Professor, School of Highway, Chang'an University, Xi'an 710064, P.R. China (E-mail: zhangsr@chd.edu.cn)

distance weighted fuzzy KNN Support Vector Machine (SVM) for traffic flow prediction. But the prediction accuracy of non-parametric models is greatly influenced by the training samples, and the fuzziness of model parameters makes them difficult to apply in practical engineering.

In purpose of improving the prediction accuracy, many hybrid algorithms combined with parametric methods and non-parametric methods were proposed. Hu *et al.* (2016) proposed a model combined with Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) for traffic flow prediction. PSO was used to search optimal Support Vector Regression (SVR) parameters. The negative effect that caused by noises of original data set was decreased. Moretti *et al.* (2015) proposed a statistical and neural network bagging ensemble hybrid model for short-term traffic flow prediction. Cong *et al.* (2016) proposed a traffic flow prediction model combined with SVM and fruit fly optimization. Yang and Hu (2016) used a parameters optimized wavelet neural network for traffic flow prediction. Szeto *et al*. (2009), Smith *et al.* (2002), Lin *et al.* (2013), and Lippi *et al.* (2013) provided a comparative analysis of parametric and nonparametric models. Although the prediction accuracy of nonparametric methods and hybrid methods are superior to parametric methods, the operation time is longer and training sample sizes are more.

Because traffic flow data has high spatio-temporal correlation characteristics, many scholars considered to improve prediction accuracy with the spatio-temporal correlations. Sun *et al.* (2006), Sun and Xu (2011) proposed Bayesian network approaches to forecast traffic flow on a link with spatial traffic flow data from adjacent road links. Min and Wynter (2011) developed an extended time-sequence-based method that incorporated temporal and spatial interactions. Bernaś *et al.* (2015) developed an improved KNN model combined pre-segmentation of detector locations area based on traffic flow patterns. Pan *et al.* (2013) used a stochastic cell transmission framework to predict short-term traffic flow by considering the spatio-temporal correlation in the network. Li *et al.* (2015) proposed a robust causal dependence data mining algorithm in big data network and its application to traffic flow predictions. They provided Granger causality strategy into processing the raw large size of traffic data to produce a highly dependent traffic flow data. However, all these methods only considered the data closed to the prediction station, which could not fully reveal the spatio-temporal characteristics of traffic flow data.

In this paper, aiming at the characteristics of the SARIMA model with shorter operation time and less satisfied prediction performance, and the nonparametric model with better prediction performance and longer operation time, a hybrid prediction methodology is proposed combined with improved SARIMA (ISARIMA) model and Genetic Algorithm (GA) according to the high spatio-temporal correlation characteristics of traffic flow data. Based on the issue that SARIMA model parameters have the great influence on prediction accuracy, a sliding-window function is introduced into SARIMA model to optimize model parameters. Consisting of high spatio-temporal correlation characteristics of traffic flow data, Genetic Algorithm (GA) is used to select mostly related neighboring stations with the testing station. A multi-input autoregressive (AR) model is used with traffic flow data in optimal selected stations. Finally, the result of ISARIMA is combined with multi-input AR model, and the final prediction results are gained. The tested results from traffic data provided by the Transportation Research Data Lab (TDRL) at the University of Minnesota Duluth (UMD) Data Center present that proposed algorithm has almost the same prediction accuracy with artificial neural networks (ANNS), but its operation time is almost the same with SARIMA model.

The rest of this paper is organized as follows. Section 2 gives details on a hybrid traffic prediction method based on ISARIMA and GA. In section 3, the dataset used is introduced for the numerical experiments. The results and performance evaluation are presented in section 4. Finally, the conclusions and the future research are stated in section 5.

2. Method

2.1 ISARIMA

A brief presentation of the SARIMA model form is given below. A time series $\{Y_t\}$ can be defined by the Eq. (1) and (2).

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)e_t \tag{1}$$

$$\begin{aligned}
&\Phi(z) = 1-\Phi_1 z-\ldots-\Phi_p z^P \\
&\varphi(z) = 1-\varphi_1 z-\ldots-\varphi_p z^p \\
&\Theta(z) = 1-\Theta_1 z-\ldots-\Theta_Q z^Q \\
&\theta(z) = 1-\theta_1 z-\ldots-\theta_q z^q
\end{aligned} \tag{2}$$

Where B is backshift operator defined by $B^a Y_t = Y_{t-a}$. The parameters p and P represent the non-periodic and periodic autoregressive polynomial order. $\varphi_i(1 \le i \le p)$ is non-periodic autoregressive model parameters, and $\Phi_i(1 \le i \le P)$ is the periodic autoregressive mode parameters. The parameters q and Q represent the non-periodic and periodic moving average polynomial order. $\theta_i(1 \le i \le q)$ is non-periodic moving average model parameters, and $\Theta_i(1 \le i \le Q)$ is periodic moving average model parameters. The parameter d and D represent the order of normal differencing and periodic differencing. S is the period of the time series. e_t is generally regarded as the Gaussian white noise with variance σ^2.

