Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, October 6-9, 2013

A Spatio-Temporal Multivariate Adaptive Regression Splines Approach for Short-Term Freeway Traffic Volume Prediction

Yanyan Xu, Qing-Jie Kong and Yuncai Liu

Abstract— **Current freeway traffic flow prediction techniques pay attention to time series prediction or introduce the upstream adjacent road segments in the short-term prediction model. In this paper, all of the road segments on the freeway are considered as candidates of the independent variables fed into the prediction model. A spatio-temporal multivariate adaptive regression splines (MARS) approach is proposed for the road network analysis and to predict the short-term traffic volume at the observation stations on the freeway. The actual traffic data are collected from a series of observation stations along a freeway in Portland every 15 minutes. In the first phase, the macroscopic dependency relationships of the stations on the freeway are investigated via MARS method. Subsequently the stations most related to the object station are selected and fed into the MARS prediction model to generate the short-term volume. The experiments are carried out on the actual traffic data and the results indicate that the proposed spatio-temporal MARS model can generate superior prediction accuracy in contrast with the historical data based MARS model, the parametric ARIMA, and the nonparametric PPR methods.**

I. INTRODUCTION

In recent years, as an efficient realization of intelligent transportation systems (ITS), parallel-transportation management systems (PtMS) have been applied to extenuate the transportation pressure in large cities by degrees [1], [2]. In PtMS, the short-term traffic flow prediction in the freeways plays a significant role in some concrete components, such as the artificial transportation systems (ATS) and the traffic information services (TIS).

Since several decades ago, various approaches have been proposed and tested to ameliorate the short-term prediction of traffic flow on freeways based on different models. From the early parametric to the subsequent non-parametric methods, historical traffic data on the object road has been considered as the most important factor to the prediction model. For instance, researchers have taken advantage of the historical data to predict short-term traffic flow through Kalman filtering [3], autoregressive integrated moving average (ARIMA) [4], nonparametric regression method such as k-nearest neighbor (k-NN) approach [5], regression trees approach [6]. These methods also can be seen as univariate methods as the univariate historical values on the single object road are fed into the prediction model. On the basis of considering the traffic flow as time series, these approaches mostly could promise well when the traffic relatively remained stable, instead of the complicated situations.

In the recent two decades, researchers perceived the important of the influence of the spatial information in the traffic flow prediction. Hobeika *et al.* [7] tried to predict short-term traffic flow based on the historical and upstream traffic states. Chandra *et al.* [8] developed a vector autoregressive model considering the spatial contributions of the upstream roads. The relationship between traffic flow on the current section and the upstream stations can be used for predicting in Zhang *et al.*'s paper [9]. Besides, machine learning approaches have also been extensively utilized to deal with the short-term traffic flow prediction, especially v-Support Vector Machines [10], Bayesian combined neural network approach [11], stochastic approach [12] and so on.

The above mentioned spatio-temporal correlation models are developed to predict the current road's traffic flow taking advantage the upstream traffic. However, the other spatial traffic states from the road segments or stations which are not immediately adjacent are neglected. In this paper, a multivariate spatio-temporal correlation model based on a collection of observation stations is developed to predict the freeway traffic flow. We first describes a spatio-temporal multivariate adaptive regression splines (MARS) model to mining the correlative dependence relationships among the observation stations. Following the variables importance investigation, the short-term traffic volume is predicted using the MARS prediction model with the data from the selected most related stations as inputs. Finally, in the experiment stage, the actual traffic data collected from a series of observation stations along a freeway in Portland every 15 minutes are exploited to verify the effectivity of the proposed predictive model. The results indicate that the proposed spatio-temporal MARS model can generate more preferable prediction in contrast with the historical data based MARS model, the parametric method ARIMA, and the nonparametric PPR methods.

The remainder of this paper is structured as follows: section II is a brief introduction on the data set used in our work; section III describes the basic theory of MARS model concisely; the details of the spatio-temporal model building and the experiment results are illustrated and analyzed in Section IV, moreover, other two prediction methods are implemented for comparison with our model; finally, some concluding remarks are given in Section V.