For a time series, if the time dependencies in the expected values and covariance are small relative to the nominal level, the series may be close to stationary, and it can be modeled as an SARIMA. Therefore, in order to yield the nearly stationary transformation of the raw series, SARIMA model begins with selecting the normal and periodic differencing scheme. The best model orders are selected by ACFs (Auto-Correlation Function), PACFs (Partial Auto-Correlation Function) and Akaike's information criteria (AIC). After the model order is determined in traditional SARIMA model, the delay factor of the difference equation is increasing item by item, which cannot fully reveal the temporal

correlation characteristics of traffic flow data. In this paper, the positions where the values of ACFs and PACFs were more than a threshold as delay factors were selected. A sliding-window function was introduced to update training set in order to synchronize with the prediction station.

In this paper, through experiment, when differencing order d and D were set 1, that was $d = D = 1$, the traffic flow series can meet the conditions for stationary. According to Williams (2003), when the P and Q were also set to 1, $P = Q = 1$, SARIMA model emerged as the preferred model. Then the Eq. (1) also can be described as

$$Y_t = \Phi_1 * Y_{t-s} + Y_{SAR} + Y_{SMA} \tag{3}$$

In Eq. (3), Φ_1 is the periodic AR parameter. We introduced a sliding-window function S_Δ, and the ISARIMA model can be expressed in Eq. (4)

$$Y_{t-\Delta} = S_\Delta \left[\Phi_1 Y_{t-\Delta-s} + Y_{SAR-\Delta} + Y_{SMA-\Delta} \right] \tag{4}$$

Where $Y_{SAR-\Delta}$ and $Y_{SMA-\Delta}$ can be calculated by Eq. (5).

$$\begin{aligned} Y_{SAR-\Delta} &= \sum_{i=0}^{p} \varphi_i (Y_{t-\Delta-t(i)} - \Phi_1 * Y_{t-\Delta-s-t(i)}) \\ Y_{SMA-\Delta} &= \sum_{j=0}^{q} \theta_j (e_{t-\Delta-t(j)} - \Theta_1 * e_{t-\Delta-s-t(j)}) \end{aligned} \tag{5}$$

Where Θ_1 is periodic moving average model parameter, and φ_i and θ_j are non-periodic autoregressive and moving average model parameters. $t(i)$ and $t(j)$ respectively are the corresponding time delay that PACFs and ACFs are larger than the threshold. The sliding-window S_Δ can be generated as

$$S_\Delta = \begin{cases} \varepsilon(t-\Delta) - \varepsilon(t-L-\Delta) & t > 0, \Delta \geq 0 \\ 0 & others \end{cases} \tag{6}$$

L is the length of training set and also the size of the sliding-window function S_Δ. Δ is the time duration of S_Δ, which ranges from 1 to L, and $\varepsilon(t)$ is the unit step function.

2.2 GA Optimization

GA is a stochastic search algorithm based on the theory of the biological evolution, including selection, crossover and mutation operations, and it is a widely used method for parameters optimization. In GA, a set of possible solutions as population are generated, and each possible solution is treated as chromosome or an individual in population. Randomly selected individuals in the population are used as the initial solution, the fitness values are calculated after crossover and mutation operation, and individuals will be kept or eliminated according to the values of the fitness function in the population. After several times iterations, the optimal solution can be finally obtained.

In this paper, GA is used to select those stations which were highly correlated with the test station. The flowchart of the optimization method is shown in Fig. 1, and the detailed calculation process is shown in algorithm 1.

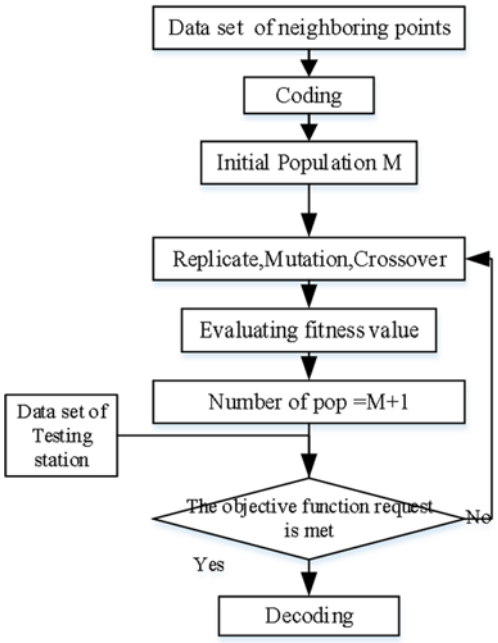

Fig. 1. The Flowchart of GA

Algorithm 1: Iteration Steps of GA optimization

Step 1: generate a $M \times N$ chromosome code matrix $C_{M\times N}$, M is the number of chromosome. N is the number of code element in per chromosome, and is the station numbers in this experiment.

Step 2: Initialization:
- i. Initialize the matrix $C_{M\times N}$ with 0 and 1 stochastically, 0 represents that the station is not selected, and 1 represents that the station is selected;
- ii. Set crossover rate p_c, mutation rate p_m and iterative threshold ε;
- iii. Generate a population and get an average traffic flow series Y_{N_0}. N_0 is the initial ID;
- iv. Calculate initial fitness value by cross-correlation $R(Y_{N_0}, Y)$, Y is the historical data in the prediction station;

Step 3: GA optimizing:

While (m < M), do
- i. Do crossover operation and mutation operation between two individuals of population, and get a new average traffic flow series Y_{N_m} ($0 < m \leq M$).
- ii. Evaluate fitness value by $R(Y_{N_m}, Y)$;
- iii. Update population according to iterative threshold ε;
- iv. m++;

Note: $R(Y_{N_m}, Y)$ expressed in Eq. (7), is cross-correlation, and it is fitness function of GA optimization. Where T is the length of traffic flow time series, $\overline{Y}_{N_m}$ is the mean value of Y_{N_m}, and $\overline{Y}$ is the mean value of Y.

$$R(Y_{N_m}, Y) = \frac{\sum_{t=1}^{T} (Y_{N_m,t} - \overline{Y}_{N_m})(Y_t - \overline{Y}) / T}{\sqrt{\sum_{t=1}^{T} (Y_{N_m,t} - \overline{Y}_{N_m})^2 / T} \sqrt{\sum_{t=1}^{T} (Y_t - \overline{Y})^2 / T}} \tag{7}$$

$R(Y_{N_m}, Y)$ represents the cross-correlation factor of station N_m and testing station. It could prove the similarity of traffic flow data between testing station and neighboring stations. It ranges from 0 to 1. In this paper, threshold ε ranges from 0.90 to 0.99 according to experiments.

2.3 Prediction Method

A hybrid prediction methodology is proposed combined with ISARIMA model and GA. ISARIMA model is used to predict

the traffic flow in test station at first. GA is used to select highly correlated detecting stations with the test station. Then a multi-input AR model is built to predict the traffic flow in test station. The final result can be calculated by Eq. (8).

$$Y_t = \lambda Y_{ISARIMA} + \gamma Y_{GA-AR} + c \tag{8}$$

Where $Y_{ISARIMA}$ is the prediction result of ISARIMA with sliding-window function, $Y_{GA\text{-}AR}$ is the prediction result of multi-input AR with neighboring detecting stations optimized by GA, c represents the constant term. λ, γ, and c represents the parameters of linear regression. Y_{ISRIMA} and $Y_{GA\text{-}AR}$ can be calculated as Eq. (9) and (10).

$$Y_{ISARIMA} = S_{\Delta}\left[\Phi_1 Y_{t-\Delta-s} + Y_{SAR-\Delta} + Y_{SMA-\Delta}\right] \tag{9}$$

$$Y_{GA-AR} = \sum_{i=1}^{N_{opt}} \alpha_i Y_i \tag{10}$$

In Eq. (9), the training set is the historical data of the testing station. In Eq. (10), N_{opt} means the optimal number of stations by GA, which reflect the spatial correlation. $Y_i = [Y_{t-1}^i, Y_{t-2}^i, \cdots, Y_{t-L}^i]^T$ represents traffic flow data in the i^{th} related station at the (t-1) time step to the (t-L) time step. $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iL}]$ is a weight vector to reflect the temporal correlation. L is the time lag.

The data is obtained immediately in the previous step is essential in traffic flow prediction. In our experiment, it is found that the performance improvement was minor when adding the time lag L. Thus, L = 1 is set and the Eq. (10) is simplified as Eq. (11).

$$Y_{GA-AR} = \sum_{i=1}^{N_{opt}} \alpha_i Y_i = \alpha_1 Y_{t-1}^1 + \alpha_2 Y_{t-1}^2 + \cdots + \alpha_{N_{opt}} Y_{t-1}^{N_{opt}} \tag{11}$$

$Y_{GA\text{-}AR}$ is the prediction traffic flow in the test station. The historical traffic flow is taken in optimized stations by GA into Eq. (11). The estimation value of model parameters $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{N_{opt}}]$ is obtained.

The flowchart of the proposed method is described in Fig. 2.

3. Data Description

The data used to evaluate the performance of the proposed model was collected in mainline detectors provided by TDRL at UMD Data Center from June 22nd, 2015 to June 26th, 2015. The sampling period of the testing dataset was 5-min. To alleviate the impact of outliers in 5-min traffic flow data, these data were aggregated into 15-min time intervals according to the recommendations of Highway Capacity Manual (USA) in this paper. There were totally 480 data points in the dataset. The data were divided into two data sets, the first 384 data points were used as the training sample, while the remaining 96 data points were served as the testing sample for measuring forecasting performance of the proposed model. Fig. 3 shows the station locations that are used. There were 26 main stations in mainline. The station S407 was located near a transportation hub in road networks in the experiments. Therefore, it was selected as the test station for the traffic flow prediction.