*This work is partly supported by China National 863 Key Program under Grant 2012AA112307, Shanghai STCSM Program under Grant 11231202801, and China NSFC Program under Grant 61104160

Yanyan Xu and Yuncai Liu are with the Department of Automation, Shanghai Jiao Tong University & Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China. Email: {*xustone, whomliu*}*@sjtu.edu.cn*.

Qing-Jie Kong is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Email: *kongqingjie@gmail.com*.

978-1-4799-2914-613/$31.00 ©2013 IEEE

II. Data Set Description

The work in this paper concentrates on the short-term prediction of the traffic volume on the freeways by taking full advantage of the spatial and temporal information. Therefore, we employ the traffic volume obtained from the observation stations along a long-distance freeway. The data is drawn from the PORTAL FHWA Test Data Set [13] developed by Portland State University. The current PORTAL system receives the traffic volume every 20 seconds from the freeway loop detectors, which are installed in the main line lanes and on-ramps on the Portland-Vancouver metropolitan region freeways. In the data set, an observation station has a set of related loop detectors.

The data set used in this paper is collected from eight adjacent stations located on freeway Interstate 205 (I-205) aligning from south to north. Figure 1 shows the distribution of the eight chosen observation stations on the I-205. More details about the number of lanes on the freeways, the specific locations, and the lengths of the stations are illustrated in Table I. Milepost denotes the location of the observation station on the freeway.

The traffic volumes were collected from February 24 to March 23, 2013. The univariate traffic volume observations were obtained over each 15 minutes interval. The data from February 24 to March 16 were the training data set; the latter one week were processed as the test data set to evaluate the developed prediction model. In addition, the traffic volume is formatted as the average number of vehicles per lane per hour (VPLPH).

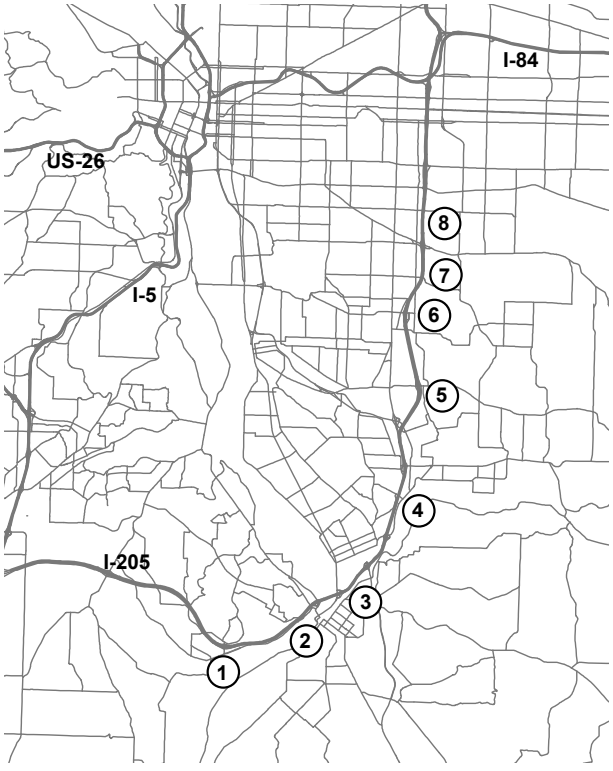

Fig. 1. Locations of the observation stations on I-205 in Portland

TABLE I
PROPERTIES OF SELECTED STATIONS

Stations	Lanes	Milepost	Length
1. 10th Street to I-205 NB	2	6.88	2.63
2. ORE 43 SB-NB	2	8.8	1.07
3. ORE 99E NB	2	9.45	1.01
4. Gladstone NB	3	11.05	1.75
5. Clackamas Hwy NB	3	12.94	1.27
6. Lawnfield NB	3	13.58	0.69
7. Sunnybrook NB	3	14.32	0.37
8. Sunnyside NB	4	14.32	0.94

III. Overview of the MARS Method

Multivariate adaptive regression splines (MARS) proposed by Friedman [14] is a hybrid nonparametric regression approach which can automatically model non-linearities and interactions between high-dimensional predictors and responses. It is a spline regression model that uses a specific class of base functions as predictors in place of the original data. MARS has been applied to a wide variety of data analyses in recent years, including traffic flow prediction [15].

The core idea of MARS is to build flexible regression function as a sum of basis functions, each of which has its support on a distinct region [16]. Within a region, the regression function reduces to a product of simple functions that are initially constant but can be chosen as splines. In particular, MARS uses expansions in piecewise linear basis functions of the form $(x-t)^+ = max(0, x-t)$, where t is a univariate constant, named the knot, and the $+$ indicates the positive part. Therefore, assuming X is composed by N predictors with p-dimensional vectors, the the collection of basis functions is

$$C = \left\{(X_j - t)^+, (t - X_j)^+\right\} \tag{1}$$

where $t \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\}$ and $j = 1, 2, \ldots, p$.

The model-building strategy of MARS is like a forward stepwise linear regression. Based on the preprocessing, functions from the set C and their products are allowed to be used in MARS, Finally, the MARS model is expressed as

$$Y = f(X) + \epsilon = \beta_0 + \sum_{j=1}^{r} \beta_m h_m(X) + \epsilon \tag{2}$$

where each $h_m(X)$ is a function in C, or a product of two or more such functions. These functions serve as a set of functions representing the relationship between the predictor variables X and the target variable Y. The error term ϵ is the Gaussian white noise produced in the data collection stage.

The "optimal" $f(X)$ in the MARS model is achieved in a two-stage process. In the first forward stepwise stage, a model is grown by adding basis functions selected from set C until an overly large model is found. In other words, the selection of basic functions from the initial set is achieved by determining a constant function $h_0(X) = 1$, and all functions in the set C are candidate functions. Meanwhile given a choice for the h_m, the coefficients β_m are estimated by minimizing the residual sum-of-squares. During the stage,

978-1-4799-2914-613/$31.00 ©2013 IEEE

new pairs of functions are considered at each phase until the model has the maximum number of terms specified at the beginning of the process.

In the second backward stepwise stage, basis functions are deleted step by step in order of least contribution to the model until an optimal balance of bias and variance is found. The backward removal is performed by suppressing those model terms that contribute to a minimal residual error. This stage consists of reducing the complexity of the model complexity by increasing its generalisability. This process can be conducted by means of generalized cross validation (GCV):

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \widehat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2} \tag{3}$$

Where $M(\lambda)$ indicates the effective number of parameters in the model and can be estimated with:

$$M(\lambda) = r + cK \tag{4}$$

where r is the number of linearly independent basic functions and K is the number of knots selected in the forward process.

Finally, by allowing for any arbitrary shape for the response function as well as for interactions, and by using the two-stage model selection method, MARS is capable of reliably tracking very complex data structures that often hide in high-dimensional data.

IV. Experiments and Discussions

In order to build the MARS model, the data set is divided into two parts: training set and testing set. The training set continues for 3 weeks, and consists of 2016 time intervals. The training set is used to build the MARS model and analyze the spatio-temporal characteristic of the traffic flow between the freeway observation stations. The testing set consists of the remaining 672 time intervals and is used to evaluate the performance of the proposed predictive model.

The input variables consist of the current traffic volume V_t and the former historical volumes at the 8 observation stations in the investigated freeway. The time lag of the historical data is equal to 4 in our project. Therefore, X is a collection of $\{V_t, V_{t-1}, \ldots, V_{t-4}\}$ from all the stations. Y is the observation average traffic volume V_{t+1} in the later 15 minutes.