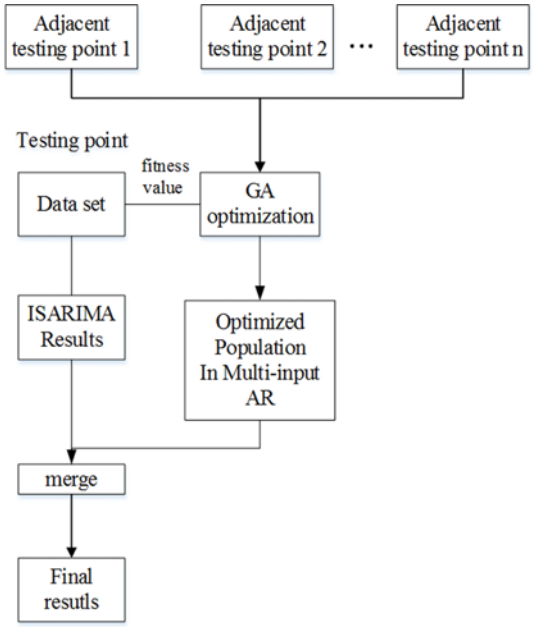

Fig. 2. The Flowchart of the Proposed Method

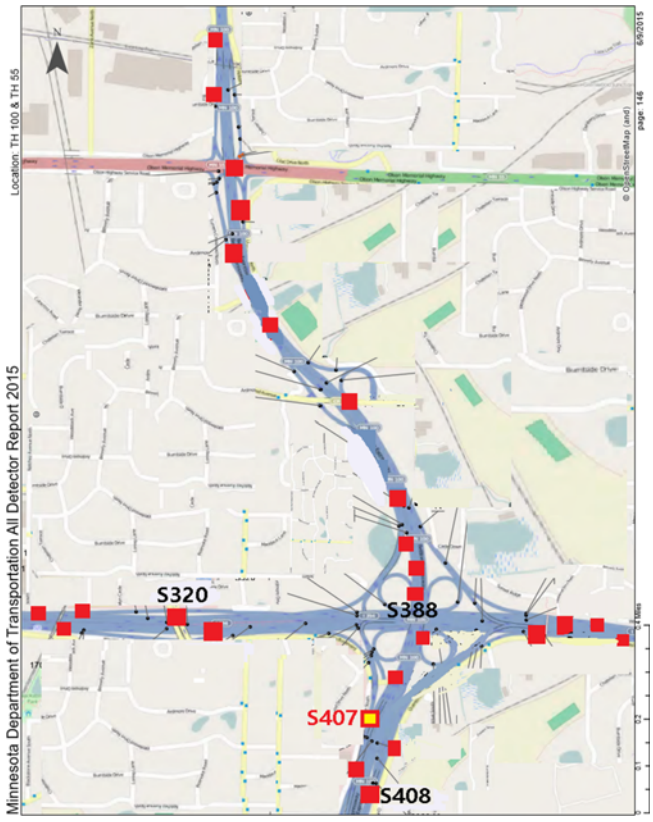

Fig. 3. The Station Locations that are Used in Experiment

Typical weekly traffic flows in the station S407 and three neighboring stations are shown in Fig. 4. From Fig. 4, it can be observed that the traffic flow distribution from Monday to Friday is almost the same mode in S407, and is obviously different from weekends. But the data distribution in neighboring stations is

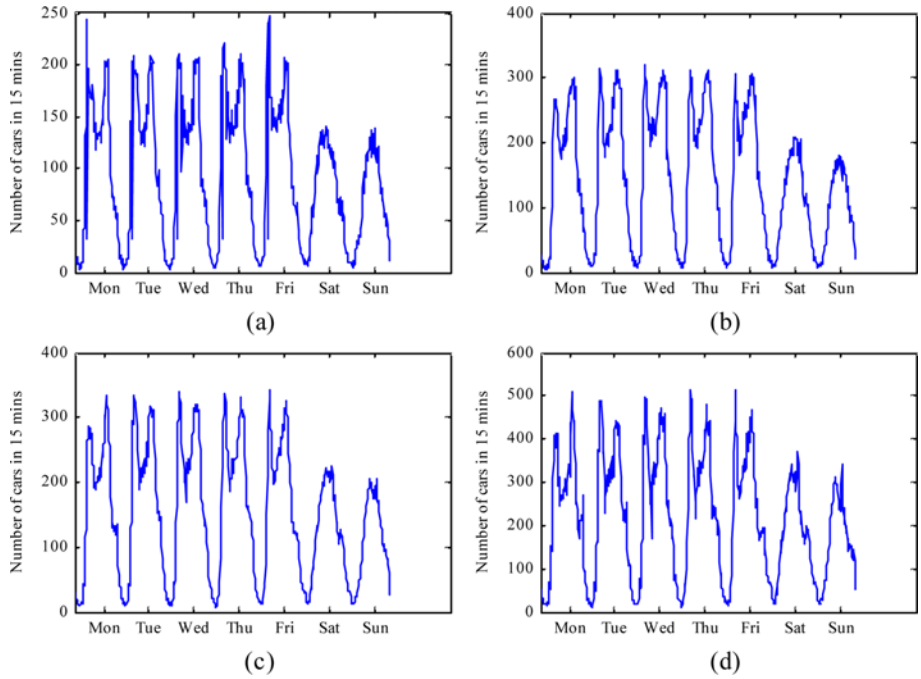

Fig. 4. Typical Weekly Patterns in the Station 407 and Other Three Neighboring Stations from June 22nd to June 28th, 2015: (a) S407, (b) S388, (c) S408, (d) S320

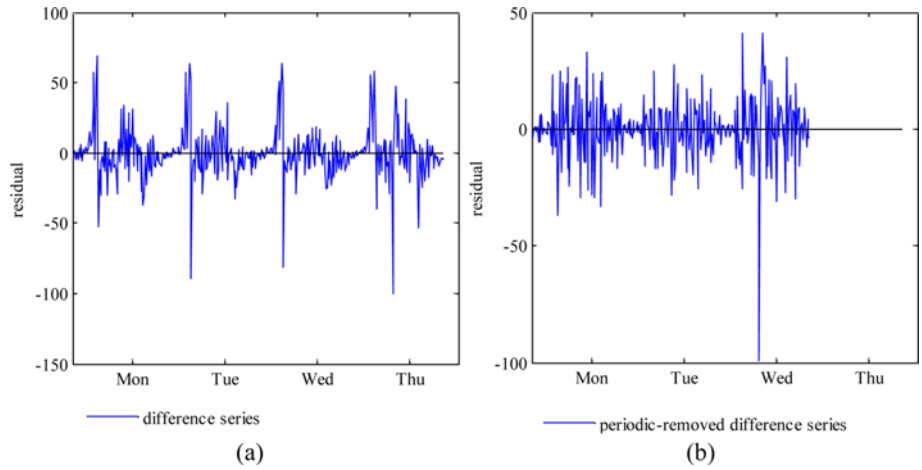

Fig. 5. Difference Residual Series in the Station S407: (a) 1st Order Difference Series from June 22nd, 2015 to June 26th, 2015, (b) Periodic-removed 1st Order Difference Series from June 22nd, 2015 to June 26th, 2015

similar with the station S407 during the whole week. There are some differences in the characteristics of traffic data between weekends and workdays. In order to ensure the stability of the experiment, we only used the data in workdays as the test data.

4. Results and Performance Evaluation

4.1 Model Parameter Estimation of ISARIMA

The 1st order difference series and periodic-removed 1st order difference of traffic flow data in station S407 is presented respectively in Fig. 5(a) and (b). It is observed that the 1st order difference series show a strong periodic pattern. As discussed above, traffic flow can be predicted as a typical time series only if it is stationary stochastic series with respect to mean and variance, so the periodic effects of the traffic flow data must be removed. As is shown in Fig. 5(a), the time interval between each adjacent 2 extreme points is 96, which means the period is 96, namely the periodic lag a for backshift operator B^a in Eq. (1). The periodic difference can be calculated by Eq. (12).

$$y'_{407t} = y_{407t}(B^{96})|_{t=a+1,\ldots,N} \tag{12}$$

Where y_{407t} is original data, and y'_{407t} is periodic-removed data set.

From Fig. 5(b), it can be found that the periodic effects were removed completely by 1st order periodic difference, and the processed periodic-removed 1st order difference series can be

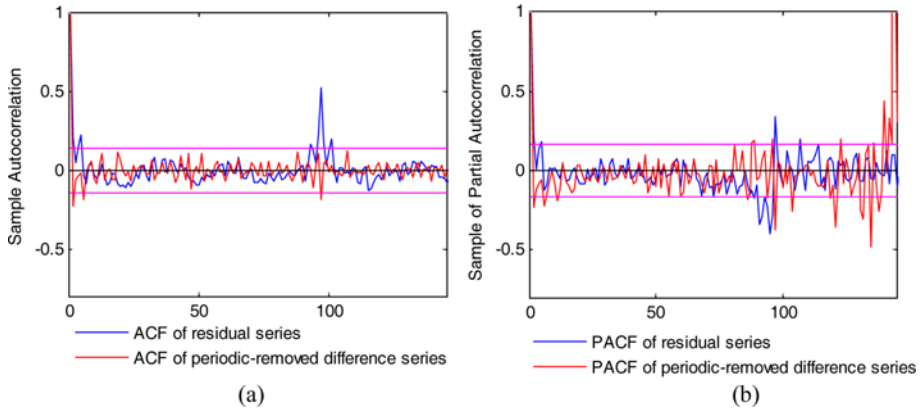

Fig. 6. ACFs and PACFs for 1st Order Difference Residual Series and Periodic-removed 1st Order Difference Residual Series in the Station 407: (a) ACFs, (b) PACFs

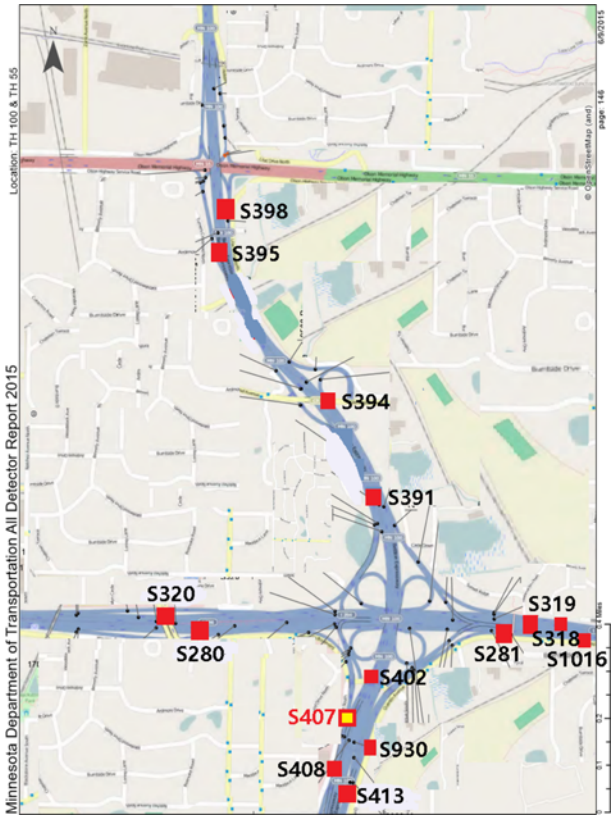

Fig. 7. Locations of Optimized Stations

analyzed as a stationary random sequence. In order to estimate the model order, we respectively calculated the ACFs and PACFs for original residual series and periodic-removed 1st order difference residual series. The results were shown in Fig. 6(a) and 6(b). From Fig. 6, it is shown that the PACF and ACF of periodic-removed residual series did not show the pronounced scattering phenomenon compared to the original residual series. The threshold value was set as twice times of the standard deviation for periodic-removed residual series, and then the parameters of p and q can be identified. In this paper, these parameters are updated with the updating of the training set by sliding-window function.