A. Spatio-Temporal Relationships Analysis

As mentioned in Section III, MARS models include a backwards elimination feature selection routine that looks at reductions in the GCV estimate of error. Therefore, we track the GCV changes during the building of the model for each predictor. The importance of the variables can be estimated via accumulating the reduction in the statistic when each predictor's feature is added to the model. If a predictor (including spatial and temporal traffic volume) was rarely or never used in any MARS basis function, it has little or no influence on the specified freeway station.

In our study, the GCV importances are normalized to 0 to 100. And hence 100 denotes that the predictor is the most important one among all of the predictors, while 0 means that the variable is useless to the response. Taking a deep look into the observation station $S3$, the variable importances for $S3\_(t+1)$ are plotted in Fig. 2. Other variables not appearing in the figure are unused in the model. It is clear from the figure that the variable $S3\_t$ certainly has the most important influence on $S3\_(t+1)$ by a comfortable margin. Other variables listed in the figure are the upstream or downstream stations, such as $S1$, $S4$, and $S8$. Therefore, from the variable importances figure, we can conclude that the most influential stations to $S3$ along the freeway are $S1$, $S4$, and $S8$. The fact that there are two downstream variables in the figure signifies that the traffic status on the downstream roads equally impact the traffic state on the current road segment.

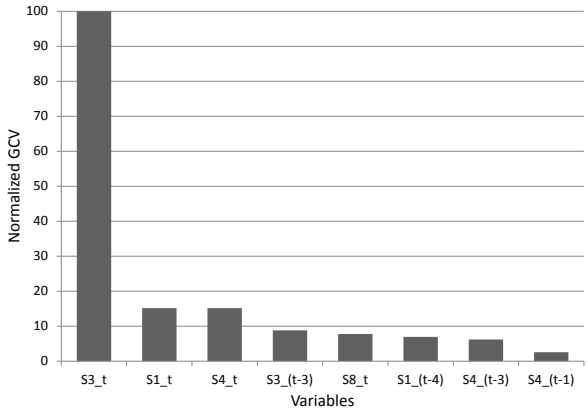

Fig. 2. The importance of the traffic varaibles related to station 3

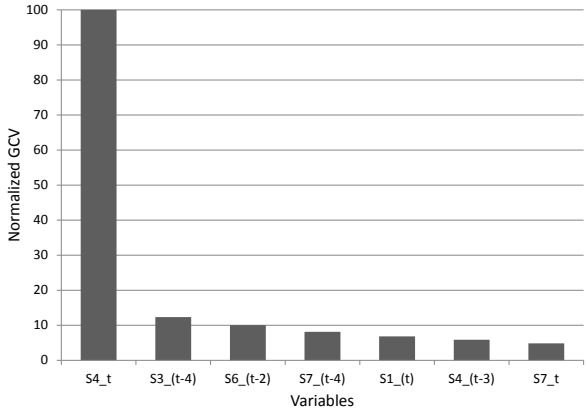

Fig. 3. The importance of the traffic variables links related to station 4

As another example the variable importances chart of station $S4$ is drawn in Fig. 3. The effective stations related to $S4$ are $S3$, $S6$, $S7$, and $S1$ with separate time lag.

B. Prediction Results Analyses

Following the spatio-temporal relationships analysis of the traffic states on the freeway, this paper predicts the short-term traffic volume on the eight observation stations. During the traffic prediction stage, the current and historical traffic volumes on current and the most related observation states are treated as the input of the MARS prediction model. Otherwise, the output of the prediction model is

the short-term volume at the current station. To reflect the contributions of the spatial traffic states to the object station, a temporal MARS model (temp-MARS) is also employed to compare with the proposed spatio-temporal MARS (ST-MARS) model. Moreover, two classic prediction models including parametric and non-parametric methods are also implemented on the testing data set and compared together. The parametric one is the 3-order autoregressive integrated moving average (ARIMA) model. The nonparametric one is the projection pursuit regression (PPR) method.