4.2 Optimization Selection of Correlated Detecting Stations

In order to select those stations highly correlated with the test station, GA is used as optimization method. After several times of experiments, crossover rate p_c is set as 0.95 and motion rate p_m as 0.05. The maximum generation number M ranged from 30 to 80 in this study.

The locations and ID of optimized stations are shown respectively in Fig. 7 and Table 1. From Fig. 7 and Table 1, it can be seen that the locations of optimized stations are a little different from classical understanding, that is, closer stations from the prediction station have more correlation than those further stations. In fact, many further stations have great correlation with the prediction station. The numbers of upstream stations are more than downstream stations, while there are some downstream stations are related with the station, which means that the effect of the upstream flow is more important in traffic flow prediction.

4.3 Prediction Results

The final prediction results are shown as follows.

$$\begin{aligned} Y_{t-\Delta} &= \lambda Y_{ISARIMA} + \gamma Y_{GA-AR} + c \\ &= 0.7316 \times S_{\Delta}\left[\Phi_1 * Y_{t-\Delta-96} + Y_{SAR-\Delta} + Y_{SMA-\Delta}\right] \\ &\quad + 0.2379 Y_{GA-AR} + 19.233 \end{aligned} \tag{13}$$

It consists of two parts of prediction results, where $Y_{SAR-\Delta}$ and $Y_{SMA-\Delta}$ can be represented as Eq. (14), and the second part is described in Eq. (15).

$$\begin{aligned} Y_{SAR-\Delta} &= \sum_{i=0}^{p} \varphi_i \left(Y_{t-\Delta-t(i)} - \Phi 1 * Y_{t-\Delta-96-t(i)}\right) \\ Y_{SMA-\Delta} &= \sum_{j=0}^{q} \theta_j \left(e_{t-\Delta-t(j)} - \Theta 1 * e_{t-\Delta-96-t(j)}\right) \end{aligned} \tag{14}$$

$$Y_{GA-AR} = \sum_{i=1}^{N_{opt}} \alpha_i Y_i = \alpha_1 Y_{t-1}^1 + \alpha_2 Y_{t-1}^2 + \cdots + \alpha_{14} Y_{t-1}^{14} \tag{15}$$

In Eq. (14), the parameters p, q etc. in $Y_{SAR-\Delta}$ and $Y_{SMA-\Delta}$ are updating with the updating of training set. The parameter λ is

Table 1. The Value of Parameter α and the Station ID we Selected

Station ID	α	Station ID	α
S280	0.9147	S395	0.5845
S281	0.7334	S398	0.8162
S318	0.171	S402	0.4881
S319	0.1832	S408	0.5702
S320	0.5993	S413	0.2555
S391	0.6538	S930	0.6425
S394	0.8263	S1016	0.6614

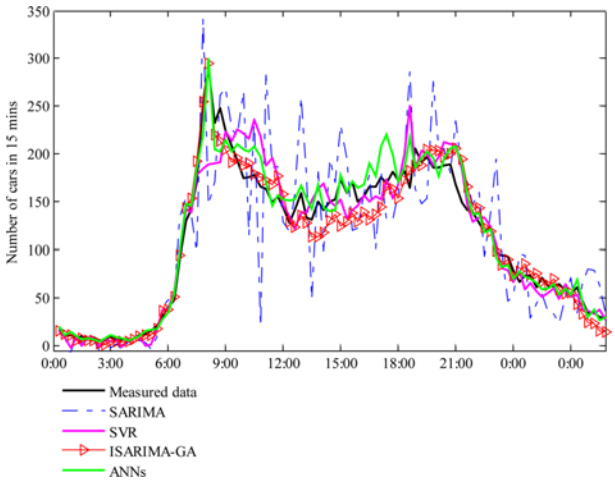

Fig. 8. Prediction Results with Different Methods Compared with Measured Traffic Flow Data

0.7316, which means the historical data of S407 plays a key role in prediction. The parameter γ is 0.2379, which reflects the impact of optimal stations on the prediction results. The constant term in Eq. (13) can be seen as a fixing value to alleviate the extreme candidate values, and the parameter N_{opt} means the number of neighboring stations optimized by GA, in the experiment N_{opt} = 14. The constant term C is 19.233.

The selected station ID and corresponding weight coefficient α_i are shown in Table 1. The weight coefficient α_i can be estimated by regression of historical data. It can be seen that α_i is almost inversely proportional to the distance from the prediction station. To verity the feasibility and efficiency of proposed algorithm, the prediction results are compared with SARIMA, SVR and ANNS. In SARIMA model, p is set as 5, q as 4, and d as 2. In ANNS model, a three-layer Back Propagation (BP) network is used with 10 hidden neurons and one output with sigmoid as activation function for all neurons. The learning rate is set as 0.05, the number of iteration as 400, and target error as 0.001. In SVR model, kernel function is set as Radial Basis Function (RBF), the penalty parameter of the error term as 300.

Figure 8 shows the 15-min traffic flow prediction results from 0:00 to 24:00 June, 26 in 2015 in station 407 with different methods and measured traffic flow data. It can be found that the proposed algorithm has better performance than others, especially in morning and evening peak hours, and the prediction value is almost coincided with the measured data.

4.4 Performance Evaluation

In order to evaluate the prediction performance, Root Mean Square Error (RMSE), which was the most frequently used metrics of prediction performance in previous work, and predicting accuracy (ACC) was chosen to evaluate the difference between the actual values with predicted values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{16}$$

$$ACC = (1-\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|)\times 100\% \tag{17}$$

Where N is the length of prediction data, y_i and $\hat{y}_i$ are the measured value and predicted for i^{th} validation sample respectively.

The performance is shown in Table 2 based on different correlated stations selected by different threshold ε. The larger ε we set, the less number of correlated stations are selected by GA. It can be seen that, if ε gets too large or too small, either ACC or RMSE is not well. Based on Table 2, threshold is set as 0.95.

It is shown in Table 3 that the prediction performance with different training series. The forecasting accuracy with proposed model gets 5.83% improvement over with one station data and 12.26% improvement over with neighboring stations without GA optimization. The traffic flow data has strong spatio-temporal

Table 2. RMSE and ACC of Different Threshold in GA

Threshold ε	Num. of correlated stations	RMSE (15-min)	ACC (15-min)
0.90	19	21.24	79.89
0.91	19	21.24	79.89
0.92	17	21.24	79.89
0.93	14	16.14	87.00
0.94	14	16.13	87.05
0.95	14	16.01	87.21
0.96	13	17.45	85.88
0.97	10	17.88	84.31
0.98	9	18.02	83.01
0.99	7	19.78	81.09

Table 3. RMSE and ACC of Different Training Set

Models	RMSE (15-min)	ACC (15-min)
ISARIMA based on one station	19.54	81.12
ISARIMA based on GA optimized stations	16.18	86.95
ISARIMA based on neighboring stations without GA	32.45	74.29

Table 4. Prediction Performance and Average Operation Time in Different Models

Models	RMSE (15-min)	ACC (15-min)	Average operation time
SVR	20.56	80.29	14.05 min
SARIMA	34.18	72.29	11.06 min
ANNs	18.77	89.05	35.5 min
ISARIMA-GA	16.18	86.95	12.4 min

correlation with neighboring detecting stations, but if all neighboring detector data are directly used to predict without optimization selection, the prediction accuracy cannot be improved, and even becomes worse.

The performance of several methods used in this paper is shown in Table 4. The 15-min ahead forecasting accuracy of SARIMA is the lowest, and ANNS method has the best prediction accuracy. The ACC of the proposed method respectively improved 14.7% and 6.7% compared with SARIMA, SVR, and was slightly less than the ANNS method. The prediction accuracy of the proposed method was close to the ANNS method, and they are superior to other methods. However, for the average operation time, ANNs method is 3 times longer than SARIMA, and is over 2 times longer than proposed method. The proposed method has almost the same prediction accuracy with ANNS, but its operation time is almost the same with SARIMA model. Therefore, the proposed method is an effective method to traffic flow prediction.

5. Conclusions

In this paper, the spatio-temporal characteristics of traffic flow data is considered, and a hybrid prediction methodology combined ISARIMA model and GA is proposed. An improved method to optimize the SARIMA model parameters is elaborated. GA is used to select highly correlated stations with the test station. A hybrid prediction model was proposed with ISARIMA multi-input AR optimized by GA. Test results with real traffic data provided by TDRL show that proposed method has good prediction performance compared with other methods. Since the traffic flow data is affected by weather, incident and other factors, the impact of these factors on traffic flow data will be further studied so as to improve the prediction accuracy. We will also further optimize the algorithm and reduce the operation time to meet the requirement of real-time prediction.

Acknowledgements

This research is partly supported by Grand Science and Technology Special of the Ministry of Transport of China (Grant no. 2011318812260). The authors express their gratitude to the UMD Data Center (TDRL) for providing the data.

References

Bernaś, M., Płaczek, B., Porwik, P., and Pamuła, T. (2015). "Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction." *IET Intelligent Transport Systems*, Vol. 9, No. 3, pp. 264-274, DOI: 10.1049/iet-its.2013.0164.

Cheng, Y. W., Wen, T. J., Cheng, H. C., and Yang, C. Y. (2016). "Distance weighted fuzzy KNN SVM." *IEEE 13ih International Conference on Networking, Sensing and Control,* Mexico City, Mexico, pp. 1-6, DOI: 10.1109/ICNSC.2016.7478976.

Cong, Y., Wang, J., and Li, X. (2016). "Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm." *Procedia Engineering*, Vol. 137, pp. 59-68, DOI: 10.1016/j.proeng.2016.01.234.

Gong, Y. and Zhang, Y. (2013). "Research of short-term traffic volume prediction based on Kalman filtering." *The 6th International Conference on Intelligent Networks and Intelligent Systems*, Shenyang, China, pp. 99-102, DOI: 10.1109/ICINIS.2013.32.

Habtemichael, F. G. and Cetin, M. (2016). "Short-term traffic flow rate forecasting based on identifying similar traffic patterns." *Transportation Research Part C: Emerging Technologies*, Vol. 66, pp. 61-78, DOI: 10.1016/j.trc.2015.08.017.

Hou, Y., Edara, P., and Sun, C. (2015). "Traffic flow forecasting for urban work zones," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 4, pp. 1761-1770, DOI: 10.1109/TITS.2014.2371993.

Hu, W., Yan, L., Liu, K., and Wang, H. (2016). "A short-term traffic flow forecasting method based on the hybrid PSO-SVR." *Neural Processing Letters,* Vol. 43, No. 1, pp. 155-172, DOI: 10.1007/s11063-015-9409-6.