As evaluating indicators, two measures for forecasting error analysis, root mean square error (RMSE) and mean absolute scaled error (MASE) proposed by Rob Hyndman [17], are adopted in this research to evaluate the performance of the proposed model. RMSE and MASE are defined as follows:

$$RMSE = \sqrt{\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_k - \hat{V}_k\right)^2\right]} \tag{5}$$

$$MASE = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{V_k - \hat{V}_k}{\frac{1}{K-1}\sum_{k=2}^{K}|V_k - V_{k-1}|}\right| \tag{6}$$

where K is the total number of intervals during the testing stage; V_k denotes the actual traffic volume; $\hat{V}_k$ is the prediction value produced by the proposed model. Different from RMSE, MASE is a sort of scaled error that takes account of the gradient of the actual values. The smaller MASE indicates better prediction.

In our experiments, all the 8 observation stations are predicted during the testing period, including weekdays and weekends. Fig. 4 and 5 plot the 15 minutes traffic volume prediction results of the predictive models together with the actual volume for station $S3$ on March 17 and 18, respectively. As Sunday, the volume on March 17 keeps a high level at midday, and our ST-MARS model can follow this state closely. Differently, the volume on Monday contains the morning and evening peak as shown in Fig. 5. At the beginning of the morning peak during 6:00 to 7:00, our ST-MARS model can follow the climbing more closely than other models. Moreover, the ST-MARS also performs much better during the descent phase after 8:00. In addition, Fig. 6 draws the prediction results of station 5 on March 18. The figures show that the proposed ST-MARS can follow the trace of the actual value during stable traffic states or peaks, on weekdays or weekends.

Furthermore, to precisely weight the proposed ST-MARS prediction model against other models, the numerical errors of the prediction approaches for comparison are exhibited in Table II and III. From the tables we can catch that the performances of the proposed ST-MARS model surpass the temp-MARS, ARIMA, and PPR method on all of the observation stations. From Table II, the average RMSE are reduced by about 7%, 13%, and 8% relative to the temp-MARS, ARIMA, and PPR methods according to RMSE, respectively. Furthermore, by the look of the MASE errors in Table III, the MASE of ST-MARS is much less than 1 and performs much better than the other three models. To sum up, we can conclude that ST-MARS also gets ahead of the temporal MARS approach, ARIMA, and PPR in general.

TABLE II
RMSE COMPARISON OF THE SELECTED STATIONS

Station ID	RMSE			
	ST-MARS	temp-MARS	ARIMA	PPR
1	89.13	99.91	103.70	98.21
2	140.03	161.48	159.44	158.35
3	95.48	110.57	120.11	112.86
4	80.22	84.83	97.21	85.57
5	79.78	82.90	93.19	83.55
6	78.99	79.86	88.49	81.47
7	92.20	96.37	102.17	99.60
8	79.92	77.27	85.63	80.06
Average	**91.97**	**99.15**	**106.24**	**99.96**

TABLE III
MASE COMPARISON OF THE SELECTED STATIONS

Station ID	MASE			
	ST-MARS	temp-MARS	ARIMA	PPR
1	0.87	0.94	1.01	0.92
2	0.87	0.96	0.99	0.95
3	0.77	0.91	1.00	0.92
4	0.84	0.86	0.99	0.87
5	0.83	0.87	0.99	0.87
6	0.84	0.89	1.01	0.90
7	0.94	0.96	1.03	1.01
8	0.88	0.90	1.01	0.92
Average	**0.855**	**0.911**	**1.004**	**0.920**

V. CONCLUSIONS

This paper has presented a spatio-temporal multivariate adaptive regression splines approach for the roads relevance analysis and prediction of the short-term traffic volume on the freeway. The traffic data set is collected from the observation stations on a freeway in Portland every 15 minutes. In the first stage, a MARS model is designed to build the dependency relationships of the average traffic volumes between the observation stations and their historical values. For each station on the freeway, a set of variables are found up with the strongest interrelated ones. Afterwards, the historical volume on current and the most interrelated volumes are fed into the MARS prediction model to predict the short-term traffic flow.