Huang, R. and Sun, S. (2013). "Kernel regression with sparse metric learning." *Journal of Intelligent & Fuzzy Systems*, Vol. 24, No. 4, pp. 775-787, DOI: 10.3233/IFS-2012-0597.

Ji, H., Xu, A., Sui, X., and Li, L. (2010). "The applied research of Kalman in the dynamic travel time prediction," *The 18th International Conference on Geoinformatics*, Beijing, China, pp. 1-5, DOI: 10.1109/GEOINFORMATICS.2010.5567722.

Li, L., Su, X., Wang, Y., Lin, Y., Li, Z., and Li, Y. (2015). "Robust causal dependence mining in big data network and its application to traffic flow predictions." *Transportation Research Part C: Emerging Technologies*, Vol. 58, pp. 292-307, DOI: 10.1016/j.trc.2015.03.003.

Lin, L., Li, Y., and Sadek, A. (2013). "A k nearest neighbor based local linear wavelet neural network model for on-line short-term traffic volume prediction," *Procedia - Social and Behavioral Sciences*, Vol. 96, pp. 2066-2077, DOI: 10.1016/j.sbspro.2013.08.233.

Lippi, M., Bertini, M., and Frasconi, P. (2013). "Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning." *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 2, pp. 871-882, DOI: 10.1109/TITS.2013.2247040.

Min, W. and Wynter, L. (2011). "Real-time road traffic prediction with spatio-temporal correlations." *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 4, pp. 606-616, DOI: 10.1016/j.trc.2010.10.002.

Moretti, F., Pizzuti, S., Panzieri, S., and Annunziato, M. (2015). "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling." *Neurocomputing*, Vol.167, No. C, pp. 3-7, DOI: 10.1016/j.neucom.2014.08.100.

Okutani, I. and Stephanedes, Y. J. (1984). "Dynamic prediction of traffic volume through Kalman filtering theory." *Transportation Research Part B: Methodological*, Vol. 18, No. 1, pp. 1-11, DOI: 10.1016/0191-2615(84)90002-X.

Pan, T. L., Sumalee, A., Zhong, R. X., and Indra-Payoong, N. (2013). "Short-term traffic state prediction based on temporal–spatial correlation." *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 3, pp. 1242-1254, DOI: 10.1109/TITS.2013.2258916.

Smith, B. L., Williams, B. M., and Oswald, R. K. (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting." *Transportation Research Part C: Emerging Technologies*, Vol. 10, No. 4, pp. 303-321, DOI: 10.1016/S0968-090X (02)00009-8.

Sun, S., Huang, R., and Gao, Y. (2012). "Network-scale traffic modeling and forecasting with graphical lasso and neural networks." *Journal of Transportation Engineering*, Vol. 138, No. 11, pp. 1358-1367, DOI: 10.1061/ (ASCE) TE.1943-5436.0000435.

Sun, S. and Xu, X. (2011). "Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 12, No. 2, pp. 466-475, DOI: 10.1109/TITS.2010.2093575.

Sun, S., Zhang, C., and Yu, G. (2006). "A Bayesian network approach to traffic flow forecasting." *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7, No. 1, 124-132, DOI: 10.1109/TITS.2006.869623.

Szeto, W. Y., Ghosh, B., Basu, B., and O'Mahony, M. (2009). "Multivariate traffic forecasting technique using cell transmission model and SARIMA model," *Journal of Transportation Engineering*, Vol. 135, No. 9, pp. 658-667, DOI: 10.1061/(ASCE)0733-947X (2009) 135:9(658).

Voort, M. V. D., Dougherty, M., and Watson, S. (1996). "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologie*s, Vol. 4, No. 5, pp. 307-318, DOI: 10.1016/S0968-090X (97)82903-8.

Wang, Y. and Papageorgiou, M. (2005). "Real-time freeway traffic state estimation based on extended Kalman filter: A general approach," *Transportation Research Part B: Methodological*, Vol. 39, No. 2, pp. 141-167, DOI: 10.1016/j.trb.2004.03.003.

Wei, C. C., Chen, T. T., and Lee, S. J. (2013). "KNN Based neuro-fuzzy system for time series prediction." *The 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/distributed Computing*, Honolulu, USA, pp. 569-574, DOI: 10.1109/SNPD.2013.68.

Williams, B. M. and Hoel, L. A. (2003). "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results." *Journal of Transportation Engineering*, Vol. 129, No. 6, pp. 664-672, DOI: 10.1061/(ASCE)0733-947X (2003)129:6(664).

Xie, Y., Zhang, Y., and Ye, Z. (2007). "Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition." *Computer-Aided Civil and Infrastructure Engineering*, Vol. 22, No. 5, pp. 326-334, DOI: 10.1111/j.1467-8667.2007.00489.x.

Yang, H. J. and Hu, X. (2016). "Wavelet neural network with improved genetic algorithm for traffic flow time series prediction." *Optik - International Journal for Light and Electron Optics*, Vol. 127, No. 19, pp. 8103-8110, DOI: 10.1016/j.ijleo.2016.06.017.

Zhang, Y. and Ye, Z. (2008). "Short-term traffic flow forecasting using fuzzy logic system methods." *Journal of Intelligent Transportation Systems*, Vol. 12, No. 3, pp. 102-112, DOI: 10.1080/15472450802262281.