Finally, in order to evaluate the performance of the proposed prediction model, the historical data based MARS, the ARIMA and the PPR methods are employed for comparisons. The experiment results indicate that the spatio-temporal MARS model is an efficient approach for short-term traffic volume prediction on freeway.

REFERENCES

[1] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 630–638, 2010.

978-1-4799-2914-613/$31.00 ©2013 IEEE

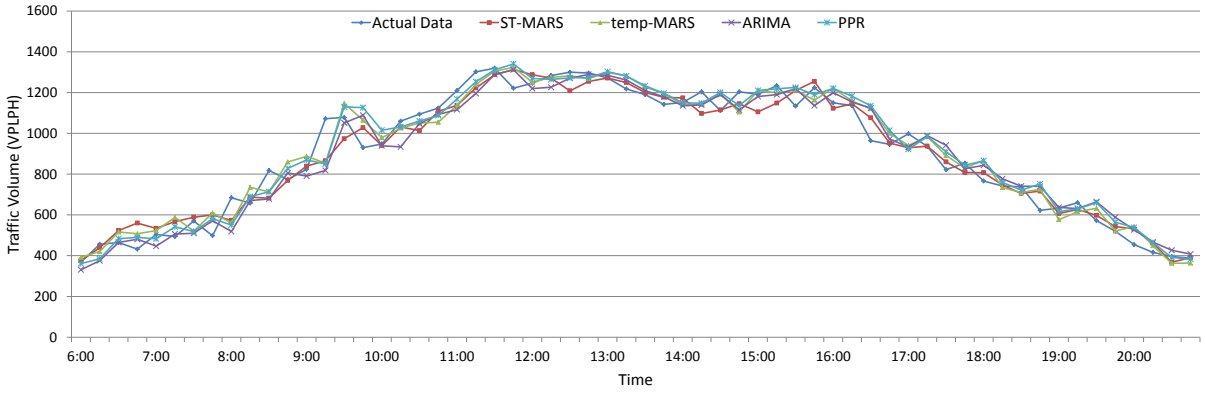

Fig. 4. Prediction of the traffic volume for station 3 on March 17

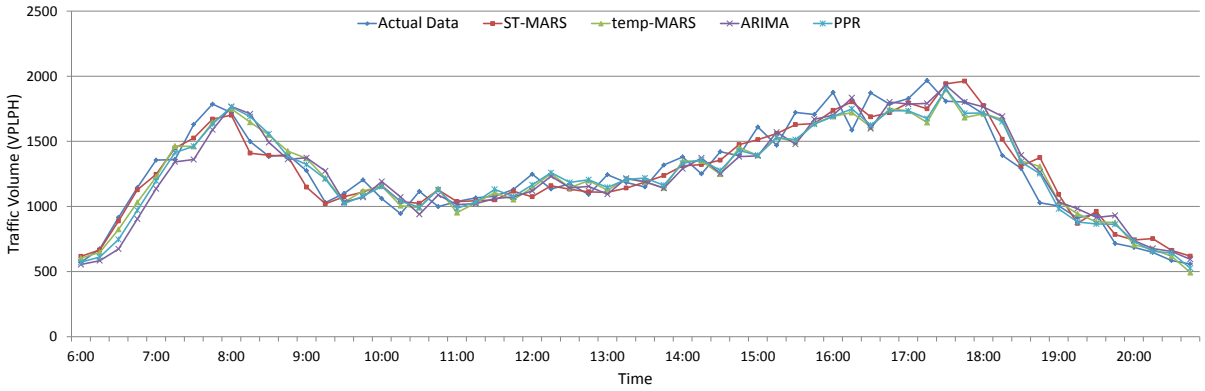

Fig. 5. Prediction of the traffic volume for station 3 on March 18

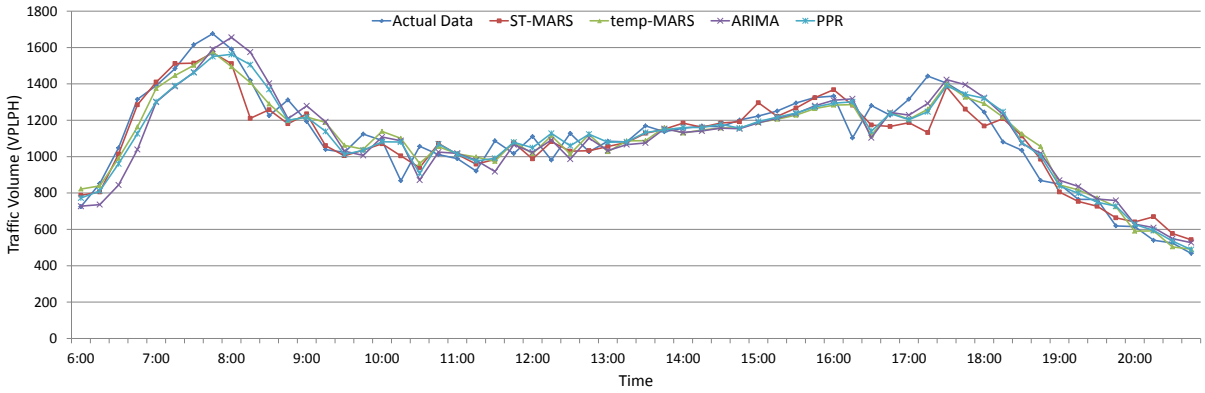

Fig. 6. Prediction of the traffic volume for station 5 on March 18

[2] G. Xiong, S. Liu, X. Dong, F. Zhu, B. Hu, D. Fan, and Z. Zhang, "Parallel traffic management system helps 16th asian games," *Intelligent Systems, IEEE*, vol. 27, no. 3, pp. 74–78, 2012.

[3] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[4] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway travel prediction: application of seasonal arima and exponential smoothing models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, pp. 132–141, 1998.

[5] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[6] Y. Xu, Q.-J. Kong, and Y. Liu, "Short-term traffic volume prediction using classification and regression trees," in *The 2013 IEEE Intelligent Vehicles Symposium*, Gold Coast, Australia, 2013, Accepted.

[7] A. Hobeika and C. K. Kim, "Traffic-flow-prediction systems based on upstream traffic," in *Proc. Vehicle Navigation and Information Systems Conference*, Yokohama, Japan, 1994, pp. 345–350.

[8] S. R. Chandra and H. Al-Deek, "Predictions of freeway traffic speeds and volumes using vector autoregressive models," *Journal of Intelligent Transportation Systems*, vol. 13, no. 2, pp. 53–72, 2009.

[9] P. Zhang, K. Xie, and G. Song, "A short-term freeway traffic flow prediction method based on road section traffic flow structure pattern," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, 2012, pp. 534–539.

[10] Y. Zhang and Y. Xie, "Forecasting of short-term freeway volume with v-support vector machines," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2024, pp. 92–99, 2007.

[11] W. Zheng, D.-H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach," *Journal of*

Transportation Engineering, vol. 132, no. 2, pp. 114–121, 2006.
[12] Y. Qi and S. Ishak, "Stochastic approach for short-term freeway traffic prediction during peak periods," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–13, 2012.
[13] (Accessed Apr. 1, 2013) The portal fhwa traffic data set. Portland State University. [Online]. Available: http://portal.its.pdx.edu/
[14] J. H. Friedman, "Multivariate adaptive regression splines," *The Annual of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.
[15] S. Ye, Y. He, J. Hu, and Z. Zhang, "Short-term traffic flow forecasting based on mars," in *Proc. 5th International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5, Jinan, Shandong, 2008, pp. 669–675.
[16] B. Clarke, E. Fokoué, and H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, ser. Springer Series in Statistics. Berlin, Germany: Springer-Verlag, 2009.
[17] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.